ABORDAGENS ADVERSARIAIS PARA EXPLICAÇÃO DE IMAGENS EM REDES NEURAIS PROFUNDAS

ANTONIO JOSE SOBRINHO JUNIOR

ABORDAGENS ADVERSARIAIS PARA EXPLICAÇÃO DE IMAGENS EM REDES NEURAIS PROFUNDAS

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Marco Antônio Pinheiro de Cristo

Manaus

Agosto de 2021

Ficha Catalográfica
Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Junior, Antonio Jose Sobrinho

J95a Abordagens adversariais para explicação de imagens em redes neurais profundas / Antonio Jose Sobrinho Junior. 2021
84 f. : il. color ; 31cm

Orientador: Marco Antônio Pinheiro de Cristo
Dissertação (Mestrado em Informática) — Universidade Federal do Amazonas

1. Explicação. 2. Aprendizagem de máquina. 3. Modelos complexos. 4. Superfície de decisão. 5. Amostras adversariais. I. Cristo, Marco Antônio Pinheiro de. II. Universidade Federal do Amazonas. III. Título.

FOLHA DE APROVAÇÃO

"Abordagens Adversariais para Explicação de Imagens em Redes Neurais Profundas"

ANTÔNIO JOSÉ SOBRINHO JÚNIOR

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Marco Antônio Pinheiro de Cristo - PRESIDENTE

Prof. Daniel Hasan Dalip - MEMBRO EXTERNO

Profa. Eulanda Miranda dos Santos - MEMBRO INTERNO

Manaus, 20 de Agosto de 2021

Dedico este trabalho aos meus pais, por serem minha base e referência, por acreditarem na educação como forma de melhoria de vida e a quem eu honro e sou grato pela existência.

Agradecimentos

Gostaria de agradecer a Deus por ter sido meu guia, através de pessoas, situações e oportunidades, num processo de construção deste trabalho que envolveu muito além do contexto acadêmico, envolvendo diversas áreas da minha vida, me fazendo valorizar mais ainda todo o processo e não apenas os resultados finais deste trabalho.

Agradeço também aos meus pais, Ivaneide e Antônio, meus verdadeiros mestres e a quem sou grato por todo o aprendizado e troca. Os primeiros a acreditar e a segurar minha mão. Meu pai, você não está aqui neste plano, mas eu sei que onde quer que você esteja, você está vibrando por mais essa etapa. Minha mãe, obrigado por, muitas vezes, sem querer, me ensinar tanto. Amo vocês! Essa conquista é nossa.

Ao meu orientador, Marco Cristo, sempre empático comigo e com meu processo de entrega deste trabalho, por não desacreditar, sempre contribuir, agregar, focar na solução, no que realmente importa.

A todos os professores do PPGI que desde a graduação instigam seus alunos a busca a qualidade nas entregas e por impactarem positivamente a minha vida acadêmica também desde a graduação. Em especial aos professores Eulanda dos Santos e Juan Colonna pelos inputs pertinentes a este trabalho ao longo de sua construção. E ao professor Eduardo Feitosa, representando a coordenação do PPGI, por ser acessível e por resolver questões administrativas que surgissem, sempre com leveza.

Ao CESAR, empresa que estou trabalhando atualmente, meu agradecimento especial pelo ambiente de trabalho plural e incrível, pelas pessoas incríveis, pela evolução nesses últimos anos e por reacender a chama da paixão pela informática que me fez voltar ao PPGI e a este trabalho, sem impedimentos. Especialmente aos gerentes que trabalharam comigo nesse período Willian Albuquerque e Aleandro Anjos, obrigado pelo apoio e incentivo na entrega do mestrado.

Aos meus amigos e apoiadores, foi incrível constatar que eu podia contar com vocês. Alguns de vocês são: Juliana Nunes, Vivian Lô, Yumi Ouchi, Adriane Almeida, Andrews Melo, Claudia Rejane, Emidio Oliveira, Fabiana Bezerra, Lennon Chaves, Rodrigo Borba, Bruna Grecia, Luana Grécia, Paloma Maquiné, entre outros...

"Não tenho um caminho novo. O que eu tenho de novo é um jeito de caminhar."
(Thiago de Mello)

Resumo

O uso de modelos de previsão cada vez mais complexos tem se tornado comum em uma variedade de aplicações. Em muitas delas, a opacidade destes modelos representa um desafio na detecção de vícios ou injustiças relacionados ao processo decisório, o que pode contribuir para as percepção de desconfiança e discordância quanto aos fatores determinantes para os resultados, impactando o entendimento de seus usuários, principalmente quando estes modelos são usados como base para decisões consideradas críticas ou sensíveis. Tal desafio motiva o desenvolvimento de estratégias de explicação globais (por que o modelo toma certas decisões?) e locais (por que o modelo toma certas decisões para uma certa instância?). Nesta última, devido à complexidade da superfície de decisão obtida por modelos complexos, têm sido comum adotar análise de vizinhança com amostras sinteticamente geradas para (a) buscar a superfície de separação mais próxima, (b) isolar as características e (c) fornecer explicações que sustentem a amostra sintética de interesse (alvo de explicação) nesta região do plano. Neste trabalho propomos abordagens que combinam a utilização dos conceitos de modelos adversariais com o objetivo de explicação de instâncias. Partimos da suposição de que uma instância adversarial é um bom ponto de partida para a análise do menor esforço necessário para uma transição entre classes. Com base nisso, propomos meios de melhorar esta descrição inicial para que ela seja vista como mais apropriada pelo usuário para a qual a explicação é dada. Ao fim, a explicação consiste de uma descrição da sensibilidade dos atributos para a decisão tomada por um modelo complexo. As explicações fornecidas por nossos métodos para justificar classes incorretamente fornecidas, por uma CNN, para digitos manuscritos da coleção MNIST, foram avaliadas por usuários em testes cegos. Estes as consideraram significativamente melhores, em 68 a 74% dos casos, que as fornecidas por outros dois baselines da literatura, um dos quais também baseado em uma estratégia adversarial.

Palavras-chave: Explicação, Aprendizagem de Máquina, Modelos Complexos, Superfície de Separação, Amostras Adversariais.

Abstract

Complex machine learning models have been increasingly adopted due its range of well-succeded applications. In many applications, the poor knowledge on the inside aspects of a complex model represents a challenge in detecting biases or injustices related to the decision-making process. That might directly affect the confidence on the model so its results may be considered not reliable, which can be risky when such models are responsible for critical or sensitive decisions. These problems motivate the development of methods to explain the reasons behind its decisions in a global ("Why did this model make this decision?") or local ("Why was this category assigned to this sample?") fashion. In this work, we focus on the local explanation problem. Given the complexity of the decision boundary on complex models, e.g. deep neural networks, it has been common to adopt explanation approaches based on neighborhood analysis, to specify the unique features and sensitivity of a sample in comparison to a neighbor class. The idea behind these approches is to find the closest separation surface which supports the samples we want to explain. Since many adversarial techniques explore this same region to fool machine learning models, it may be possible to use similar ideas to explain a class transition of a sample in terms of feature importance. Thus, in this work, we propose an approach that combines adversarial model concepts with instance explaining requirements. We asssume that an adversarial instance is a good starting point to estimate the minimum effort for a class transition. Then, we propose ways to improve this initial description to obtain one that is more suitable for users according to their perception. By means of blind tests, users have evaluated the explanations provided by our methods for justifying errors of a CNN when used to classify images of the MNIST dataset. In 68 to 74% of the judments, they considered the provided explanations significantly better than those provided by two other baselines published in the literature, one of which is also based on an adversarial strategy.

Keywords: Explainable Artificial Intelligence (XAI), Machine Learning, Complex Models, Decision Boundary, Adversarial Examples.

Lista de Tabelas

Lista de Abreviações e Siglas

AGN Agnóstico

CAV *Concept Activation Vectors*

C&W *Carlini & Wagner*

DNN *Deep Neural Network*

DVE *Deep Visual Explanation*

EBR Explicador Baseado em Regras

ER Engenharia Reversa

FGSM *Fast Gradient Sign Method*

GBL Global

GDPR *General Data Protection Regulation*

IHC Interação Humano-Computador

IMC Importância de Características

IMG Imagem

ISP Inspeção de Modelo

JSMA *Jacobian-Based Saliency Map Attack*

LACE *Local Agnostic Attribute Contribution Explanation*

LCL Local

LIME *Local Interpretable Model-Agnostic Explanations*

LORE *Local Rule-Based Explanations*

LGPD Lei Geral de Proteção dos Dados

LRP *Layer-wise Relevance Propagation*

MDS Máscara de Saliência

MJSMA *Maximal Jacobian-Based Saliency Map Attack*

N Não

ORM Orientado a Modelo

LARE-2M *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy*

LARE-MS *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy*

PAG Perturbação por Algoritmo Genético

PBG Perturbação Baseada em Gradiente

PBV Perturbação Baseada em Vizinhança

PE Projeção de Explicação

PMT Projeção de Modelo Transparente

PRD Perturbação Randômica

RCN *Rule-Constrained Networks*

RNCP Rede Neural Convolucional Profunda

RNP Rede Neural Profunda

S Sim

SA *Sensitive Analysis*

SHAP *SHapley Additive exPlanations*

SVM *Support Vector Machine*

TAB Tabular

TCAV *Testing with Concept Activation Vectors*

TXT Texto

VD Vídeo

XAI *Explainable Artificial Intelligence*

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

Neste capítulo serão apresentados o Contexto da área estudada por esta pesquisa, a Motivação que envolve o estudo da área de explicação, bem como uma descrição do Problema e Hipótese deste trabalho, esclarecendo os Objetivos e Metodologia adotados. Nós finalizamos descrevendo a Organização desta dissertação.

1.1 Contexto

A evolução recente da área de aprendizagem de máquina permitiu a criação de modelos bem-sucedidos cada vez mais complexos, como Rede Neural Profunda (RNP) ou *Deep Neural Network (DNN)*. Por conta de sua alta complexidade e capacidade de abstração, eles desafiam a identificação dos padrões aprendidos e compreensão dos mecanismos internos para a tomada de decisão, o que se torna um problema maior quando consideramos que seus resultados geralmente constituem o estado-da-arte em muitas aplicações (Figura 1.1).

A aplicação efetiva destes modelos de previsão baseados em aprendizagem de máquina torna evidente a necessidade de compreensão dos mecanismos e processos internos de tomada de decisão com o objetivo de facilitar a compreensão do conhecimento aprendido e garantir que eles se baseiam em aspectos aceitáveis de comportamento, de acordo com o domínio de aplicação. Por exemplo, para aplicações que envolvem a discriminação de seres humanos tipicamente se espera que o processo decisório seja rastreável, resultando em decisões consistentes e que não envolvam aspectos sensíveis como seleção baseada em raça, gênero ou crenças.

Dada a sua importância, tais aspectos têm sido alvo de regulamentações estatais, como é o caso de decisões do parlamento europeu e do congresso brasileiro, que sancionaram leis de proteção às informações dos cidadãos (Lei Geral de Proteção dos Dados

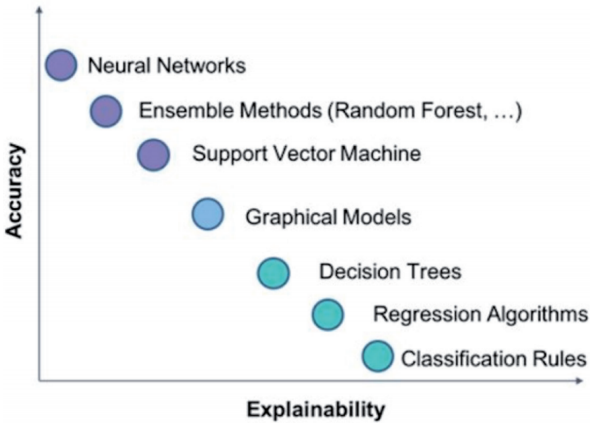

Figura 1.1. Relação Acurácia e Explicabilidade em Modelos de Aprendizagem de Máquina [Dağlarli, 2020].

(LGPD) e *General Data Protection Regulation (GDPR)*). Tais regulamentações concedem aos indivíduos, a oportunidade de se ter acesso à lógica envolvida em decisões de soluções de aprendizagem de máquina que lhe dizem respeito. Uma consequência disso é o interesse crescente em pesquisa na área, como indicado na Figura 1.2, onde podemos observar que, somente na década atual, foram registradas mais de 4000 publicações no portal *ACM Digital Library*[1] relacionadas ao termo *Explainable Artificial Intelligence (XAI)*, por exemplo.

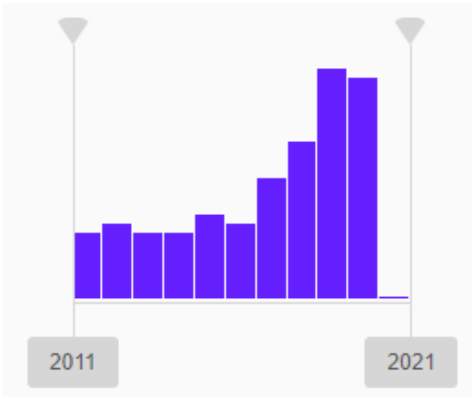

Figura 1.2. Pesquisa pelo termo *Explainable Artificial Intelligence* em ACM Digital Library[1]

Um desafio notório na área de explicação em aprendizagem de máquina é a dificuldade em avaliar uma explicação e determinar seu grau de confiança. As abordagens usadas na literatura dependem de assistência humana para avaliação e consequente aumento de confiabilidade. Estas avaliações dependem de cuidadoso acompanhamento

[1] http://dl.acm.org/

para detectar, por exemplo, possíveis inconsistências e incoerências de um explicador em relação ao modelo original e ao próprio explicador.

Quando se analisa a baixa explicabilidade de um modelo, percebe-se que a complexidade deste modelo pode interferir diretamente, uma vez que em modelos de Rede Neural Profunda (RNP), por exemplo, cada camada do modelo adiciona um novo nível de abstração e não linearidade, o que dificulta a compreensão das transformações sofridas pela informação que flui pelo modelo, conforme demonstrado na Figura 1.3. Como resultado, abordagens da literatura exploram diferentes perspectivas do problema, seja focando na explicação de uma instância particular (modelagem local), seja focando em uma explicação do mecanismo geral de decisão de um modelo (modelagem global).

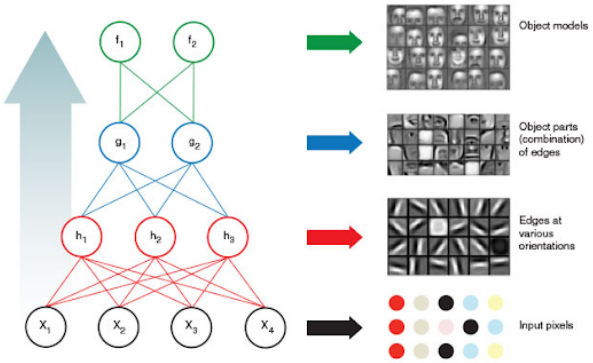

Figura 1.3. Complexidade interna de uma Rede Neural Profunda (RNP)[2]

1.2 Motivação

Com o sucesso de modelos de aprendizagem de máquina cada vez mais complexos, muitas tarefas exclusivas de seres humanos passaram a ser automatizadas. Contudo, a dificuldade em entender as razões pelas quais as decisões são tomadas por tais modelos, os tornam inadequados em áreas onde a tomada de decisão é considerada crítica e causa alto impacto. Por conta disso, muitos trabalhos têm optado por modelos mais simples de forma a privilegiar sua explicabilidade e, por consequência, melhorar sua confiabilidade [Biran & Cotton, 2017; Schuessler & Weiß, 2019; Adadi & Berrada, 2018; Guidotti et al., 2018b]. Uma forma de avaliar a confiabilidade de um modelo é através da receptividade do ser humano a uma explicação. Assim, a alta precisão humana, ligada à sua interpretabilidade, se apresenta como um fator significativo para medição da explicabilidade de modelos complexos [Ribeiro et al., 2018a].

[2] https://str.llnl.gov/june-2016/chen

Além disso, de um ponto de vista científico, ético e legal, há a necessidade do estudo de explicação de modelos complexos [Adadi & Berrada, 2018] de forma a desmistificar o que acontece nestas "caixas-pretas" (Figura 1.4). Com maior transparência, é possível evitar que tais modelos reforcem ou perpetuem formas de injustiça ao aprender maus hábitos a partir dos dados [Guidotti et al., 2018a].

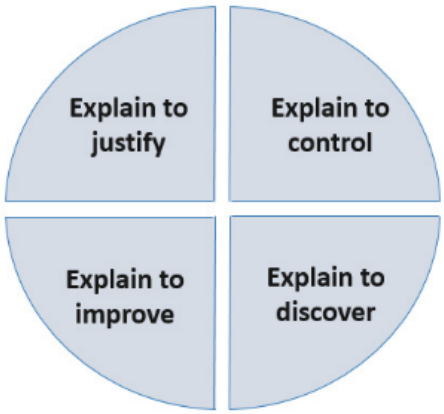

Figura 1.4. Razões para Explicação de Modelos Complexos, extraída de Adadi & Berrada [2018]

Sob a perspectiva de explicação de instâncias (modelagem local), é crescente o interesse na compreensão das superfícies de separação que indicam as regiões de transição entre as classes do modelo e em que localizações poderíamos esperar a instâncias de uma ou outra classe de forma razoável. Portanto, do ponto de vista de modelagem local, é importante investigar a natureza da transição entre classes de forma a entender como executar perturbações que aumentem nossa compreensão sobre a distribuição de instâncias em uma região do modelo. Uma forma possível de ser estudada é agregar os conceitos de ataques adversariais à área de explicação, por seu importante conhecimento sobre tais superfícies de separação, uma vez que seu estudo pode demonstrar áreas e características sensíveis à transição de classes.

Em linhas gerais, modelos explicadores devem contribuir para melhorar a percepção de confiabilidade dos modelos explicados, principalmente, se estes são muito complexos [Schuessler & Weiß, 2019; Adadi & Berrada, 2018; Guidotti et al., 2018b].

1.3 Problema

Embora muitas estratégias de explicação tenham sido desenvolvidas, neste trabalho estamos interessados em métodos locais não agnósticos[3], baseados em técnicas adversariais, aplicados a modelos complexos de classificação (Figura 1.5).

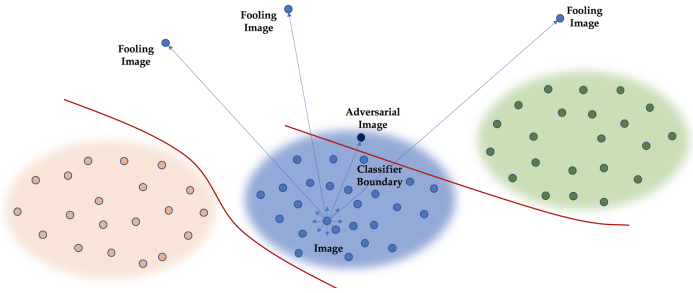

Figura 1.5. Amostras Adversariais e Superfície de Decisão [4]

Formalmente, seja M um modelo complexo, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ uma instância representada com n atributos, para o qual M fornece uma previsão $M(\mathbf{x})$. Considere o interesse em se determinar porque $z \neq M(\mathbf{x})$ não foi a previsão obtida para $\mathbf{x}$. Sem perda de generalidade, uma explicação local pode ser vista como um vetor de perturbação $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$, em nosso caso a ser obtido via técnicas adversariais, combinado a um vetor de reforço de informação $\mathbf{r} = \{r_1, r_2, \ldots, r_n\}$ utilizado para afastamento da área de confusão entre o que é entendido pelo modelo M e o que é percebido pelo usuário, para obtenção de uma amostra da classe z tal que $M(\mathbf{x} + p + r) = z$. Ou seja, uma perturbação $\mathbf{p}$ pode ser usada para determinar as mudanças nos valores dos atributos de $\mathbf{x}$ que produziriam a previsão esperada z e o reforço $\mathbf{r}$ pode ser utilizado como aproximação de características médias da classe z. Assim, a sensibilidade dos atributos de $\mathbf{x}$, inferida através de $\mathbf{p}$ combinado a $\mathbf{r}$, pode ser usada como uma justificativa para a previsão $M(\mathbf{x}) \neq z$. Neste sentido, o problema estudado neste trabalho consiste em se determinar uma perturbação $\mathbf{p}$ e o reforço necessário $\mathbf{r}$, via técnicas de geração de amostras adversariais, para entendimento dos atributos sensíveis a mudança de classe, que possa ser usada para determinar a relevância dos atributos de $\mathbf{x}$ para a previsão $M(\mathbf{x})$.

Para este tipo de estratégia, é necessário estar atento à qualidade do processo de perturbação utilizado, uma vez que este possui impacto direto na qualidade da expli-

[3] Em nosso caso, não esperamos que os modelos propostos neste trabalho sejam aplicáveis a quaisquer tipos de modelos e tipos de explicação, independente da disponibilidade de parâmetros e detalhes dos modelos explicados. Neste sentido, a proposta é não agnóstica.

[4] https://towardsdatascience.com/resisting-adversarial-attacks-using-gaussian-mixture-variational-autoencoders-be98e69b5070

cação. Perturbações inadequadas pouco contribuem para a aquisição do conhecimento sobre os mecanismos internos de previsor, bem como a forma de tomada de decisão de um modelo [Laugel et al., 2018; Jia et al., 2019].

1.4 Hipótese

Uma estratégia possível para obter perturbações adequadas é considerar a distância entre a amostra de interesse e a superfície de decisão. Neste sentido, um $\mathbf{p}$ adequado é aquele tal que $M(\mathbf{x}+p) = z \neq M(\mathbf{x})$ (Figura 1.6), caracterizando a mínima distorção possível.

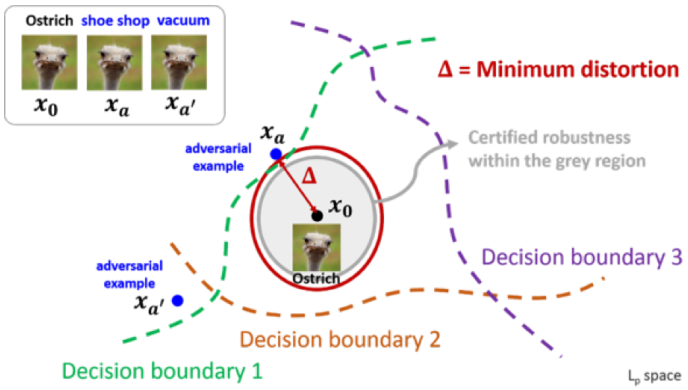

Figura 1.6. Relação entre Distorção e Amostras Adversariais [5]

No entanto, uma estratégia mais interessante ainda seria considerar não apenas as mínimas alterações necessárias para transição de classes, mas considerar as mínimas alterações necessárias para alcançar o que seria considerado uma amostra mais significativa da classe z, obtida com o afastamento desta amostra em relação à superfície de separação de classes vizinhas, através de uma perturbação p e reforço r. Neste sentido, um $\mathbf{p}$ e um $\mathbf{r}$ são adequados tal que $M(\mathbf{x}+\mathbf{p}+\mathbf{r}) = z \neq M(\mathbf{x})$, onde $\mathbf{z}$ representa uma amostra média mais representativa da classe.

Propomos a inclusão deste reforço porque amostras adversarias atendem ao requisito de transição entre classes, mas geralmente *sem* que o usuário perceba esta transição. Por exemplo, um ataque adversarial pode fazer um modelo treinado nos dígitos 4 e 9 da MNIST compreender um dígito 4 como sendo 9 apenas por colocar alguns poucos pixel na região superior do 4 que só seriam usados no 9. Em geral, um ser humano pode ter dificuldade em perceber estes poucos pixels, ainda interpretando a imagem como um 4. Isto é um problema na área de explicação pois precisamos que

[5] https://www.ibm.com/blogs/research/2018/05/clever-adversarial-attack/

o usuário perceba a transição de classes para validar uma explicação fornecida a partir deste conhecimento agregado. Ou seja, tomando por base o exemplo anterior, a quantidade de pixels na parte superior do 4 deveria ser maior para que o usuário entendesse que quando a parte superior do 4 não está "aberta", o modelo entende se tratar de um 9. Desta forma, a perturbação adequada representa o menor esforço necessário para transformar a previsão obtida na previsão esperada e o reforço aproxima ao que seria uma amostra mais representativa desta classe-alvo, facilitando a percepção do usuário sobre a sensibilidade de $\mathbf{x}$ para uma previsão $M(\mathbf{x})$. Esta é uma hipótese razoável se considerarmos que a topologia do espaço de decisão é aquela que facilita a separação entre as classes de interesse. Neste caso, o esforço em se perturbar a posição de uma amostra na direção da superfície de decisão é diretamente associada com a sensibilidade do modelo na região do espaço onde a perturbação ocorre.

1.5 Objetivo

O objetivo desta dissertação é desenvolver uma abordagem de explicação baseada em técnicas adversariais. Mais especificamente, a utilização de técnicas de ataques adversariais é explorada como evidência para determinar a sensibilidade dos atributos de uma instância para uma certa previsão. Esta sensibilidade será então usada como justificativa para a previsão realizada.

Para tanto, os seguintes objetivos específicos devem ser alcançados:

1. Elencar diferentes técnicas adversariais e, de preferência, complementares que possam fornecer evidências sobre o esforço de transição entre classes no espaço de decisão;

2. Propor um método de explicação que utilize os conceitos das técnicas adversariais e seja útil para determinar a relevância dos atributos, interpretáveis por seres humanos, para uma certa previsão de interesse. Tal método deve fornecer explicações consistentes, coerentes e consideradas adequadas por usuários;

3. Propor um método de explicação que ajuste os conceitos das técnicas adversariais para o objetivo da explicação, caracterizando suas vantagens, desvantagens e campos de aplicação adequados;

1.6 Metodologia

Nesta pesquisa estamos direcionando esforços para o estudo de explicação de amostras em modelos de RNP, tendo como entrada imagens dentro do contexto da base clássica MNIST[6]. Escolhemos o contexto de imagens porque permite a interpretação visual de resultados, algo intuitivo para seres humanos e, por isso, muito explorado em trabalhos da área.

É comum entre estratégias de explicação a obtenção de conhecimento sobre a região vizinha de uma amostra através de técnicas de perturbação e geração de amostras sintéticas. Neste trabalho utilizamos técnicas adversariais como forma de obter esta informação, pois elas tipicamente usam conhecimentos consistentes sobre transição de classes e superfícies de decisão de modelos complexos.

Para avaliar nossos resultados, comparamos os métodos propostos com outros similares usados na literatura. Como não estamos preocupados com a construção dos modelos de classificação, escolhemos um único modelo profundo para gerar todas as decisões. Então, usamos as diferentes técnicas a comparar para explicar as decisões tomadas pelo modelo. Estas explicações foram submetidas à avaliação de usuários em testes cegos onde eles determinaram quais as melhores explicações fornecidas, sem saber que técnicas forneceram que explicações.

1.7 Contribuição

Acreditamos que as principais contribuições deste trabalho para a área de técnicas de explicação de modelos de aprendizagem de máquina são:

- Condução de um estudo sobre técnicas de explicação de amostras em modelos de RNP no contexto de imagens;
- Apresentação de estratégias para sinalização de características sensíveis à mudança de classes;
- Utilização de uma base de dados altamente compreensível para validação das inferências proposta por nossa técnica de explicação;
- Combinação de conhecimento de outras áreas para validação de intuições da área de explicação e simplificação de tarefas de alguns destes algoritmos.

[6] http://yann.lecun.com/exdb/mnist/

1.8 Organização

No restante do documento está organizado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica que descreve os conceitos necessários para o entendimento dos aspectos gerais que compõem a abordagem proposta; o Capítulo 3 apresenta uma breve revisão da literatura recente explicação em aprendizagem de máquina e amostras adversariais; o Capítulo 4 detalha os métodos propostos neste trabalho; o Capítulo 5 apresenta a avaliação experimental da abordagem, bem como os resultados obtidos; e por fim, o Capítulo 6 apresenta as conclusões, limitações e trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo serão apresentados os conceitos necessários para o entendimento deste trabalho, revisitando informações sobre Modelos de Aprendizagem de Máquina, conceitos relacionados a Amostras Adversariais além das principais informações da área Explicação de Decisão em Modelos de Aprendizagem de Máquina, finalizando com as Considerações Finais revisitando a importância destes tópicos para este trabalho.

2.1 Modelos de Aprendizagem de Máquina

Modelos de aprendizagem de máquina são artefatos utilizados para aprendizado de padrões em amostras, de forma que o que foi observado seja útil em futuras decisões tomadas pelo modelo. Exemplos de tais decisões incluem enquadrar amostras em determinados grupos coesos (as chamadas classes) ou prever algum comportamento futuro (previsão do tempo). Nestes casos em particular, tarefas conhecidas como classificação e regressão temporal, são necessárias informações anteriores, dadas como exemplo por um agente externo, o supervisor, para que o modelo possa aprender os padrões e depois ser aplicado em amostras desconhecidas. Modelos que usam tais estratégias são descritos como Modelos de Aprendizagem Supervisionada. Ao longo desta dissertação vamos focar em modelos de aprendizagem supervisionada usados para classificação.

Segundo Cohen [2021], estes modelos evoluíram significativamente em termos de acurácia principalmente por terem se tornado muito complexos. Este é o caso de redes neurais. Tais arquiteturas complexas contribuem para que as decisões tomadas não sejam de fácil entendimento, sendo chamados de "Caixa-Preta" (Tabela 2.1).

Modelos simples são aqueles cuja arquitetura permite uma inferência clara sobre o comportamento do modelo e por isso são mais transparentes. No entanto, essa transparência vem ao custo de desempenhos inferiores, muitas vezes, quando compa-

ramos com o resultado de modelos mais complexos. São exemplos de modelos mais transparentes as Árvores de Decisão(Figura 2.1), a Regressão Linear e o Naïve Bayes.

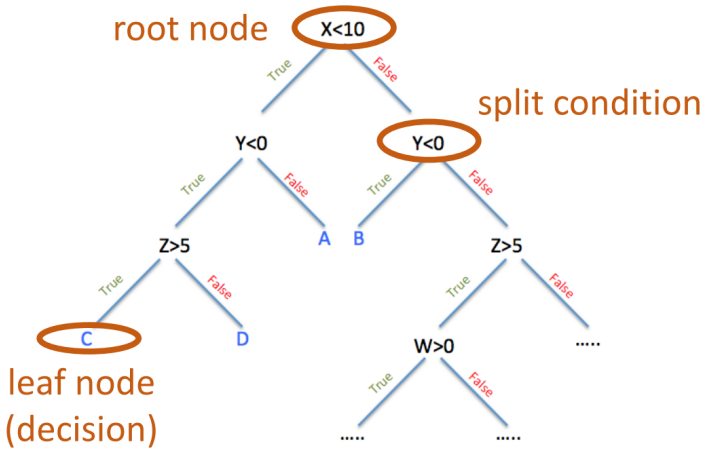

Figura 2.1. Árvore de Decisão, extraída de Yang et al. [2019a].

Modelos complexos são aqueles cuja arquitetura é mais elaborada e envolvem um plano de múltiplas dimensões para que a separação entre classes seja feita (Tabela 2.1). Segundo Guidotti et al. [2018b], muitas aplicações de sucesso, como a que envolvem em visão computacional e processamento de linguagem natural têm se beneficiado da recente evolução nos processos de aprendizado automático. O uso destes modelos é particularmente problemático em muitas aplicações em que é necessário verificar a cadeia de eventos causais envolvidos na tomada de uma decisão, de forma compreensível por seres humanos. O entendimento da "lógica" interna do processo de decisão permite a formulação de justificativas, algo intrínseco à própria definição de explicabilidade, uma vez que a alta complexidade desses modelos leva à baixa interpretabilidade dos mesmos. Nesse caso, é comum a formulação de estratégias de aproximação de modelos mais complexos através de outros mais simples que capturem o comportamento dos primeiros. Os modelos mais simples são então usados como modelos de explicação para os modelos originais.

Modelos Únicos	**Combinação de Modelos**
Rede Neural Profunda (RNP)	*Random Forests*
Rede Neural Convolucional Profunda (RNCP)	*Boosted Trees*
Support Vector Machine (SVM)	*Bagging Trees*

Tabela 2.1. Exemplos de Modelos Caixa Preta.

2.2 Amostras Adversariais

Sem perda de generalidade, seja S uma superfície de decisão que separa classes c_i e c_j. Dada uma instância I de classe c_i, é possível transformá-la em um exemplo de c_j movendo-a para o outro lado de S. De forma equivalente, a aplicação de um conjunto mínimo de pertubações na localização de I leva à criação de uma instância I' de classe c_j se a pertubação aplicada for suficiente para que I' se localize do outro lado de S.

Dada S, I' deve ser classificada como c_j mesmo sendo muito similar ao exemplo I da classe c_i. Ela é conhecida como um exemplo adversarial por ser um exemplo de difícil classificação, uma vez que consiste da amostra original perturbada o suficiente apenas para ser de outra classe [Oyallon, 2017]. No contexto de imagens, por exemplo, Yuan et al. [2019] demonstram que imagens adversariais, obtidas por meio de pequenas alterações igualmente distribuídas nas imagens originais, são difíceis de perceber por humanos. Tais alterações são conhecidas como distorção, ou seja, a diferença entre a amostra original e a adversarial (Figura 2.2).

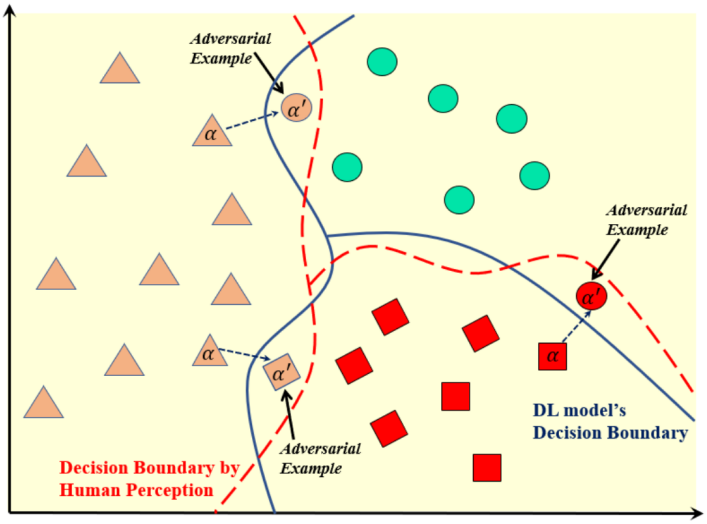

Figura 2.2. Superfície de Separação e Amostras Adversariais, extraída de Ruan et al. [2020]. Seja este um problema de 3 classes; seja a linha azul a superfície de seperação dessas classes aprendida por um modelo; seja a linha vermelha pontilhada a superfície de decisão percebida por seres humamos. α' é o que chamamos de amostra adversarial pois se utilizam da superfície de decisão aprendida para enganar o própio modelo, mas não são coerentes com as percepções humanas, além de estarem sempre próximas às superfícies de separação de classes.

As técnicas responsáveis pela geração dessas amostras são chamadas de ataques adversariais, cujo objetivo é enganar um modelo de aprendizagem de máquina, não necessariamente de forma imperceptível pelo usuário, embora isso geralmente ocorra. Enquanto algumas destas técnicas aplicam mínimas alterações nos sinais de gradientes,

outras utilizam critérios de otimização onde se busca o menor esforço necessário para gerar uma amostra adversarial, com a máxima similaridade quando comparada com uma amostra genuína [Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016; Papernot et al., 2016; Carlini & Wagner, 2017].

2.2.1 *Fast Gradient Sign Method (FGSM)*

Goodfellow et al. [2014] propuseram um ataque adversarial com foco em perturbações sobre toda a imagem. Este método usa pequenas alterações no sinal de gradiente de cada pixel de uma imagem, de forma a gerar uma máscara que pode ser imperceptível ao olho humano. Para encontrar as pertubações que constituem a máscara, basta estimar as dimensões do espaço de entrada que são mais sensíveis à mudança na classe calculando o gradiente da função de custo em relação à entrada. Quando a entrada é modificada por alterar os valores destas dimensões na direção oposta ao do gradiente, se maximiza o erro da rede. O vetor de perturbações pode ser calculado, portanto, como:

$$\eta = \epsilon \times \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$$

onde ϵ é um fator usado para definir a escala da perturbação, J é a função de custo, $\mathbf{x}$ é o vetor de entrada, θ são os pesos do modelo e y representa o rótulo. Ao sobrepor as pertubações na imagem original se obtém a imagem adversarial (Figura 2.3), caracterizada por um rótulo diferente do esperado.

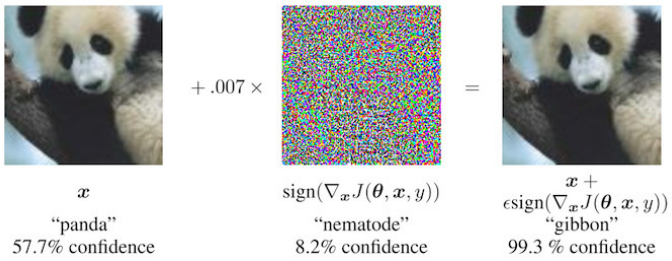

Figura 2.3. Exemplo de aplicação do FGSM, extraído de Goodfellow et al. [2014]. Na figura acima, as imagens correspondem, respectivamente a uma imgem original (rótulo *"panda"*), a máscara de perturbações e a imagem adversarial (rótulo *"gibbon"*).

Note que é possível mudar este método para que ele produza um ataque direcionado a uma classe particular por maximizar a probabilidade daquela classe específica.

2.2.2 *DeepFool*

Moosavi-Dezfooli et al. [2016] propuseram um ataque adversarial que tinha por objetivo se aproximar da superfície de separação com o mínimo de esforço possível, para obter as mínimas perturbações capazes de gerar uma amostra adversarial e por consequência, a geração de uma amostra adversarial com o mínimo de diferença de uma amostra original (Figura 2.4).

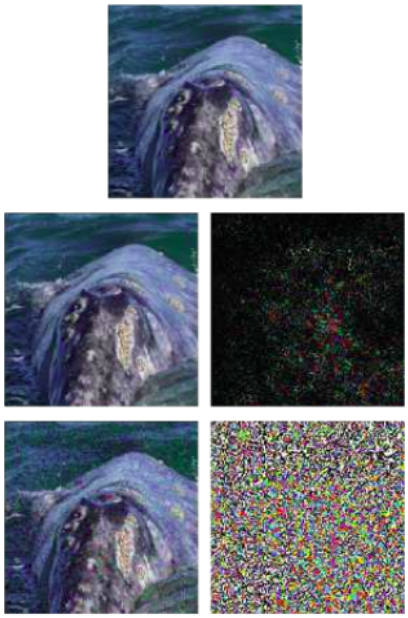

Figura 2.4. Comparação do *DeepFool* com o FGSM, extraída de Moosavi-Dezfooli et al. [2016]. Na figura acima, a primeira linha corresponde à imagem original, a segunda linha corresponde à imagem adversarial gerada pelo DeepFool e sua respectiva máscara de perturbação e a terceira linha corresponde à imagem adversarial gerada pelo FGSM e sua máscara de perturbação. Claramente, o DeepFool emprega um ataque com muito menos perturbações da imagem original.

A intuição deste método é que a robustez de um modelo M para uma instância x pode ser aproximada pela distância desta instância para o hiperplano de separação entre as classes. Ou seja, a mínima perturbação que poderia mudar a classe de x corresponde à projeção ortogonal de x no hiperplano. Logo, supondo que r é a projeção ortogonal, enquanto a classe de $x + r$ não for diferente de x, um novo r deve ser calculado até que $M(x + r) \neq M(x)$. Neste caso, $x + r$ é a imagem adversarial.

2.2.3 *Carlini & Wagner (C&W)*

Sem perda de generalidade, podemos definir uma imagem adversarial $\mathbf{x}'$ como uma imagem muito similar a uma imagem $\mathbf{x}$ (que se pretende atacar), de forma que se

$M(\mathbf{x}) = y$, então $M(\mathbf{x}') \neq y$. Carlini & Wagner [2017] discutem diversos ataques adversariais, entre os quais um baseado na tradução direta da definição anterior para o seguinte critério de otimização:

$$\begin{aligned} \min \quad & c\,\|\mathbf{x} - \mathbf{x}'\|_2^2 + J(\theta, \mathbf{x}', y) \\ \text{s.t.} \quad & \mathbf{x}' \in [0,1]^n \end{aligned}$$

onde c é um fator que determina quão importante é a similaridade entre as imagens original e adversária, J é a função de custo, $\mathbf{x}$ é imagem atacada, $\mathbf{x}'$ é imagem adversarial, θ são os pesos do modelo e y é o rótulo de $\mathbf{x}'$, distinto do de $\mathbf{x}$. Imagens adversariais geradas pelo método propostos podem ser vistas na Figura 2.5.

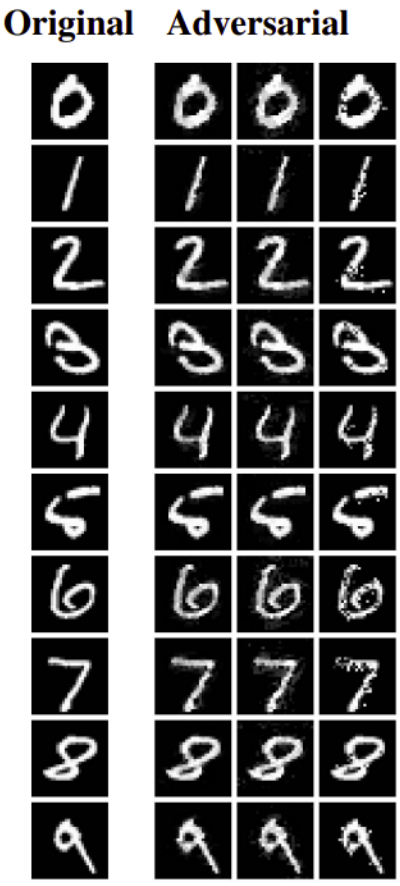

Figura 2.5. Ataque *Carlini & Wagner* (C&W), extraída de Carlini & Wagner [2017].

2.2.4 *Jacobian-Based Saliency Map Attack (JSMA)*

Papernot et al. [2016] propuseram um ataque adversarial com foco na descoberta de

uma máscara de saliência que mapeasse as áreas ou características sensíveis à alteração de classe, combinando adição ou remoção de pixels, no caso de imagens.

A intuição é que deveriam ser escolhidas as perturbações (inserção ou remoção de pontos) que produzem máximo erro de classificação. Este erro pode ser estimado pelo gradiente da função de custo (para cada classe) com respeito a todas as dimensões da entrada, ou seja, através da matriz Jacobiana. O mapa de saliência obtido é usado então para ranquear as dimensões conforme o erro de classificação que elas produzem. Os pontos são então removidos até alcançar a iteração em que se observa a transição de classe. A imagem obtida nesta iteração é a adversarial.

Note que como a imagem muda a cada perturbação, a matriz Jacobiana deve ser recalculada de forma que este método é bem mais caro que os vistos anteriormente. Contudo, ao aplicar o mapa de saliência, espera-se que um número menor de perturbações sejam necessários, o que torna a mudança empregada menos perceptível. Dada a sua natureza, este método é facilmente aplicável a ataques direcionados para certas classes. A Figura 2.6 demonstra a geração de amostras adversariais utilizando a estratégia de remoção de pixels.

2.2.5 *Maximal Jacobian-Based Saliency Map Attack (MJSMA)*

Uma variante interessante do JSMA é o método direcionado MJSMA, proposto por Wiyatno & Xu [2018]. No MJSMA, a saliência do i-ésimo atributo é calculada como:

$$S^{(i)} = -\frac{\partial h_y}{\partial X_i} \frac{\partial h_t}{\partial X_i}$$

onde h_y e h_t correspondem aos logits (valores imediatamente anteriores à camada softmax da rede neural sendo explicada) da classe real y e da classe alvo t, respectivamente. Ou seja, a importância (saliência) de um atributo é dada pela intensidade e diferença dos gradientes entre as classes real e alvo.

Embora conceitualmente interessante, o JSMA tem o inconveniente de gerar perturbações usando operações de inclusão ou remoção de pixels, mas nunca ambas. Ao contrário o MJSMA produz ataques em que operações de inclusão ou exclusão podem ser intercaladas. Isto pode ser visto na Figura 2.7 que demonstra a geração de amostras adversariais utilizando as estratégias JSMA com remoção de pixels, JSMA com inserção de pixels e MJSMA.

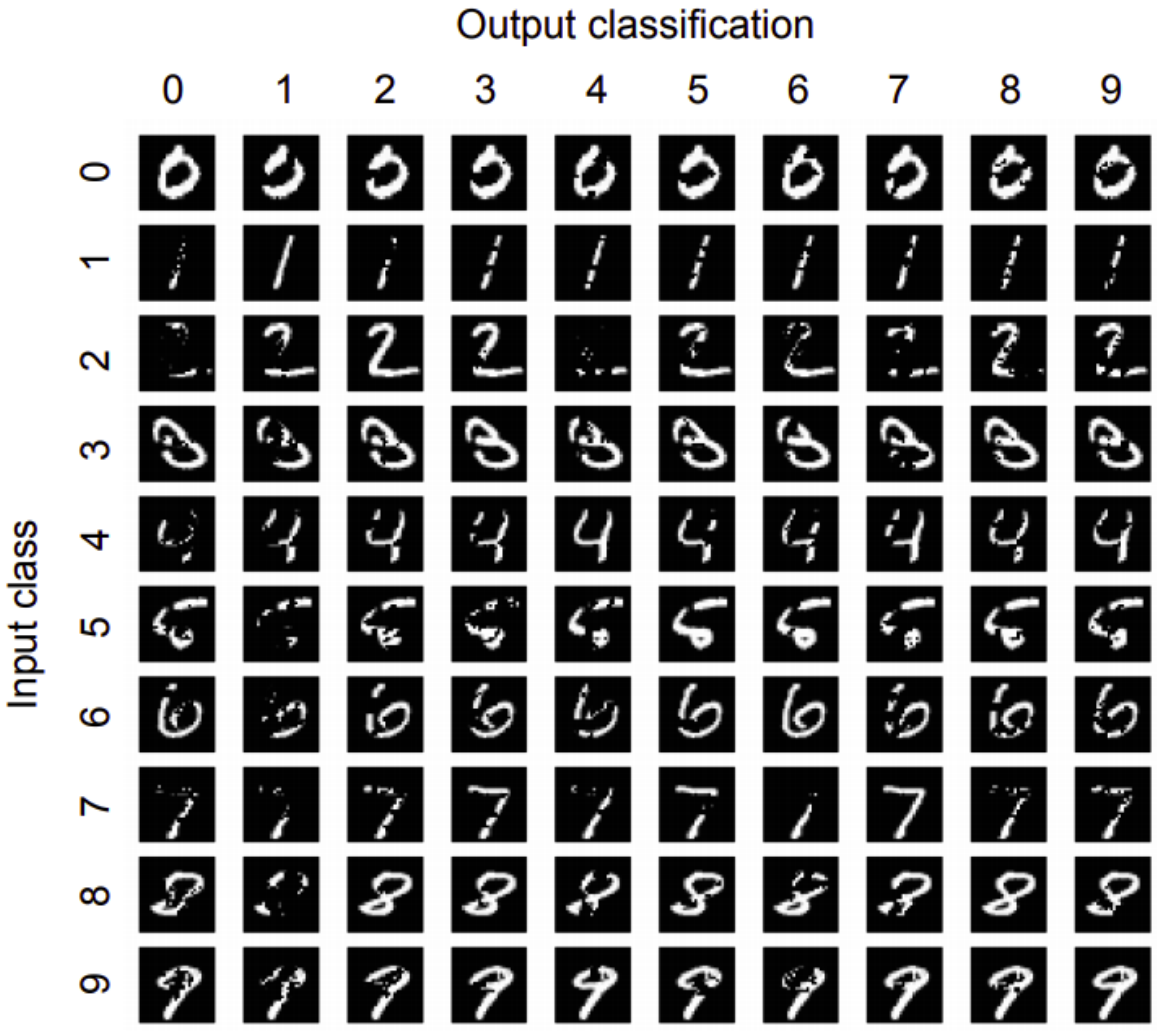

Figura 2.6. Estratégia do *Jacobian-Based Saliency Map Attack* (JSMA) aplicando remoção de pixels, extraída de Papernot et al. [2016]. Na figura acima, utilizando exemplos da base MNIST, as imagens da diagonal representam as imagens originais, as demais representam imagens adversariais geradas pelo método.

2.3 Explicação de Decisão em Modelos de Aprendizagem de Máquina

Ao aprofundar o estudo de técnicas de explicação em modelos de aprendizagem de máquina, Guidotti et al. [2018b] concluem que a natureza da experiência do usuário foco da técnica desenvolvida impacta diretamente na receptividade de um método de explicação, uma vez que os usuários de um modelo podem ter diferentes níveis de conhecimentos e experiência em uma determinada tarefa. Conhecer a experiência do usuário na tarefa que o modelo se propõe a resolver é um aspecto crucial para entender a interpretabilidade de um modelo. Dependendo da área, usuários com maior

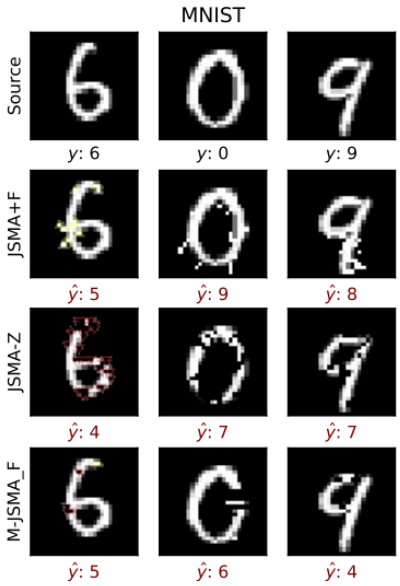

Figura 2.7. Estratégia do *Maximal Jacobian-Based Saliency Map Attack* (MJSMA), extraída de Wiyatno & Xu [2018]. Na figura acima, utilizando exemplos da base MNIST, as imagens da primeira linha representam imagens originais; as imagens da segunda linha representam amostras adversariais geradas pelo JSMA com estratégia de adição de pixels; as imagens da terceira linha representam amostras adversariais geradas pelo JSMA com estratégia de remoção de pixels; e a quarta linha apresenta amostras adversariais geradas pelo MJSMA, combinando as estratégias de adição e remoção de pixels.

experiência podem preferir uma técnica de explicação mais ou menos complexa, uma vez que diferentes tipos de dados, originados de diversas áreas, apresentam um nível de interpretabilidade igualmente variável para o ser humano.

Além da percepção humana, o desenvolvimento de novas técnicas de explicação requer atenção para outros aspectos importantes como (a) clareza do tipo de explicação fornecida, (b) se alguma técnica de reforço de informação vai ser necessária, (c) qual o tipo de explicador vai ser entregue e (d) como esse explicador será avaliado.

2.3.1 Aspectos das Explicações

Ribeiro et al. [2016] orientam que o entendimento da relação entre os elementos de entrada e a saída de um modelo é característica crucial para que uma explicação seja interpretável. Em aplicação de imagens, por exemplo, o nível de interpretabilidade de uma explicação pode ser medido através da avaliação sobre a sinalização de áreas de uma imagem que melhor contribuem para a predição de uma determinada classe desta imagem; essas sinalizações em imagens também podem ser chamadas de máscaras de

interpretação [Fong & Vedaldi, 2017].

Igualmente importante é a representação da explicação sob a qual será avaliada a interpretabilidade, como citam Schneider & Handali [2019]; Guidotti et al. [2018b]; Ribeiro et al. [2016]. Segundo estes últimos, explicar uma predição é a apresentação textual ou visualização de artefatos que fornece um entendimento do relacionamento entre os componentes da instância e sua predição, seja através de palavras-chaves ou partes de uma imagem, por exemplo.

2.3.1.1 Características de uma Explicação

A explicação de decisão em modelos de aprendizagem de máquina pode ser caracterizada pela presença de Justiça, Confiança, Fidelidade, Acurácia e Personalização.

Justiça Para Schneider & Handali [2019], justiça em explicação é a característica das explicações que possuem a mesma qualidade (fidelidade, interpretabilidade, generalização e esforço) para cada indivíduo.

Confiança Para Ribeiro et al. [2016], confiança é a característica garantida por uma confiança depositada no modelo, bem como em suas predições individuais que endossam a tomada de decisão em todos, ou na grande maioria, dos seus casos aplicáveis.

Fidelidade Para Fong & Vedaldi [2017], fidelidade pode ser considerada como a medida do erro de predição esperado por um modelo, isto é, a medida da capacidade de um modelo de explicação em imitar um modelo caixa-preta, podendo ser percebida através das medidas de precisão, revocação e F1 [Guidotti et al., 2018b], em função da saída de um modelo caixa-preta. Schneider & Handali [2019] sintetizam a fidelidade como o grau de mapeamento obtido entre entrada e saída. Ribeiro et al. [2016] especificam a importância da Fidelidade Local como sendo a capacidade de um explicador representar localmente a explicação de forma coerente com a representação do modelo original.

Acurácia Para Guidotti et al. [2018b], acurácia é a relação de predição do modelo para amostras desconhecidas, cuja avaliação pode ser feita pela medida de acurácia e F1.

Personalização Em Aprendizagem de Máquina, a personalização tem sido apontada como uma tendência recente e diz respeito a levar em conta as interações do usuário

para disponibilizar a uma explicação personalizada. É possível encontrar a personalização como um fim, tal como em sistemas de recomendação, personalização em pesquisa web, etc. Mas também é possível estudar a personalização da própria ferramenta de aprendizagem de máquina para resolver um determinado problema, como em aprendizagem de máquina iterativa [Schneider & Handali, 2019].

2.3.1.2 Propósito de uma Explicação

Antes do desenvolvimento de uma técnica de explicação em modelos de aprendizagem de máquina, após tomar nota dos requisitos de uma explicação, é preciso atentar para o objetivo da explicação, podendo ser Explicação de Modelo, Explicação de Amostra, Inspeção de Modelo e Construção de um Modelo Transparente.

Explicação de Modelo Consiste em entender a lógica geral do modelo caixa-preta, para interpretação global através de um **Explicador Global** [Guidotti et al., 2018b; Jia et al., 2019]. São exemplos de explicadores de globais: *Testing with Concept Activation Vectors (TCAV)* [Kim et al., 2018] e *Rule-Constrained Networks (RCN)* [Okajima & Sadamasa, 2019].

Explicação de Amostra Busca a correlação entre os dados de um registro de entrada e a decisão tomada por um modelo caixa-preta, caracterizando interpretação local, através de um **Explicador Local** [Guidotti et al., 2018b; Jia et al., 2019]. Para Fong & Vedaldi [2017], uma razão para o aumento do estudo desse problema são as especificidades de instâncias de modelos complexos que podem ser inconsistentes com uma explicação global, além do fato que modelos complexos podem caracterizar, em sua solução, um plano de múltiplas dimensões. São exemplos de explicadores locais o *Local Rule-Based Explanations (LORE)* [Guidotti et al., 2018a], o *Local Interpretable Model-Agnostic Explanations (LIME)* [Ribeiro et al., 2016] e o *Anchors* [Ribeiro et al., 2018a].

Inspeção de Modelo Para Guidotti et al. [2018b], esta variação combina as duas abordagens anteriores e está relacionada à visualização de aspectos do modelo, seja sob uma perspectiva global ou local. Exemplos incluem a *Sensitive Analysis (SA)* e *Layer-wise Relevance Propagation (LRP)* [Samek et al., 2017], sinalizando a sensibilidade de um atributo ou identificando neurônios responsáveis por decisões específicas.

Construção de um Modelo Transparente Esta abordagem tem por objetivo a construção de um modelo inicial já transparente, a partir do qual é possível entender seus

mecanismos de tomada de decisão sem o auxílio de um modelo substituto. Sendo assim, a partir dessa abordagem, é possível obter interpretabilidade em nível local ou global. Exemplos de Construção de um Modelo Transparente incluem técnicas baseadas em modelos transparentes como a extração de regras de associação (*Rule-Constrained Networks (RCN)* [Okajima & Sadamasa, 2019]) ou seleção de amostras representativas (protótipos) para serem usadas como exemplos.

2.3.2 Reforço de Informação e Perturbações

Quando uma técnica de explicação está focada em Explicação de Amostra, considerando que estas técnicas costumam dar atenção principalmente para as regiões próximas a amostra analisada, é comum constatar a escassez de amostras vizinhas conhecidas ou limitação de informação sobre a vizinhança de uma amostra analisada. Sendo assim, estas técnicas adotam meios para o enriquecimento de informações através da geração de amostras sintéticas, inseridas no contexto do modelo, nas proximidades de uma amostra analisada para geração de uma explicação (Figura 2.8).

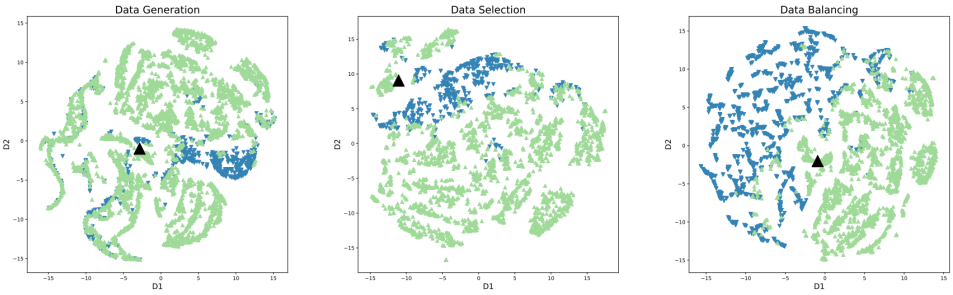

Figura 2.8. Processo de Construção de Vizinhança adotado na técnica *EXPLaining black-box classifiers using Adaptive Neighborhood generation* (EXPLAN), extraído de Rasouli & Yu [2020], onde há a geração e balanceamento da população de amostras vizinhas a uma amostra-alvo, como parte de seu processo de fornecimento de explicação para tal amostra específica.

Por conta disso, técnicas de explicação com foco em amostras, precisam visitar este tópico e ponderar a necessidade de enriquecimento de informação nas proximidades de uma amostra alvo de explicação, uma vez que trabalhos recentes têm atribuído a qualidade da explicação à qualidade deste conhecimento sobre a vizinhança de uma amostra alvo de explicação [Ribeiro et al., 2016; Guidotti et al., 2018b; Ribeiro et al., 2018b]. Esse enriquecimento de informação pode ser feito através de alguns métodos de perturbação, tais como a Perturbação Randômica (PRD), a Perturbação Baseada em Vizinhança (PBV), a Perturbação Baseada em Gradiente (PBG) e a Perturbação por Algoritmo Genético (PAG).

Perturbação Randômica (PRD) Nesta perturbação, as amostras são geradas aleatoriamente dentro de uma distribuição normal centrada na amostra de teste [Jia et al., 2019]. Ribeiro et al. [2016, 2018a] adotam essa estratégia nas técnicas de explicação LIME e *Anchors*.

Perturbação Baseada em Vizinhança (PBV) Esta perturbação acontece dentro de um raio cujo limite é determinado pela amostra adversarial mais próxima, onde todas as amostras inseridas neste raio são perturbadas [Jia et al., 2019] de modo a aumentar o conhecimento desta região em um modelo.

Perturbação Baseada em Gradiente (PBG) Neste método, apenas uma amostra perturbada é gerada, geralmente em direção a um local ótimo Jia et al. [2019]. Normalmente aplicado a perturbação em imagens, a exemplo da técnica de geração de amostras adversariais *FGSM* [Pang et al., 2018].

Perturbação por Algoritmo Genético (PAG) Adota algoritmos genéticos para simular a população próxima a uma instância. Essa estratégia é adotada para evidenciar a superfície de separação [Guidotti et al., 2018a] e compor a solução da técnica de explicação LORE, por exemplo.

2.3.3 Aspectos dos Explicadores

Com os pré-requisitos alinhados a respeito das características e propósito da explicação, é possível aprofundar os conhecimentos sobre os tipos de explicadores para depois trabalhar no desenvolvimento de uma técnica própria. Em termos gerais, o processo de explicação de modelos de aprendizagem de máquina pode ser visto na Figura 2.9 e para alcançar o resultado esperado precisa levar em conta os seguintes aspectos: Estratégia do Explicador, a Natureza do Explicador e o Tipo do Explicador.

2.3.3.1 Estratégia do Explicador

A estratégia do explicador consiste na escolha da base da técnica de explicação desenvolvida. Guidotti et al. [2018a] indicam que as principais estratégias de uma técnica de explicação são Engenharia Reversa (ER) e Projeção de Explicação (PE).

Engenharia Reversa (ER) É a técnica que usa um modelo auxiliar para simular o comportamento do modelo de aprendizagem de máquina com menor complexidade e

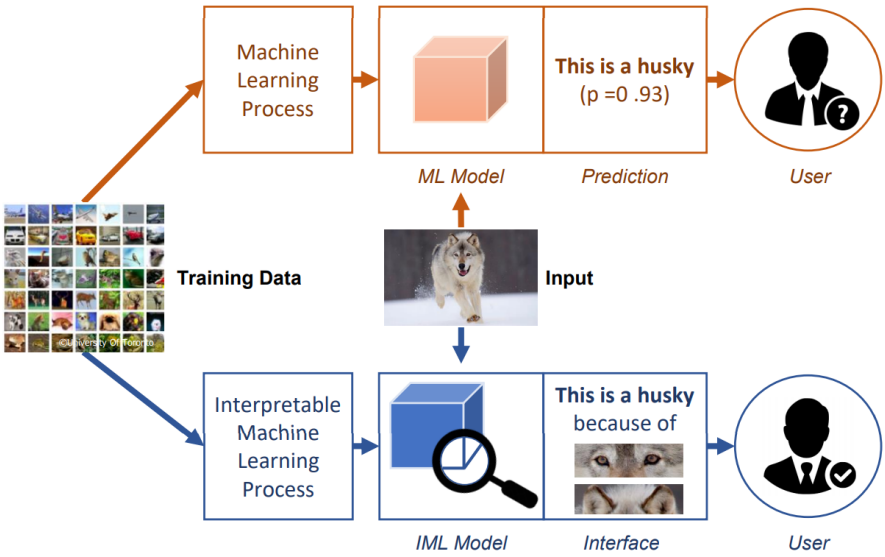

Figura 2.9. Fluxo de uma Técnica de Explicação, extraído de Yang et al. [2019b]. Nesta imagem, há a separação de dois processos que envolvem a área de explicação. O primeiro processo, demonstrado na parte superior da imagem, é a criação de conhecimento sobre um modelo de aprendizagem de máquina e suas predições informadas; no exemplo da imagem, a amostra de entrada foi dita "husky"com 93% de certeza. O segundo processo, demonstrado na parte superior da imagem, é a investigação e busca por artefatos que sinalizem ou justifiquem os motivos pelos quais o modelo do passo anterior tomou tal decisão; neste caso, a mesma amostra de entrada foi fornecida a uma técnica de explicação que retornou as evidências que sustentam a decisão do modelo conhecido.

maior transparência. Está diretamente ligado aos seguintes propósito de explicação: Explicação de Modelo e Explicação de Amostra

Projeção de Explicação (PE) É a técnica que visa utilizar o próprio modelo para inspecionar e apurar as ativações internas de um modelo de aprendizagem de máquina. Uma forma de atender esta estratégia é trabalhar na construção de um modelo complexo de alta acurácia e alta transparência a exemplo da Construção de um Modelo Transparente. Outra forma de também atender esta estratégia é investigar a sensibilidade de um modelo de aprendizagem de máquina através da Inspeção de Modelo.

2.3.3.2 Natureza do Explicador

A observação da literatura permite a classificação dos explicadores em Explicador Agnóstico (AGN) e Explicador Orientado a Modelo (ORM).

Explicador Orientado a Modelo (ORM) Para Pastor & Baralis [2019], representa o grupo de explicadores que procuram explicar modelos caixa-preta específicos como o DeapLIFT [Shrikumar et al., 2017] e Grad-CAM [Selvaraju et al., 2017], que são focados em RNP, por exemplo.

Explicador Agnóstico (AGN) Para Ribeiro et al. [2016], é um explicador que possui a habilidade de lidar com qualquer tipo de modelo e ainda assim fornecer explicação. Para Guidotti et al. [2018b], é um preditor compreensível que não está limitado a uma solução específica, podendo suportar como entrada dados Tabular (TAB) (numérico, categórico ou lógico), Imagem (IMG), Vídeo (VD) e Texto (TXT) (detecção de spam ou classificação de tópico).

2.3.3.3 Tipo do Explicador

Na revisão de literatura feita por Guidotti et al. [2018b], são apontados como possíveis explicadores as Árvores de Decisão, o Explicador Baseado em Regras (EBR), a Importância de Características (IMC), a Máscara de Saliência (MDS), a *Sensitive Analysis* (SA), a Plotagem de Dependência Parcial, a Seleção de Protótipo e a Maximização de Ativação.

Árvores de Decisão Um explicador deste tipo pode ser enquadrado um explicador global ou local, representado por uma árvore complexa ou simples como demonstração da explicação (Figura 2.1).

Explicador Baseado em Regras (EBR) Um explicador deste tipo utiliza regras de decisão por serem amigáveis ao entendimento do ser humano, facilitando a aceitação e entendimento de uma explicação. Utilizam Explicador Baseado em Regras (EBR) as seguintes técnicas: *Anchors* [Ribeiro et al., 2018a], *Local Rule-Based Explanations* (LORE) [Guidotti et al., 2018a], *EXPLaining black-box classifiers using Adaptive Neighborhood generation* (EXPLAN) [Rasouli & Yu, 2020].

Importância de Características (IMC) É uma técnica de explicação local (ou global) que consiste em fornecer pesos às características utilizadas por um modelo de modo semelhante a coeficiente em modelos lineares usados como modelos de interpretação, sendo assim é possível perceber o impacto de uma característica (ou um grupo de características) na predição de uma amostra. Lundberg & Lee [2017] indicam que boas técnicas de explicação deste tipo precisam apresentar precisão local, perda e consistência de acordo com a Tabela 2.2. Utilizam Importância de Características (IMC)

as seguintes técnicas: *Local Interpretable Model-Agnostic Explanations* (LIME) [Ribeiro et al., 2016], *SHapley Additive exPlanations* (SHAP) [Lundberg & Lee, 2017] e *Adversarial* [Marino et al., 2018].

Máscara de Saliência (MDS) Considerada uma visualização da técninca anterior, é utilizada para problemas envolvendo imagens, buscando destacar os pontos importantes em uma imagem que possuem alta contribuição em uma predição. A técnica *Meaningful Perturbation* [Fong & Vedaldi, 2017] utiliza Máscara de Saliência (MDS) como forma de explicação.

***Sensitive Analysis* (SA)** Esta técnica avalia a incerteza na saída a partir da busca de fontes de incerteza na entrada, destacando pontos importantes para predição e sinalizando como as partes de uma entrada impactam na predição de uma amostra. Geralmente, é utilizada para desenvolver técnicas de inspeção de modelo [Samek et al., 2017].

Plotagem da Dependência Parcial Esta técnica é orientada a visualização e entendimento do relacionamento da saída com a entrada em um espaço de características reduzidas, sendo possível analisar como as variações de uma característica afetam a predição do modelo, é um tipo específico da técnica anterior.

Maximização de Ativação Esta técnica é um desdobramento da técnica *Sensitive Analysis* (SA) e visa, dentro dos pontos sensíveis, evidenciar aqueles cuja contribuição é significativa, melhorando a percepção de uma explicação. A técnica *Layer-wise Relevance Propagation* (LRP) [Samek et al., 2017] utiliza esta forma de inspeção para sinalizar o que impacta significativamente na predição de uma amostra.

Seleção de Protótipo Esta técnica consiste no retorno de amostras representativas junto com a saída do classificador para embasar decisão do classificador. Um desdobramento da técnica *Testing with Concept Activation Vectors* (TCAV) [Kim et al., 2018] permite a utilização de conceitos de alto nível para seleção de protótipos como forma de explicação de um modelo.

2.3.4 Avaliação de Técnicas de Explicação

Doshi-Velez & Kim [2017]; Biran & Cotton [2017] organizaram as formas de avaliação de técnicas de explicação de modelos em aprendizagem de máquina sob três aspec-

Característica	Descrição
Precisão Local	É a capacidade de um modelo local conseguir representar amostras locais tais quais são representadas em seu modelo original.
Perda	Dada a existência de um grupo simples de características para o fornecimento de uma explicação, a perda de uma dessas características terá impacto direto na explicação.
Consistência	Indica que uma vez selecionadas as características relevantes, caso o modelo sofra alterações, o conjunto de características ou permanecerá o mesmo, ou aumentará, de acordo com sua contribuição.

Tabela 2.2. Características comuns a Métodos de Atribuição de Características Aditivas utilizados em modelos de explicação recentes.

tos: Baseada na Funcionalidade, Baseada no Conhecimento Humano ou Baseada na Aplicação (Figura 2.10).

Baseada na Funcionalidade Com baixo custo de implementação, essa avaliação não exige participação humana, utilizando-se de definições de interpretabilidade como representação da qualidade da explicação. Esse tipo de avaliação é apropriada quando o problema avaliado possui classes de modelos validadas por humanos anteriormente, quando um método ainda não estiver maduro suficiente ou quando experimentos com seres humanos forem anti-éticos. As representações de interpretabilidade podem ser obtidas através da utilização de árvores de decisão, cujas regras tornam tal modelo altamente interpretável, uma vez que as regras descritas ao longo de uma árvore de decisão possuem alta receptividade por seres humanos, o que contribui para que estes através de tais regras descritas tenham um entendimento transparente sobre o processo decisório deste modelo. Desta forma, utilizar uma árvore de decisão para realizar tarefas parecidas a de um modelo anterior previamente conhecido, pode agregar no entendimento do processo decisório deste.

Baseada no Conhecimento Humano Com médio custo de implementação, essa avaliação exige a participação humana, utilizando-se de tarefas simples que corroborem o entendimentos simples do contexto do problema estudado, por não exigir a participação de especialistas. Esse tipo de avaliação é apropriada quando se objetiva testar, por exemplo, que tipos de explicação são melhores compreendidas, dadas restrições de tempo. Um exemplo de avaliação Baseada no Conhecimento Humano é a Escolha

Binária Forçada, onde são apresentados para seres humanos duas explicações e eles devem indicar aquela que julgam ser a melhor.

Baseada na Aplicação Com alto custo de implementação, essa avaliação exige a participação humana, utilizando-se de tarefas reais, como o auxílio de médicos na realização de diagnóstico de doenças específicas. Dessa forma através dessa abordagem é dada a oportunidade de análise de desempenho de explicadores na identificação de erros ou novos fatos.

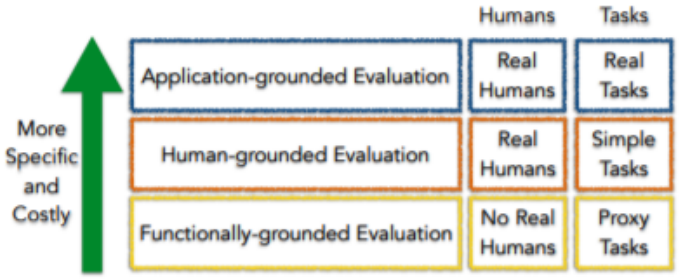

Figura 2.10. Representação dos tipos de avaliação de explicação, de acordo com o grau de complexidade e envolvimento do ser humano, extraída de Doshi-Velez & Kim [2017]

Lage et al. [2019] abordaram aspectos práticos sobre a avaliação de técnicas de explicação com a validação de usuários. Eles investigaram o método Explicador Baseado em Regras (EBR), estudando o efeito das regras interpretáveis para os usuários dependendo da interação desses usuários com os modelos.

Uma interação entre usuário e modelo pode ser descrita como Verificação, Simulação ou Contrafactual. No primeiro caso, há a comparação de comportamento de um modelo com o comportamento de usuário. Por exemplo, dadas as recomendações dadas por um modelo, verifica-se a consistência destas comparadas às recomendações dadas por um usuário. No segundo caso, dada uma explicação, tenta-se adivinhar o comportamento do modelo. Já o último caso acontece quando são analisadas as mudanças de opinião dos modelos a partir de alterações indicadas nas explicações, evidenciando a sensibilidade entre amostras e classes de um modelo (Figura 2.11).

Na análise do Explicador Baseado em Regras (EBR) feita por Lage et al. [2019] foram apontados aspectos que impactavam na facilidade de entendimento por humanos: tamanho da explicação, variação de blocos cognitivos (Figura 2.12) das explicações e utilização de termos repetidos, no bloco magenta.

Sendo assim, considerada a relação usuário, método de explicação e variação da explicação, Lage et al. [2019] concluíram que técnicas de explicação poderiam ser

Figura 2.11. Representação de Avaliação de Atividade Contrafactual, extraída de Lage et al. [2019]. Nesta atividade, os usuários devem indicar se as observações no quadro magenta alterariam as preferências do 'alien'.

avaliadas em relação ao tempo de responsta para execução de uma determinada tarefa, acurácia das explicações ou satisfação do usuário deste modelo.

2.4 Considerações Finais

Neste capítulos visitamos conceitos necessários para o entendimento deste trabalho, passando por conceitos sobre modelos de aprendizagem de máquina, superfície de separação e técnicas de explicação.

Alguns aspectos sobre modelos de aprendizagem de máquina foram levantados, como a sua evolução arquitetural e elevada acurácia por causa da viabilização de modelos cada vez mais complexos, chamados de modelos caixa-preta. Tais modelos evidenciam a relação antagônica entre complexidade e transparência, uma vez que modelos simples costumam possuir alta transparência e baixa acurácia quando comparados com modelos complexos de baixa transparência e alta acurácia. Isto torna particularmente desafiador o fornecimento de explicações e/ou justificativas para as decisões desses modelos.

Quando os aspectos de complexidade de modelos de aprendizagem de máquina são levantados, é importante salientar que a complexidade de um modelo não está

Figura 2.12. Representação de Blocos Cognitivos, extraída de Lage et al. [2019]. Nesta atividade, os blocos cognitivos são sinalizados pela explícita e coesa separação de informações sobre as observações das informações sobre as recomendações.

apenas na complexidade arquitetural a exemplo de RNP e Rede Neural Convolucional Profunda (RNCP), mas também no plano de separação de classes desta solução que envolvem múltiplas dimensões. Sabendo disso, é importante notar que técnicas adversariais detêm um conhecimento que pode agregar significativamente para a área de técnicas de explicação por conta da experiência no aspecto da transição entre classes e conhecimento sólido da superfície de separação de classes de modelos complexos. Inclusive Marino et al. [2018] apresentaram orientações para a combinação desse conhecimento na área de explicação.

No que diz respeito a técnicas de explicação em modelos de aprendizagem de máquina, revisitamos os conceitos comuns encontrados na área, além de pontos relevantes no desenvolvimento de técnicas de explicação, respondendo às seguintes questões: (a) O que é uma explicação? (b) Quais os objetivos de uma explicação? (c) É necessário algum tipo de enriquecimento de uma área de uma amostra para geração de uma explicação? (d) O que é um explicador? (e) Qual a estratégia do explicador? (f) Qual é a sua natureza? (g) Como este explicador representará uma explicação?

A resposta destas perguntas dá um direcionamento consistente para o desenvolvimento de uma técnica nesse sentido. A avaliação destas técnicas depende do tipo

de informação que se deseja obter, podendo abranger aspectos da funcionalidade da técnica, aceitação dos usuários ou mesmo coerência diante de cenários reais. Tendências desta área têm mostrado o direcionamento para a avaliação da receptividade dos usuários diante destas técnicas de explicação por serem os maiores clientes e maiores validadores das técnicas de explicação, focando, portanto, nos dois últimos tipos de avaliação.

Sendo assim, queremos neste trabalho estabelecer relações entre esses conceitos, percepções e intuições sobre modelos de aprendizagem de máquina, superfície de separação e transição de classes através do desenvolvimento, implementação e experimentação de uma técnica de explicação de decisão em modelos de aprendizagem de máquina, validando-a junto a usuários.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentados alguns dos trabalhos desenvolvidos na área de explicação de modelos de aprendizagem de máquina, a partir de 2016. Os trabalhos foram divididos em Métodos de Explicação com Propósito Global e em Métodos de Explicação com Propósito Local (Figura 3.1[1]). Este último, foi organizado de acordo com os tipos de explicações fornecidas, podendo apresentar Explicações baseadas em Máscara de Saliência, Explicações baseadas em Regras de Decisão ou Explicações baseadas em Importância de Características. Depois disso é apresentada uma Síntese dos Trabalho Relacionados, seguida das Considerações Finais do capítulo.

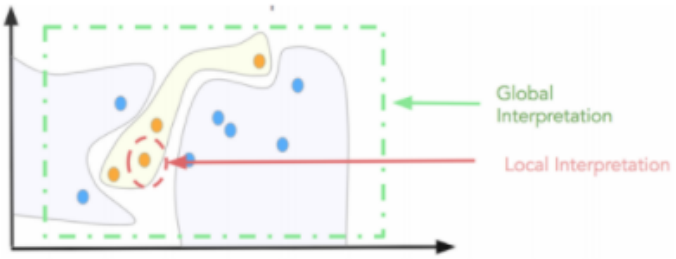

Figura 3.1. Explicação Global versus Explicação Local. Nesta imagem, a explicação local da amostra laranja destacada em círculo pontilhado é dada pelas superfícies de separação próximas a esta amostra, enquanto a explicação global da classe de amostras da cor laranja é dada a partir de uma generalização de suas amostras.

[1] https://www.kdnuggets.com/2018/06/human-interpretable-machine-learning-need-importance-model-interpretation.html

3.1 Métodos de Explicação com Propósito Global

Os métodos de explicação com propósito global têm o objetivo de explicar o modelo, de forma genérica, independente da amostra. Estes métodos podem ser úteis para obter inferências sobre o comportamento do modelo, mas em certas amostras, ou mesmo *outliers*, a utilização de tais métodos podem encontrar um resultado inesperado. A seguir, descrevemos alguns trabalhos recentes com propósito global.

Kim et al. [2018] apresentaram o *Testing with Concept Activation Vectors (TCAV)*, um método de explicação pra RNP, usando Engenharia Reversa (ER) para gerar explicações para imagens. Dada uma RNP, são repassados a esta rede dois conjuntos de conceitos de alto nível compreensíveis por seres humanos e verificados os sinais que esta rede ativa para estes conceitos. Estes sinais conhecidos, quando colocados em uma superfície de separação, permitem o conhecimento dos *Concept Activation Vectors* (CAV) e, a partir disso, é possível fornecer uma explicação para uma amostra, em função destes conceitos de alto nível fornecidos no início. A Figura 3.2 demonstra os conceitos do TCAV para amostras da classe zebra; o conhecimento do impacto dos CAV em uma RNP permitem também a explicação através da seleção de amostras, como na Figura 3.3.

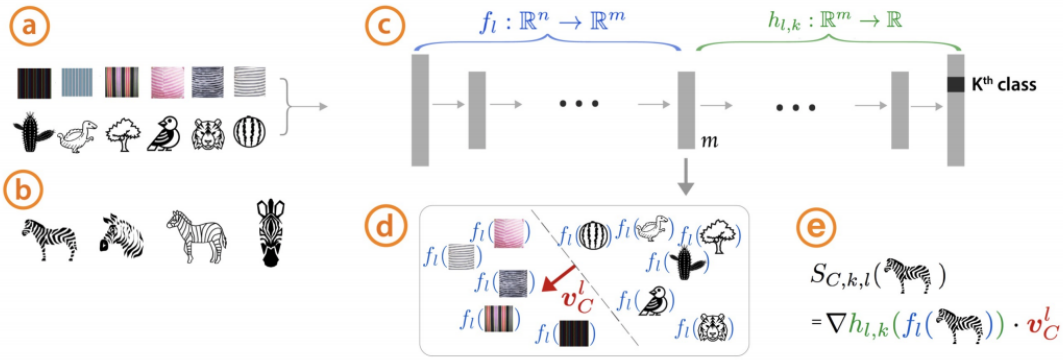

Figura 3.2. Representação da Intuição do TCAV, extraída de Kim et al. [2018]. Dado um conjunto de exemplos definidos pelo usuário para um conceito (por exemplo, 'listras') e exemplos aleatórios [a], exemplos de dados de treinamento rotulados para a classe estudada [b] uma rede treinada [c], o TCAV pode avaliar a sensibilidade do modelo ao conceito dessa classe. Os CAVs são aprendidos pelo treinamento de um classificador linear para fazer a separação entre as ativações produzidas pelos exemplos aleatórios e os exemplos de conceito em qualquer camada [d]. O CAV é o vetor ortogonal ao limite de classificação, representado por uma seta vermelha. Para a classe estudada (zebras), TCAV usa a derivada direcional S para representar a sensibilidade ao conceito [e].

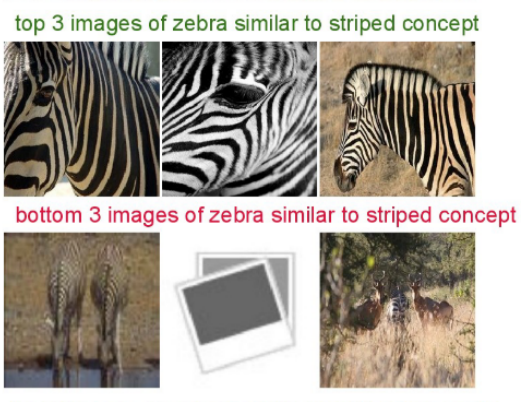

Figura 3.3. A aplicação dos conhecimentos aprendidos é representada, nesta figura extraída de Kim et al. [2018], depois de aprendido o conceito de alto nível "listras" e sua relação com a classe zebra. A primeira linha representa as amostras mais parecidas com o conceito "listras em zebra" e a segunda linha, as imagens menos parecidas.

Uma outra forma de construir um modelo complexo e de decisões explicáveis é fazer um modelo transparante desde a sua construção, evitando as complexas abstrações feitas pelas RNP, como é o caso do *Rule-Constrained Networks (RCN)* proposto por Okajima & Sadamasa [2019]. Neste método, os autores constroem uma RNP de regras de decisão, combinando sua alta explicabilidade à alta acurácia das RNP. O RCN é um método de explicação baseado na Projeção de Modelo Transparente (PMT), pois ele gera uma Projeção de Explicação (PE) para dados de tabelas. Dado um conjunto de regras de decisão e as probabilidades das classes de um modelo (Figura 3.4), uma rede neural escolhe as melhores regras deste conjunto inicial. Desta forma, ao invés de prever uma classe, a rede neural prevê uma regra que satisfaça as informações iniciais e forneça a classe correta com máxima probabilidade.

Rule	**Antecedent (Condition)**	**Consequent (Class probability)**
r_0	$f_0 =$ "A"	$[0.2, 0.8]$
r_1	$f_0 =$ "B"	$[0.8, 0.2]$
r_2	$f_1 > 2$	$[0.3, 0.7]$
r_3	$f_1 > 2.5$	$[0.2, 0.8]$
r_4	$f_0 =$ "A" AND $f_1 > 2$	$[0.1, 0.9]$
r_5	$f_0 =$ "B" AND $f_1 \leq 2$	$[0.9, 0.1]$

Figura 3.4. Representação de Conjunto de Regras Fornecido para Método RCN, extraída de Okajima & Sadamasa [2019]. Esse exemplo de regras de decisão representa a permutação dos valores de duas características impactando a probabilidade de tal regra representar uma classe ou outra em um problema de classificação de duas classes.

Ambos os trabalhos trazem contribuições importantes e discussões válidas para

essa área de conhecimento em crescimento. Ao focar na clareza de suas explicações, os autores deixam em evidência a importância do entendimento do usuário desses métodos, se comparado ao entendimento interno do modelo, ao aproximar abstrações de alto nível e utilização de regras de decisão como forma de demonstrar explicações.

3.2 Métodos de Explicação com Propósito Local

A generalização do entendimento das classes em Métodos de Explicação com Propósito Global pode ser importante para entendimento genérico de um modelo, mas quando é necessária uma análise mais detalhada de amostras especiais que fujam deste padrão genérico de uma classe de um determinado modelo, são necessários métodos de explicação que analisem sua vizinhança e especificidades. Tais métodos podem ponderar sobre o que torna uma amostra especial dentro do plano no qual o modelo está inserido. Seguindo esta linha, foram desenvolvidos Métodos de Explicação com Propósito Local, aqui nesta seção distribuídos de acordo com a forma que mostram suas explicações: Explicações baseadas em Máscara de Saliência, Explicações baseadas em Regras de Decisão ou Explicações baseadas em Importância de Características.

3.2.1 Explicações baseadas em Máscara de Saliência

Usadas exclusivamente com imagens, máscaras de saliência evidenciam pontos de uma imagem (pixels) que justificam determinada classificação. Fong & Vedaldi [2017] propuseram um método próprio de detecção de Máscara de Saliência (MDS), usando Engenharia Reversa (ER), combinando Perturbação Baseada em Gradiente (PBG). Dada uma amostra inicial, o trabalho usa PBG para detectar as mínimas informações desta amostra necessárias para justificar a classe que o modelo classificou. A Figura 3.5 mostra o esquema de aprendizado da máscara, enquanto a Figura 3.6 mostra como essa máscara é usada para explicação.

Samek et al. [2017] apresentaram como métodos de explicação o *Sensitive Analysis (SA)* e o *Layer-wise Relevance Propagation (LRP)*. Ambos os métodos usam RNP e buscam, a partir de ER, evidenciar justificativas de predição para dados de imagens, textos e vídeos. A análise de sensibilidade é usada para estimar quão sensível é a previsão em relação a alterações na amostra de entrada, enquanto a propagação de relevância em camadas decompõe a decisão em função da amostra de entrada. A Figura 3.7 demonstra esta forma de explicação no contexto de classificação de imagem.

Khalifa et al. [2018] apresentaram o *Deep Visual Explanation (DVE)*, um *framework* orientado a Rede Neural Profunda (RNP) para justificativa de predições, a

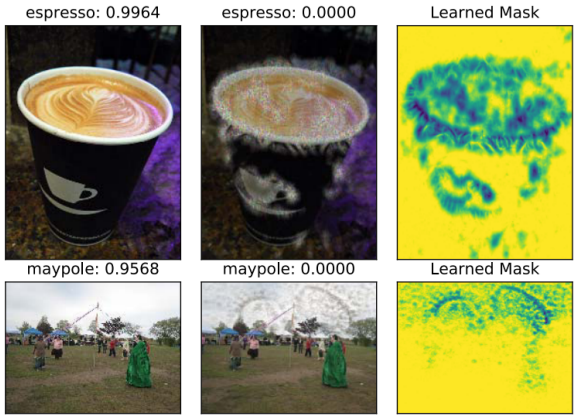

Figura 3.5. Representação do Esquema de Detecção de uma Máscara de Saliência (MDS), extraída de Fong & Vedaldi [2017]. A primeira linha representa a detecção da classe "expresso" e a segunda linha representa a detecção da classe "mastro".

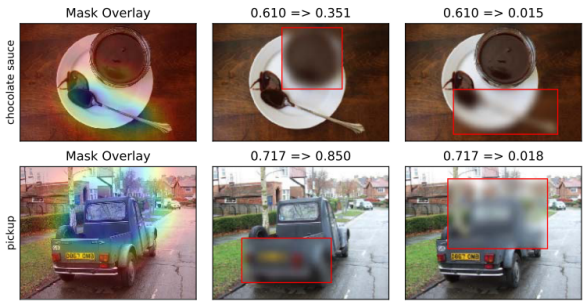

Figura 3.6. Representação de uma Explicação em Formato de Máscara de Saliência (MDS), extraída de Fong & Vedaldi [2017]. A primeira coluna indica a explicação, as colunas seguintes mostram como a inserção de ruídos na imagem afeta a sua predição e valida a explicação inicial.

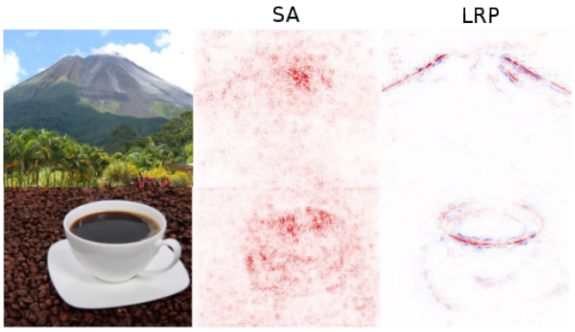

Figura 3.7. Comparação das Sinalizações de *Sensitive Analysis (SA)* e *Layer-wise Relevance Propagation (LRP)*, extraída de Samek et al. [2017].

partir de ER para dados de imagens. Dada uma RNCP, são analisadas as ativações feitas na camada anterior à camada de predição de forma a evidenciar o que, na amostra inicial, leva a determinada predição, como ilustra a Figura 3.8.

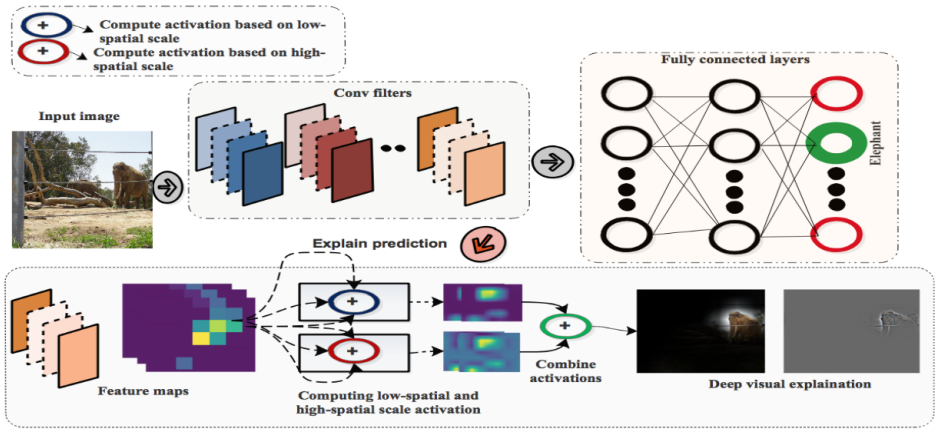

Figura 3.8. *Deep Visual Explanation (DVE)*, extraída de Khalifa et al. [2018], analisa as ativações máximas e mínimas de modo a destacar as região de maior relevância e contribuição para a predição de uma classe.

3.2.2 Explicações baseadas em Regras de Decisão

Com alto poder de explicabilidade, regras de decisão são um forte artifício de explicação e num contexto de explicação com propósito local, exigem um reforço da vizinhança da amostra analisada para que sejam extraídas justificativas e explicações de amostra, conforme demonstrado nas técnicas a seguir.

Ribeiro et al. [2018a] apresentaram o *Anchors*, um método desenvolvido a partir de Engenharia Reversa (ER), utilizando Perturbação Randômica (PRD) na vizinhança das amostras analisadas e mostrando explicações através de Explicador Baseado em Regras (EBR) para dados de imagens, textos e tabelas. Dada uma amostra de determinada classe, o método busca isolar uma área que possua a mesma predição e realiza Perturbação Randômica para reforçar a quantidade de amostras na região e aumentar o conhecimento da vizinhança. A partir disso, regras de decisão são geradas dentro da área escolhida. A Figura 3.9 demonstra um exemplo de uma intuição da técnica, seguida de um exemplo concreto de regras geradas pela técnica. O trabalho também sugere que a geração destas regras de decisão permite aos usuários inferir o comportamento de um modelo em amostras desconhecidas e, por conta disso, este método também pode ser considerado um método com propósito de explicação global.

Guidotti et al. [2018a] apresentaram o *Local Rule-Based Explanations (LORE)*, um método desenvolvido a partir de Engenharia Reversa (ER), utilizando Perturbação

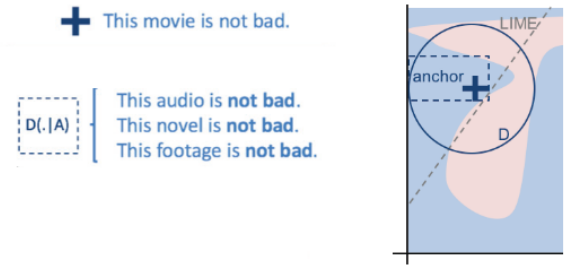

Figura 3.9. Representação da Intuição de *Anchors*, extraída de Ribeiro et al. [2018a]. A função de decisão do modelo complexo é representada pelo plano de fundo azul e rosa. A cruz azul em negrito é a instância que está sendo explicada. A linha tracejada representa a superfície de separação de classes mais próxima da amostra cruz. O quadrado pontilhado representa a área exclusiva da área azul que será trabalhada pela técnica. Os *Anchors* são definidos por um conjunto de regras exclusivamente pertencentes à classe azul.

por Algoritmo Genético (PAG) na vizinhança das amostras analisadas e mostrando explicações através de Explicador Baseado em Regras (EBR) para dados de tabelas. Em contraste à técnica anterior, esta técnica apresenta dois grupos de regras de decisão, um grupo que justifica a predição de uma amostra analisada e outro grupo que indica que mudanças seriam necessárias para a mudança de classe, como na Figura 3.10.

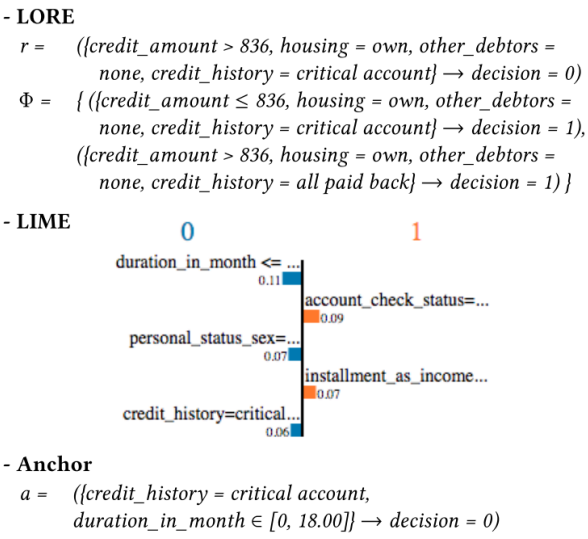

Figura 3.10. Representação do *Local Rule-Based Explanations (LORE)*, extraída de Guidotti et al. [2018a]. Comparando com as representações de outros explicadores (LIME e *Anchors*)

Rasouli & Yu [2020] apresentaram o *EXPLaining black-box classifiers using Adaptive Neighborhood generation (EXPLAN)* um método desenvolvido a partir de Engenharia Reversa (ER), utilizando Perturbação Baseada em Vizinhança (PBV) nas pro-

ximidades das amostras analisadas e mostrando explicações através de Explicador Baseado em Regras (EBR) para dados de tabelas. A técnica de perturbação de amostras utilizada nesta técnica busca gerar amostras vizinhas coerentes com o plano no qual estão inseridas. Um exemplo de explicação gerada pela técnica pode ser encontrada na Figura 3.11.

> $\boldsymbol{x}$ = {**age:** 30; **workclass:** Private; **education:** 11th; **marital-status:** Never-married; **occupation:** Prof-specialty; **relationship:** Unmarried; **race:** White; **sex:** Male; **capital-gain:** 0; **capital-loss:** 0; **hours-per-week:** 40; **native-country:** United-States — **class:** $\leq 50K$}
> $\boldsymbol{e}$ = {**age:** $\leq$ 30 $\wedge$ **capital-gain:** $\leq$ 0 $\wedge$ **hours-per-week:** ≤ 44} $\rightarrow$ **class:** $\leq 50K$

Figura 3.11. Representação de uma explicação gerada pela técnica *EXPLaining black-box classifiers using Adaptive Neighborhood generation (EXPLAN)*, extraída de Rasouli & Yu [2020], onde x é a entrada e e é a explicação.

3.2.3 Explicações baseadas em Importância de Características

A Importância de Características (IMC) é geralmente descrita em termos de uma combinação linear das características, como observado na Equação 3.1,

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.1}$$

onde $g(z')$ é a explicação de uma entrada z', $z' \in \{0,1\}^M$, M é o número de características simplificadas utilizadas na explicação e $\phi_i \in \mathbb{R}$ indica o peso (importância) associado a cada característica, com ϕ_0 indicando um termo livre (e, portanto, trata-se da equação de um plano). Ao abordar o problema de explicação de modelos desta forma é possível mensurar o impacto das características de uma amostra na sua predição. A seguir, descrevemos alguns dos trabalhos relevantes que têm estudado este problema através da abordagem de importância de características.

Ribeiro et al. [2016] apresentaram o *Local Interpretable Model-Agnostic Explanations (LIME)*, um método desenvolvido a partir de Engenharia Reversa (ER), utilizando Perturbação Randômica (PRD) na vizinhança das amostras analisadas e fornecendo explicações através de Importância de Características (IMC) para dados de imagens, textos e tabelas. Dada uma amostra a ser analisada, esta técnica busca a superfície de separação mais próxima para evidenciar o que mais contribui e o que pouco contribui

para a predição de uma determinada amostra. Para enriquecer a vizinhança e obter mais informações, é feita a Perturbação Randômica. A Figura 3.12 mostra esquema de separação e a Figura 3.13 demonstra uma forma de exibição de uma explicação segundo o LIME.

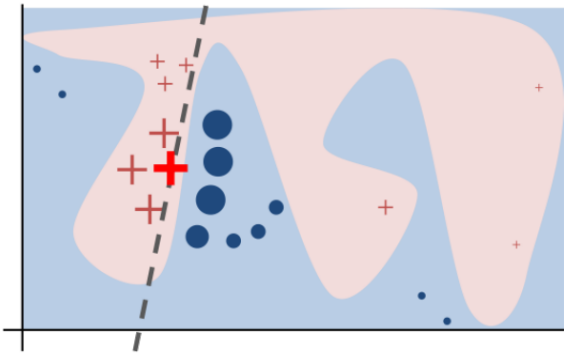

Figura 3.12. Representação da intuição do *LIME*, extraída de Ribeiro et al. [2016]. A função de decisão do modelo complexo é representada pelo plano de fundo azul e rosa, que não pode ser bem aproximada por um modelo linear. A cruz vermelha em negrito é a instância que está sendo explicada. Os exemplos adjacentes são obtidos a partir do modelo original e seus pesos são dados pela proximidade da instância que está sendo explicada, representada aqui pelo tamanho dos pontos. A linha tracejada é a explicação aprendida que é localmente, mas não globalmente, fiel à superfície de decisão complexa usada pelo modelo explicado.

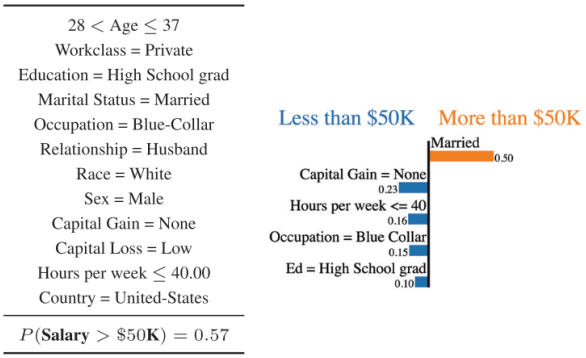

Figura 3.13. Representação de uma explicação gerada pelo LIME, extraída de Ribeiro et al. [2018a]. Onde a imagem da esquerda representa uma entrada e a predição fornecida pelo classificador, enquanto a imagem da direita representa a explicação da predição.

Lundberg & Lee [2017] apresentaram o *SHapley Additive exPlanations (SHAP)*, um método desenvolvido a partir de Engenharia Reversa (ER) e fornecendo explicações através de Importância de Características (IMC) para dados de imagens, textos e tabelas. Nesta técnica, os autores utilizam valores de *Shapley* para melhoria de

técnicas clássicas no estudo de explicação, tais como LIME [Ribeiro et al., 2016] e DeapLIFT [Shrikumar et al., 2017]. A maior contribuição desse trabalho foi unificar teoricamente os trabalhos clássicos de explicação de modelos que utilizam Importância de Características (IMC) como forma de demonstrar explicação através de de uma função linear (Equação 3.1), além de unificar características como precisão local, perda e consistência (Tabela 2.2) como sendo inerentes a soluções que envolvem importância de características. Um exemplo de explicação no contexto de imagens pode se encontrado na Figura 3.14.

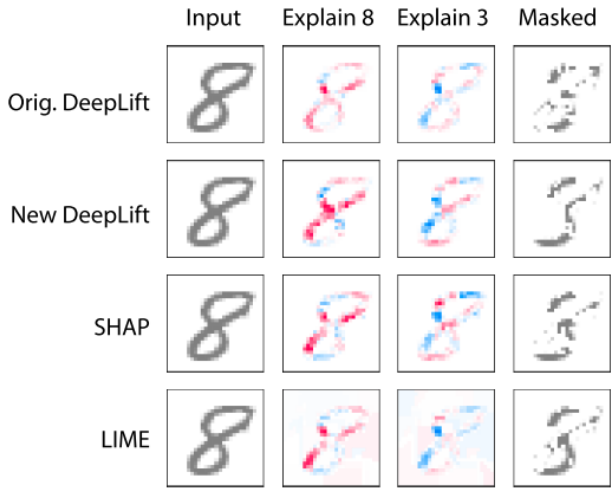

Figura 3.14. Representação de uma explicação gerada pelo SHAP em comparação com outras técnicas de explicação (DeepLift e LIME), extraída de Lundberg & Lee [2017]. A coluna da esquerda representa a imagem original, as colunas do meio representam as explicações para o 8 ser classificado como 8 e como 3, respectivamente. As sinalizações em tons vermelhos indicam forte contribuição para a predição, enquanto azul indica pouca contribuição. A última coluna utiliza a coluna anterior para gerar o dígito-alvo, no caso, o 3.

Marino et al. [2018] apresentaram uma abordagem adversarial, orientada a modelo ORM com gradientes definidos, utilizando Perturbação Baseada em Gradiente (PBG) para exibir explicação em formato de Importância de Características (IMC) para dados de tabela. Esta técnica é focada em amostras cuja predição foi mal informada por um modelo de aprendizagem de máquina, ou seja, houve erro por parte do classificador e o objetivo é apontar quais as mínimas mudanças seriam necessárias na amostra para que esta fosse corretamente classificada pelo modelo. A Figura 3.15 ilustra o método proposto.

Mais formalmente, a intuição usada pode ser traduzida como encontrar o mínimo de modificações necessárias para mudar a saída do classificador para a amostra real $\mathbf{x}$, incorretamente classificada. Assim, busca-se uma amostra adversarial $\mathbf{x}'$ classificada

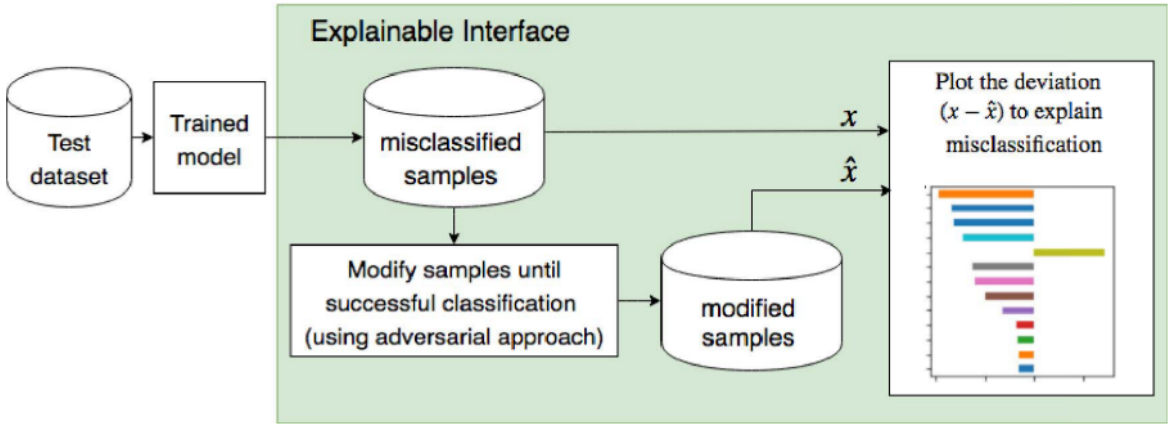

Figura 3.15. Representação da intuição da Técnica de Explicação Adversarial, extraída de Marino et al. [2018]. Esta técnica usa amostras adversariais de um modelo para determinar as mínimas alterações necessárias para a classificação correta desta instância. Ela fornece como explicação a diferença entre a amostra explicada e sua amostra adversarial.

como $\hat{y}$ (aquela que seria a classe correta de x) enquanto minimizando a distância entre $\mathbf{x}$ e $\mathbf{x}'$:

$$\begin{aligned} \min \quad & (\mathbf{x}-\mathbf{x}')^\top Q(\mathbf{x}-\mathbf{x}') \\ \text{s.t.} \quad & \arg\max_k P(y=k|\mathbf{x}',\theta) = \hat{y} \\ & \mathbf{x}_{min} \leq \mathbf{x}' \leq \mathbf{x}_{max} \end{aligned} \tag{3.2}$$

onde Q é uma matriz definida simétrica positiva, que permite ao usuário especificar um peso na métrica de diferença quadrática a ser usada para comparar a instância original e a adversarial, enquanto $\mathbf{x}_{min}$ e $\mathbf{x}_{max}$ estabelecem os limites de valores válidos (para a base de dados usada) que garantem que a imagem adversarial pertence ao domínio da distribuição de dados.

Desde que o problema estabelecido pela Equação 3.2 é difícil de resolver, o objetivo é modificado para simplificar a restrição imposta, usando uma estratégia similar àquela apresentada por Carlini & Wagner [2017]:

$$\begin{aligned} \min_{\mathbf{x}'} \quad & c\ H(\hat{y}, P(y|\mathbf{x}',\theta))\ I(\mathbf{x}',\hat{y}) + (\mathbf{x}'-\mathbf{x})^\top Q(\mathbf{x}'-\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x}_{min} \leq \mathbf{x}' \leq \mathbf{x}_{max} \end{aligned} \tag{3.3}$$

onde c é um fator que pondera a importância da amostra adversarial ser da classe alvo (a correta, ou seja, diferente da classe de $\mathbf{x}$), $H(\hat{y}, P(y|\mathbf{x}',\theta))$ é a entropia entre a previsão do modelo (parametrizado por θ) e a classe alvo e $I(\mathbf{x}',\hat{y})$ é 1 se a classe de $\mathbf{x}'$ não for $\hat{y}$ (0, caso contrário). I fornece um critério de parada para a busca de $\mathbf{x}'$, ou

seja, a amostra adversarial é considerada adequada quando ela tem a classe desejada $\hat{y}$. Comparada à estratégia de Carlini & Wagner [2017], a principal diferença está na restrição da otimização que, no caso de Marino et al. [2018], força a imagem gerada a ser compatível com a distribuição dos dados.

Pastor & Baralis [2019] apresentaram a técnica *Local Agnostic Attribute Contribution Explanation (LACE)*, um método desenvolvido a partir de Engenharia Reversa (ER), utilizando Perturbação por Algoritmo Genético (PAG) na vizinhança das amostras analisadas e fornecendo explicações através de Importância de Características (IMC), combinada com Explicador Baseado em Regras (EBR) para dados de tabelas. Nesta técnica, a partir de uma amostra, um modelo e dos dados de treinamentos, são geradas amostras sintéticas próximas à amostra inicial, para então obter o conjunto de características que traz a melhor predição, bem como o impacto isolado de cada característica da amostra inicial (Figura 3.16).

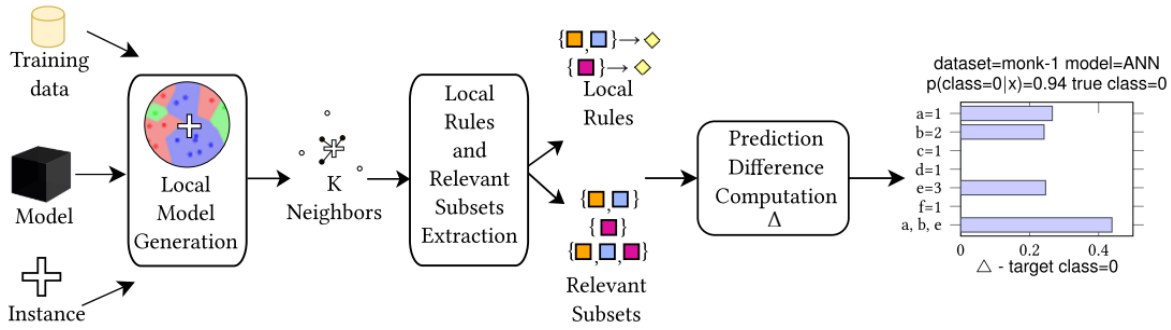

Figura 3.16. Representação do LACE, extraída de Pastor & Baralis [2019]. Um modelo local é gerado a partir da perturbação aleatória centrada na instância a se analisar. A partir disso, uma análise da vizinhança é feita e os conjuntos de regras são escolhidos para mostrar o impacto das regras na predição.

Devido seu nível de detalhamento e orientação à amostra, o interesse em métodos de explicação com propósito local tem sido foco de várias pesquisas. Estas, por sua vez, têm direcionado seus esforços na forma como seus usuários receberão suas explicações, seja através de máscaras visuais, regras de decisão de alto nível ou indicação dos dados de entrada de uma amostra que têm maior impacto em sua predição.

3.3 Síntese dos Trabalho Relacionados

Uma organização dos trabalhos da área de explicação descritos neste capítulo pode ser encontrado na Tabela 3.2. Para um melhor entendimento desta, é preciso acompanhar também a Tabela 3.1 que apresenta uma descrição das siglas presentes na tabela prin-

cipal. O que diferencia estes trabalhos é o objetivo para o qual eles se destinam, seja através de inferências sobre o comportamento de um modelo em amostras desconhecidas, seja na visualização destas explicações, na exposição de regras de alto nível de compreensão para um ser humano, ou na descoberta do impacto de partes da entrada inicial em sua predição final.

Coluna	Descrição
Referência	Trabalho onde o modelo de explicação foi proposto
Autor	Autor da proposta do modelo de explicação
Ano	Ano da publicação da proposta de modelo de explicação, variando de 2016 a 2019.
Propósito	Diz respeiro ao Propósito de uma Explicação, podendo receber os valores Local (LCL), Global (GBL), Inspeção de Modelo (ISP) ou Projeção de Modelo Transparente (PMT)
Perturbação	Diz respeito ao Reforço de Informação e Perturbações adotada pela abordagem, podendo receber os seguintes valores Perturbação Randômica (PRD), Perturbação Baseada em Vizinhança (PBV), Perturbação Baseada em Gradiente (PBG) ou Perturbação por Algoritmo Genético (PAG)
Estratégia	Diz respeito às Estratégia do Explicador, podendo receber os valores Engenharia Reversa (ER) ou Projeção de Explicação (PE)
Abordagem	Nome do modelo de explicação proposto usado neste texto
Natureza	Diz respeito à Natureza do Explicador, podendo receber os valores Agnóstico (AGN) ou Orientado a Modelo (ORM)
Explicador	Diz respeito aos Tipo do Explicador, podendo receber os valores Importância de Características (IMC), Máscara de Saliência (MDS) ou Explicador Baseado em Regras (EBR)
Entrada	Diz respeito ao tipo de entrada suportada pelo modelo de explicação, podendo receber os seguintes valores TAB, Texto (TXT), Imagem (IMG) ou Vídeo (VD)
Código	Diz respeito à disponibidade da implementação da abordagem, podendo receber os seguintes valores Sim (S) ou Não (N)

Tabela 3.1. Detalhamento das colunas da Tabela 3.2

Autor(es)/Método	Ano	Abordagem	Propósito	Perturb.	Estratégia	Natureza	Explicador	Entrada	Cód.
Kim et al.	2018	TCAV	GBL	-	ER	ORM	-	IMG	S
Okajima & Sadamasa	2019	RCN	PMT/GBL	-	PE	-	-	TAB	N
Fong & Vedaldi	2017	Meaningful P.	LCL	PBG	ER	AGN	MDS	IMG	N
Samek et al.	2017	SA	LCL	-	ER	ORM	MDS	IMG/TXT/VD	N
Samek et al.	2017	LRP	LCL	-	ER	ORM	MDS	IMG/TXT/VD	N
Khalifa et al.	2018	DVE	LCL	-	ER	ORM	MDS	IMG	N
Ribeiro et al.	2018a	Anchors	LCL/GBL	PRD	ER	AGN	EBR	TAB/TXT/IMG	S
Guidotti et al.	2018a	LORE	LCL	PAG	ER	AGN	EBR	TAB	S
Rasouli & Yu	2020	EXPLAN	LCL	PBV	ER	AGN	EBR	TAB	S
Ribeiro et al.	2016	LIME	LCL	PRD	ER	AGN	IMC	TAB/TXT/IMG	S
Lundberg & Lee	2017	SHAP	LCL	-	ER	AGN	IMC	TAB/TXT/IMG	S
Pastor & Baralis	2019	LACE	LCL	PRD	ER	AGN	IMC/EBR	TAB	N
Marino et al.	2018	Adversarial	LCL	PBG	ER	ORM	IMC	TAB	N
LARE-2M	2021	Adv. com Reforço	LCL	PBG	ER	ORM	IMC	IMG	N
LARE-MS	2021	Adv. com Reforço	LCL	PBG	ER	ORM	IMC	IMG	N

Tabela 3.2. Síntese das Abordagens Recentes para Explicação de Modelos em Aprendizagem de Máquina

3.4 Considerações Finais

Neste capítulo foram apresentados alguns os métodos de explicação de relevância recente, a partir de 2016. Eles foram distribuídos primeiramente em função do propósito de explicação, sendo globais ou locais. Os globais almejam um entendimento genérico do modelo e até uma inferência de comportamento em amostras desconhecidas enquanto os locais buscam justificativas no detalhe, orientado à amostra.

A adoção de método de um propósito ou outro vai depender apenas do objetivo de seu uso. Por exemplo, Abdul et al. [2018] apontaram a negligência no entendimento humano por parte das técnicas de explicação dos modelos caixa-preta em aprendizagem de máquina. Para o aumento da receptibilidade e confiança de técnicas de explicação, o trabalho sugere que as técnicas contemporâneas adotem as regras de decisão como parte de sua solução e a estratégia de explicadores locais para melhorar a confiança individual em uma explicação, sem esquecer de conceitos de Interação Humano-Computador (IHC).

Neste sentido, os explicadores locais apresentados neste capítulo trazem informações bem recebidas por usuários quando focam em aspectos visuais, regras de decisão de alto nível ou em indicação da importância de características. Em destaque apontamos as técnicas *Local Rule-Based Explanations (LORE)* [Guidotti et al., 2018a], *SHapley Additive exPlanations (SHAP)* [Lundberg & Lee, 2017] e *Adversarial* [Marino et al., 2018]. A primeira traz aspectos importantes ao apontar, além dos aspectos que contribuem para uma amostra pertencer a uma determinada classe, os aspectos que fazem uma amostra sair da predição sinalizada pelo modelo através de um enriquecimento da vizinhança da amostra para obtenção destas inferências de explicação. A segunda traz contribuições de outras área, ao utilizar as inferências dos valores de Shapley da teoria dos jogos, na unificação de técnicas clássicas de explicação através de indicação de importância de características. Já a última traz como contribuição a indicação das mudanças necessárias em uma amostra para que esta seja corretamente entendida pelo modelo, ao combinar técnicas adversariais focadas em amostras cujas predições foram incorretamente classificadas pelo modelo.

Sobre a técnica adversarial aplicada à explicação, é importante levantar alguns pontos. O objetivo de técnicas adversariais está ligado ao sucesso destas em enganar o modelo sem que o usuário perceba. Quando se pensa em amostra adversarial ótima e sua localização dentro do plano de um modelo, é possível inferir que tal amostra encontra-se na fronteira entre classes desse modelo, pois é nesta área que as amostras adversariais encontram maior sucesso por se tratar de uma região de transição de classes.

A inferência da utilização das técnicas adversarias como indicadoras das transições de classes é uma contribuição significativa. No entanto, é importante para várias destas técnicas que o usuário não perceba a mudança realizada. Isto contraria os objetivos na área de explicação que depende da confiança do usuário final para validação e entendimento das decisões tomadas em relação a uma amostra. Desta forma, neste trabalho, propomos a utilização de técnicas adversariais em ataques prolongados (o que chamamos de ataque adversarial com reforço), de modo que elas se afastem da fronteira de separação de classes e caracterizem melhor a classe alvo de análise. Assim, os métodos propostos neste trabalho (LARE-2M e LARE-MS, a serem apresentados no próximo capítulo) usam abordagens adversariais, como em Marino et al. [2018], porém não se restringindo às mínimas modificações necessárias para transição de classes e explorando uma técnica de ataque diferente da usada por Marino et al. [2018] e mais adequada a um processo de prolongamento iterativo do ataque.

Capítulo 4

Métodos Propostos

Neste capítulo serão apresentadas duas técnicas de explicação de amostras em RNP: a *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy (LARE-2M)* e a *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy (LARE-MS)*. Estas são técnicas de explicação local para amostras incorretamente classificadas que combinam uma estratégia de ataque adversarial com dois critérios de reforço de informação para gerar uma explicação baseada na identificação do impacto sobre certas características da mudança de classe.

A solução proposta, usada nas duas técnicas, consiste em duas etapas. Na primeira, é usada a estratégia adversarial *Maximal Jacobian-Based Saliency Map Attack (MJSMA)* para a criação de uma imagem sintética. Na segunda, a imagem gerada continua sendo perturbada, usando dois diferentes critérios de parada, para determinar um esforço de transição de classe que seja útil para justificar a classificação dada incorretamente. O primeiro critério relaxa o método MJSMA enquanto o segundo explora a superfície de decisão com base na ideia de margem ótima utilizada pelo classificador *Support Vector Machine (SVM)*. Em ambas, a ideia geral consiste em se afastar da superfície de decisão em direção à região da classe alvo.

Iniciamos este capítulo discutindo como fornecer Explicações com Técnicas Adversariais e o Reforço de Informação necessário para produzir explicações razoáveis. Em seguida, descrevemos os Algoritmos Propostos, para então apresentarmos nossas Considerações finais sobre as técnicas.

4.1 Explicações com Técnicas Adversariais

Tipos de explicação local relativamente comuns são por que motivo um modelo decidiu que uma instância x não pertence à classe y ou porque x foi incorretamente classificada

como y. Exemplos deste tipo de decisão incluem determinar porque não é concedido crédito a um cliente, porque uma pessoa foi considerada terrorista, porque uma planta não foi classificada como comestível ou porque um dígito de sete barras  não é um  .

Tomando este último caso como exemplo, é possível ilustrar a relação entre ataques e explicações. Um ataque adversarial ao dígito , direcionado para a classe , deveria produzir uma imagem similar a que o modelo acredita ser um . Do ponto de vista do modelo, as perturbações aplicadas ao representam um mínimo esforço de modificação dos atributos do para se tornar . Este mínimo esforço pode ser usado então como uma explicação do porque o não é um , neste caso, "há um traço na parte superior do que o não tem".

De fato, estratégias baseadas na geração de amostras sintéticas como mecanismo de explicação local são bastante comuns. Estas amostras sintéticas possibilitam a obtenção de informação sobre a vizinhança da superfície de decisão, delimitando as fronteiras de análise para geração de uma explicação [Ribeiro et al., 2018a, 2016; Pastor & Baralis, 2019]. Neste sentido, métodos adversariais podem representar uma forma relativamente barata de explorar o espaço. Este fato somado à intuição de que a perturbação produzida por um ataque adversarial produz uma explicação sintética motivou o seu emprego por [Marino et al., 2018].

Contudo, ao levar em conta apenas amostras adversarias na fronteira da superfície de decisão, pode-se perder informação importante relacionada com a similaridade entre o modelo aprendido e o modelo real (Figura 2.2) além de ignorar o impacto da (possível falta de) percepção humana sobre esta transição entre essas classes para geração de uma explicação. Neste sentido, é importante entender como diferentes técnicas de ataque podem ser usadas para produzir explicações já que o mínimo esforço pode não produzir necessariamente uma boa explicação.

4.1.1 Desempenho de Diferentes Estratégias de Ataque

Intuitivamente, uma ideia inicial de explicação para uma decisão incorreta de classificação binária seria visualizar a diferença entre a instância real e uma sintética obtida por um método adversarial. No caso de classificação de imagens, a explicação seria fornecida por uma imagem obtida como a diferença entre uma imagem sintética adversarial e a imagem original. Como visto anteriormente, a diferença entre a imagem original e a sintética representaria os elementos essenciais que precisariam ser observados (ou deveriam ser retirados) da imagem original para que ela fosse classificada corretamente conforme proposto por [Marino et al., 2018]. Nesta seção, sem perda de generalidade, mostramos exemplos da aplicação desta ideia ao verificar o custo de transformar um

dígito 9, do dataset de referência MNIST[1], em um 4, usando as seguintes técnicas de ataque adversarial: *DeepFool* (Figura 4.1), *Fast Gradient Sign Method (FGSM)* (Figura 4.2), *Carlini & Wagner (C&W)* (Figura 4.3) e *Jacobian-Based Saliency Map Attack (JSMA)* (Figura 4.4).

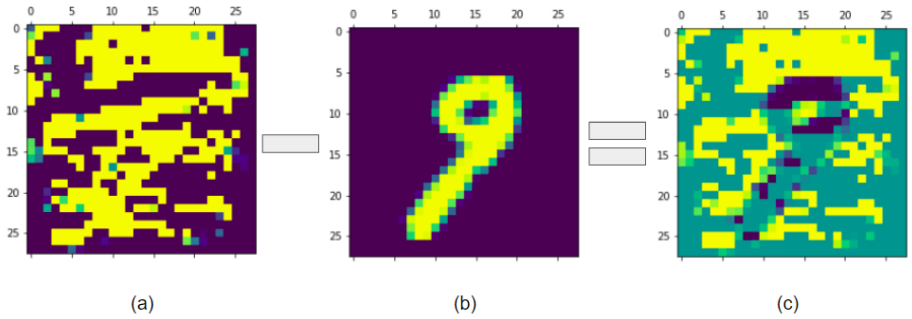

Figura 4.1. Explicando com *DeepFool*. [a] imagem adversarial obtida pelo método *DeepFool*, [b] imagem original e [c] diferença entre (a) e (b). Neste caso, [c] indica que perturbações, se inseridas na imagem original, fariam o modelo classificá-la como 4. Entre as perturbações, nota-se que a imagem adversarial não inclui a parte superior do 9. O DeepFool, por outro lado, introduz muito ruído.

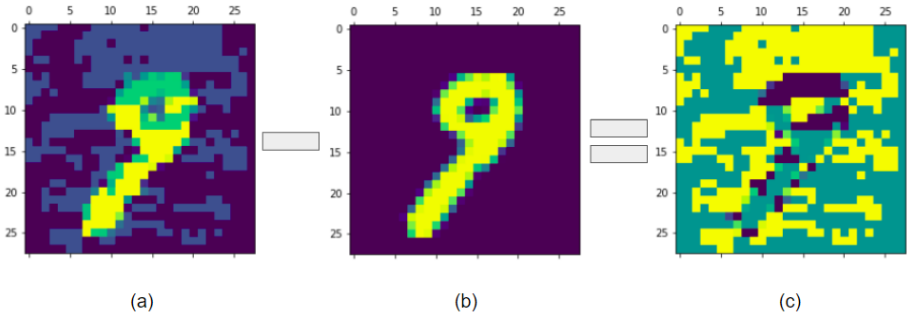

Figura 4.2. Explicando com *Fast Gradient Sign Method (FGSM)*. [a] imagem adversarial obtida pelo método *FGSM*, [b] imagem original e [c] diferença entre (a) e (b). Neste caso, [c] indica que perturbações, se inseridas na imagem original, fariam o modelo classificá-la como 4. Embora perturbações na parte superior do 9 sejam importantes, elas são suaves. Como no DeepFool, o JSMA introduz muito ruído na imagem.

[1] http://yann.lecun.com/exdb/mnist/

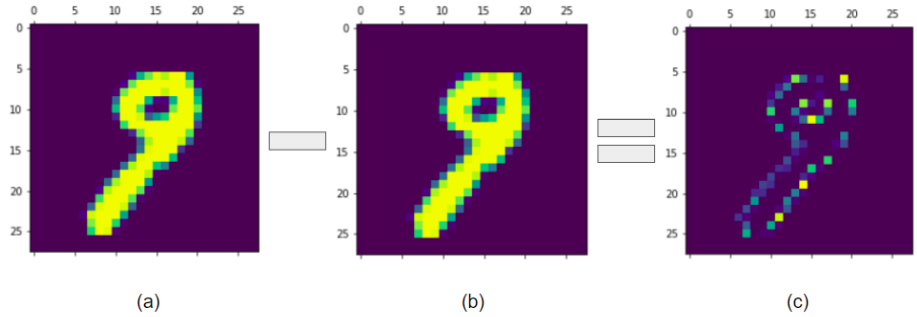

Figura 4.3. Explicando com *Carlini & Wagner (C&W)*. [a] imagem adversarial obtida pelo método *C&W*, [b] imagem original e [c] diferença entre (a) e (b). Neste caso, [c] indica que perturbações, se inseridas na imagem original, fariam o modelo classificá-la como 4. Embora o número de pertubações inseridas seja pequeno, ele se distribui por uma grande região da imagem.

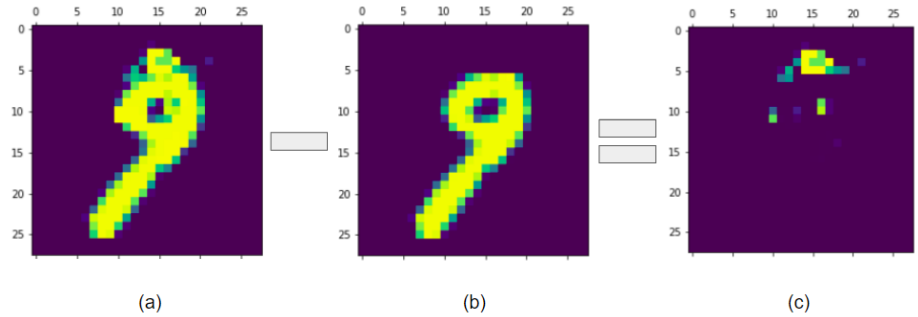

Figura 4.4. Explicando com *Jacobian-Based Saliency Map Attack (JSMA)*. [a] imagem adversarial obtida pelo método *JSMA*, [b] imagem original e [c] diferença entre (a) e (b). Neste caso, [c] indica que perturbações, se inseridas na imagem original, fariam o modelo classificá-la como 4. Neste caso, todas as perturbações são no topo da imagem, parte dela onde mais se reconhece a diferença entre 9 e 4.

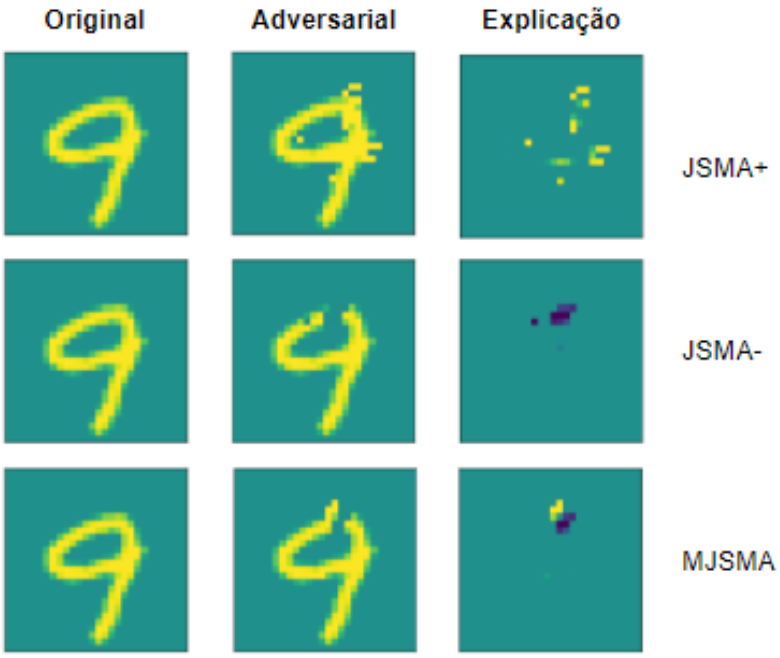

Figura 4.5. Comparação entre o *Jacobian-Based Saliency Map Attack (JSMA)* e o *Maximal Jacobian-Based Saliency Map Attack (MJSMA)* em uma abordagem de explicação. Nesta figura, a primeira linha representa o ataque JSMA com inclusão de pixels, a segunda linha representa o JSMA com exclusão de pixels e a terceira linha representa o MJSMA que combina remoção e inclusão de pixels. A coluna "Explicação" indica que pixels devem ser incluídos (amarelos) ou removidos (azuis) na imagem original (9) para que ela seja classificada como 4.

Percebemos nas figuras que os ataques estudados geram diferentes níveis de perturbação, onde métodos DeepFool e FGSM são os que mais modificam as imagens. Em imagens coloridas naturais, para os quais estes métodos foram projetados, a introdução de ruído branco na imagem é pouco perceptível por um ser humano. Em imagens de duas cores, como nas figuras, estas pertubações podem ser vistas como um excesso de ruído. De um ponto de vista de explicação, a justificativa fornecida envolveria pequenas mudanças em muitas variáveis.

Por outro lado, o método C&W, além de perturbar pouco a imagem, faz isso de forma que a imagem resultante é muito similar à original, dado o critério de otimização empregado. Este critério funciona igualmente bem para imagens coloridas e preto-e-branco. Note que, embora em pequena quantidade, as modificações se distribuem por uma ampla região da imagem. De um ponto de vista de explicação, teríamos uma justificativa envolvendo pequenas mudanças por diversas variáveis.

Finalmente, o métodos JSMA e *Maximal Jacobian-Based Saliency Map Attack* (MJSMA) produzem um número intermediário de modificações, em geral, localizadas

em regiões específicas das imagens. Isto ocorre porque estes métodos primeiro ordenam os pixels das imagens de acordo com quão bom eles distinguem uma classe de outra (quão salientes eles são). Usando esta informação, os métodos podem incluir ou excluir pixels. Por exemplo, para o exemplo envolvendo dígitos 4 e 9 (Figura 4.4), o JSMA concentra as pertubações no topo da imagem, onde intuitivamente esperamos encontrar o elemento de maior distinção entre os dígitos 4 e 9. De fato, na Figura 4.6 são apresentados os mapas de saliência para os dígitos 4 e 9 calculados para o dígito 9 e a diferença entre os mapas ressalta a importância do topo da imagem.

De um ponto de vista adversarial, esta ideia parece um tanto contra-intuitiva uma vez que modificar as regiões de máxima saliência pode ser mais perceptível pelo usuário. Contudo, a ideia do ataque é que poucas mudanças serão necessárias, o que dificulta a percepção do usuário. No contexto de explicação, teríamos uma justificativa envolvendo maiores mudanças em um grupo concentrado de variáveis, ou seja, importantes modificações em poucos conceitos. De um ponto de vista cognitivo, isto pode ser vantajoso pois um ser humano pode compreender melhor mudanças importantes em poucos conceitos que pequenas mudanças em muitos conceitos.

A Figura 4.5 apresenta uma comparação entre as estratégias da família JSMA na classificação entre digitos 4 e 9. Note que o JSMA+ indica que para o digito ser 4 é importante que ele seja mais angulado no topo e na esquerda, enquanto possui um traço mais evidente no centro se estendendo à direita. O JSMA- indica que o topo do 9 deve ser removido para ele se tornar um 4. O mesmo é indicado pelo MJSMA, que ainda ressalta a abertura no topo por incluir alguns pixels.

Como observado, nem todas as técnicas de ataque parecem igualmente efetivas no contexto de explicação. Além disso, mesmo para as técnicas baseadas em saliência, a quantidade e localização das perturbações não necessariamente é adequada para a construção de uma explicação percebida como compreensível pelo usuário. Por exemplo, na Figura 4.7, ilustramos exemplos de imagens adversariais obtidas de um ataque à imagem do dígito 9, extraída da coleção MNIST, rotulada como “original”. Os exemplos correspondem a imagens obtidas após as iterações 0, 5, 10, 12, 15, 20, 25, 30, 35, 40 e 45 usando uma versão do algoritmo MJSMA que faz uma única perturbação por iteração. Abaixo de cada imagem, uma barra colorida indica as probabilidades estimadas pelo modelo para as classes 9 (em azul) e 4 (em vermelho).

Para o exemplo ilustrado na figura, a imagem adversarial que seria escolhida pelo MJSMA é aquela obtida após 12 iterações, uma vez que neste ponto o modelo passou a acreditar que a imagem não mais pertencia à classe original 9. Embora seja possível interpretar a partir desta imagem que para um 9 se tornar um 4, é necessário remover pixels da parte superior do 9, vários usuários podem considerar que esta interpretação

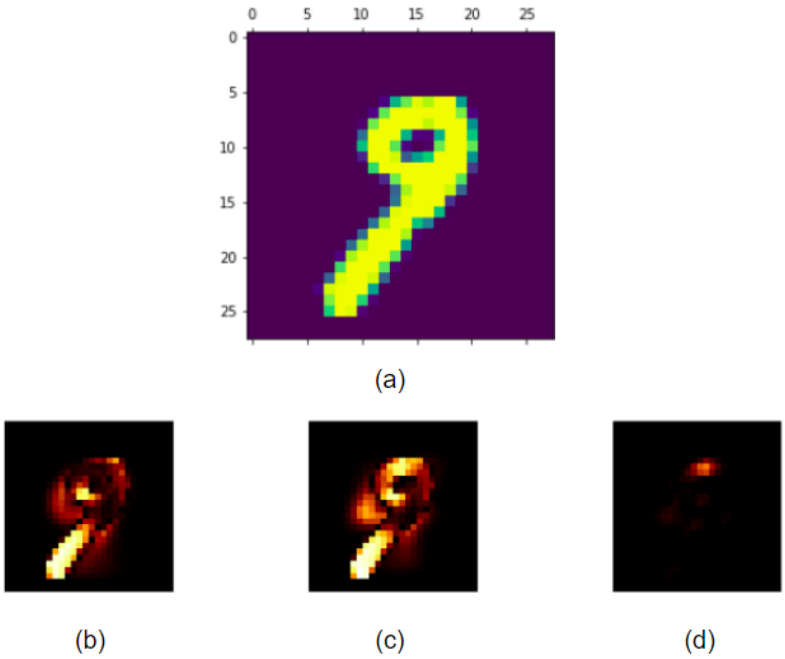

Figura 4.6. Representação de Mapas de Saliência. A imagem [a] representa um número manuscrito 9. A imagem [b] representa o mapa de saliência ativado para esta imagem ser classificada como 9, enquanto a imagem [c] representa o mapa de saliência ativado para a imagem ser classificada como 4. Sendo assim, uma possível explicação para a diferença entre os números 9 e 4, de acordo com este modelo, pode ser dado pela diferença entre [c] e [b], representada por [d]. Em outras palavras, a presença ou ausência da máscara de saliência [d] determina se um número deveria ser classificado como 4 ou 9.

é mais óbvia após a iteração 25, onde a lacuna é mais óbvia. Outros usuários podem considerar que além da parte superior do 9, a parte oeste do 9 deveria ter um aspecto mais angulado como comumente observado no número 4.

Como observado, determinar qual das imagens á mais adequada para fornecer uma explicação depende da percepção do usuário para o qual a explicação é fornecida. Com base nas observações feitas nesta seção, optamos por adotar a técnica adversarial *Maximal Jacobian-Based Saliency Map Attack* (MJSMA), baseada em mapas de saliência. Contudo, é necessário estudar até que ponto aplicar pertubações, de forma a se obter uma explicação considerada adequada para os usuários.

4.2 Reforço de Informação

Como descrito por Marino et al. [2018], um aspecto inerente a amostras adversariais, sua localização no plano e sua relação com a percepção que os usuários têm sobre

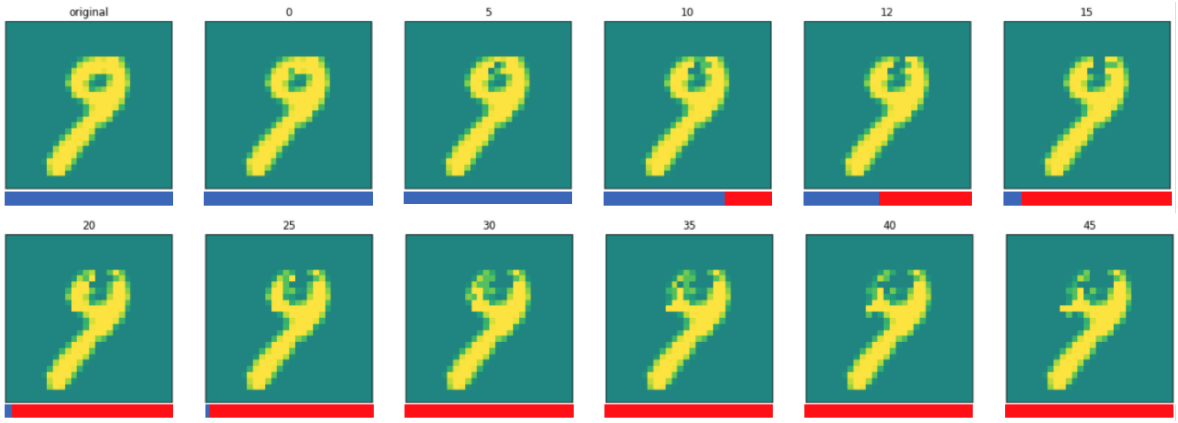

Figura 4.7. Exemplos de imagens adversariais obtidas de ataque em imagem "original" (da coleção MNIST) após as iterações 0, 5, 10, 12, 15, 20, 25, 30, 35, 40 e 45 de uma versão do algoritmo MJSMA que faz uma única perturbação por iteração. Abaixo de cada imagem a barra colorida indica as probabilidades do modelo para as classes 9 (azul) e 4 (vermelho). A imagem adversarial que seria escolhida pelo MJSMA corresponde à obtida após 12 iterações, quando a certeza do modelo é maior para a classe 4.

elas, pode ser usado como estratégia de explicação. Contudo, ao adaptar estratégias adversariais para a área de explicação, é necessário conciliarmos os objetivos destas técnicas de modo que se consiga agregar os conhecimentos sobre transição entre classes e a percepção dos usuários sobre estes processos para justificar ou explicar o processo decisório empregado pelo modelo. Para tanto, é importante compreendermos melhor a relação entre imagens adversariais e a percepção que seres humanos têm delas.

Durante o treinamento do modelo, o processo de otimização, guiado pelas amostras de treino, irá produzir um plano de separação restrito à informação observada. Por exemplo, em problemas de classificação de imagens, a informação observada normalmente se limita à distribuição de pixels (e cores) em um espaço bidimensional. Dada esta limitação inerente, este plano deve ser mais simples que aquele percebido pelo usuário do modelo, especialmente nas regiões de separação de classes onde a transição de informação pode ser abrupta, dada a natureza dos conjuntos de dados usados e as metodologias de estudo adotadas[2]. Ao contrário do modelo, a percepção do usuário envolve uma maior variedade de informação, o que inclui, mas não está limitado a, aspectos da imagem bi- e tri-dimensional e conceitos abstratos relacionados ao conteúdo

[2]Em uma coleção como a Imagenet, por exemplo, a informação conceitual de que tigres são mais similares a gatos que a cachorros está presente mas raramente é usada uma vez que a taxonomia hierárquica disponibilizada não é explorada. Assim, ela só poderá ser capturada por similaridade visual. E mesmo que não fosse assim, a informação conceitual disponível é naturalmente limitada por aspectos práticos de representação, dada as enormes possibilidades de interpretação conceitual que poderiam ser incluídas, por exemplo.

das imagens.

Assim, são nestas regiões que os ataques adversariais encontram maior sucesso ao enganar o modelo sem que usuário deste modelo perceba. A transição abrupta entre duas classes possibilita que o modelo classifique como a imagem de um cão a imagem de um gato que sofreu apenas algumas pertubações. Para um usuário, estas perturbações são imperceptíveis pois ele nota uma grande quantidade de outras evidências visuais e conceituais na imagem que lhe dão certeza de tratar-se de um gato.

Desta forma, uma amostra adversarial presente nesta região crítica tende a não representar, de fato, uma transição entre classes do ponto de vista do usuário e, portanto, pode comprometer uma possível explicação. Isto sugere que amostras adversariais ótimas podem não conter informação suficiente para produzir uma explicação considerada razoável por um usuário.

Contudo, esta amostra adversarial pode ser um ponto de partida para um processo de explicação de transição de classes se mais pertubações forem feitas até que a imagem resultante comece a ser percebida pelo usuário como uma transição suave, porém efetiva (cf. Figura 4.7). Em termos geométricos, uma amostra adversarial que explique a transição de classes deve estar mais afastada da superfície de separação encontrada pelo modelo e mais próxima dos demais pontos da classe "adversária" (Figura 4.8). Assim, uma perturbação adicional, que chamamos neste trabalho de reforço, poderia produzir uma amostra mais adequada para um mecanismos de explicação.

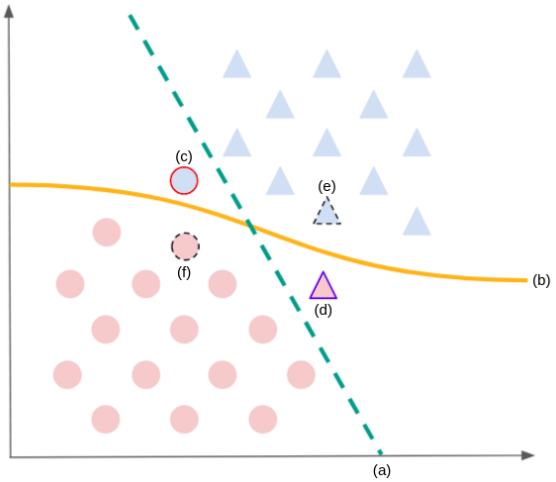

Figura 4.8. Intuição da Abordagem deste Trabalho. Considerando [a] a superfície de separação sob a percepção humana, [b] a superfície de separação aprendida por um modelo de aprendizagem de máquina, [c] e [d] amostras adversariais das classes "círculo" e "triângulo", respectivamente. As explicações a respeito das classes seriam melhor compreensíveis quando comparadas às amostras mais genéricas [e] e [f], por afastar o entendimento das classes das regiões de fronteira.

Neste trabalho investigamos duas estratégias para estimar o reforço necessário. A primeira é baseada em uma modificação da técnica adversarial MJSMA, relaxando o critério de término de forma que um número maior de perturbações seja realizada. Desta forma, permitimos que o processo continue mesmo após a detecção de uma amostra adversarial adequada para um ataque. A segunda é baseada no conceito de margem ótima de separação estabelecida por vetores de suporte como no método *Support Vector Machine* (SVM) (Figura 4.9). Um ataque adversarial tende a criar uma imagem sintética dentro da margem ótima. Uma amostra mais útil para explicação deveria ser encontrada ao sair da margem ótima para dentro do espaço da classe "adversária".

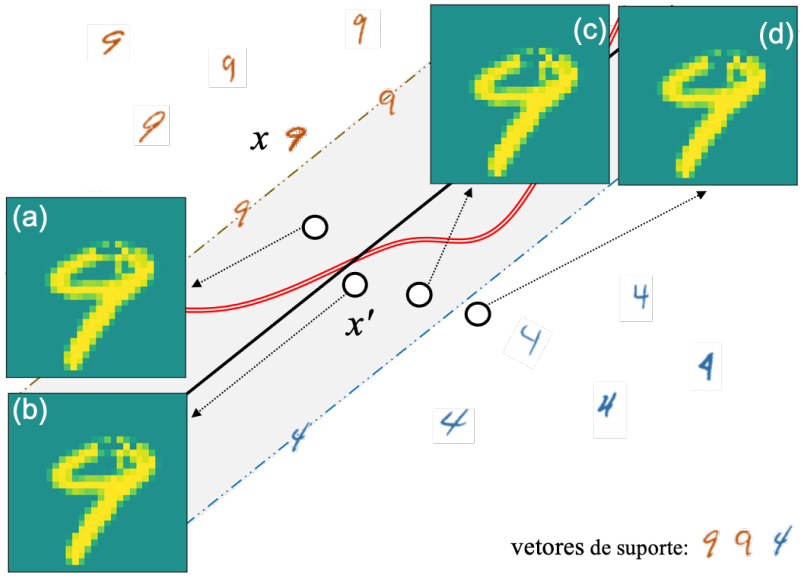

Figura 4.9. Projeção em um espaço bidimensional do uso da margem ótima de separação como critério auxiliar na explicação. A linha dupla vermelha representa a superfície de decisão do modelo a ser explicado que separa dígitos 9 de 4. A imagem x é a que deve ser explicada (e, portanto, a atacada). Após algumas iterações do MJSMA são produzidas as imagens (a) e (b), sendo (b) a primeira a ultrapassar a superfície de decisão. Esta seria, portanto, a imagem adversarial x'. Contudo, se analisarmos a superfície de separação local de (b) (ou seja, aquela obtida com base na vizinhança de (b) e representada por uma linha preta contínua), notamos que x' está dentro da margem ótima (região em cinza), definida pelos vetores de suporte das classes 9 e 4. Se continuarmos a permitir perturbações do MJSMA, são produzidas as imagens (c) e (d), cada vez mais distantes da superfície de separação local e próximas da superfície onde estão os vetores de suporte da classe 4. Intuitivamente, espera-se que a imagem obtida seja cada vez mais representativa da classe 4.

4.3 Algoritmos Propostos

4.3.1 *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy*

A técnica *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy* (LARE-2M) é baseada na técnica MJSMA para explicar a atribuição de uma amostra a uma determinada classe. A técnica proposta tem três etapas: (a) fase adversarial, (b) fase de reforço e (c) fase de explicação. Em (a), aplica-se o algoritmo do ataque adversarial MJSMA. A amostra obtida é então repassada para a etapa (b) que reforça o conhecimento em uma classe-alvo a partir de uma adaptação do próprio algoritmo MJSMA, em particular, seguindo com as iterações deste algoritmo enquanto a amostra pertencer à classe-alvo. Na etapa (c), a diferença entre as imagens em (a) e (b) é usada para determinar a sensibilidade das características modificadas no processo de transição de classes.

Diferente da técnica de explicação proposta por Marino et al., esta técnica, além de utilizar outro ataque adversarial (MJSMA), propõe uma fase adicional de reforço da amostra adversarial de forma a obter uma imagem mais representativa da classe-alvo. Para tanto, ela utiliza o mesmo ranking de alterações definidos pelo próprio MJSMA. Desta forma, enquanto Marino et al. respondem à pergunta "Qual o mínimo esforço para que uma amostra seja atribuída a uma outra classe?", neste trabalho tentamos responder a pergunta "Qual o mínimo esforço para que uma amostra seja atribuída a uma outra classe de forma perceptível para o usuário?". Para tanto, usamos o próprio MJSMA para reforçar o conhecimento sobre a classe-alvo através da forma como este ataque escolhe que mudanças de maior impacto deveriam ser realizadas.

4.3.1.1 Fase Adversarial Utilizando Ataque MJSMA

A fase adversarial da nossa técnica é descrita no Algoritmo 1, com os parâmetros de entrada e saída apresentados na Tabela 4.1. Em particular, para obter a imagem de saída, são feitas as iterações descritas nas linhas 3 a 22 (para o máximo de I_{max} iterações) enquanto a classe-alvo t não for obtida e houver pixels disponíveis a serem perturbados. A cada passo, realizamos outra iteração que busca o melhor par de pixels que podem aproximar a amostra adversarial da classe-alvo t (linhas 5 a 13). Para tanto, são usadas as derivadas parciais relativas à classe-alvo (α) e o somatório das derivadas parciais relativas às outras classes (β). A intuição é que a melhor perturbação é aquela que afeta pixels que distinguem mais a classe alvo das demais classes (-α.β). Uma vez escolhido o melhor par de pixels da iteração, são aplicadas as perturbações à amostra

adversarial x', bem como validada a necessidade de remoção do melhor par de pixels da iteração atual (linhas 17 a 21). Atingida a condição de parada principal, é retornada a amostra adiversarial gerada x', a lista de pixels não utilizados Γ na geração desta amostra e a lista de perturbação η em cada pixel, responsável pela geração da amostra x', além do valor da última iteração i realizada.

Parâmetro	**Tipo**	**Descrição**
n	Entrada	Quantidade de características das amostras. No caso da MINIST, $28 \times 28 = 784$ pixels.
x	Entrada	Amostra $x \in [0,1]^n$
y	Entrada	Classe da amostra x
t	Entrada	Classe-alvo de explicação
f	Entrada	Classificador cujo sistema decisório quer se conhecer, tal que $f(x) = y$
I_{max}	Entrada	Quantidade máxima de iterações. A fórmula para obtenção das quantidade máxima de iterações do MJSMA, descrita pelos autores do JSMA [Papernot et al., 2016], é dada por $I_{max} \leftarrow [\frac{784 \cdot \gamma}{2 \cdot 100}]$. Onde o valor 784 representa a quantidade de pixels das amostras da base MNIST e γ representa a máxima distorção que uma amostra pode sofrer. Papernot et al. [2016] observaram que o máximo valor desta distorção para se obter bons resultados seria $14,29\%$.
θ	Entrada	Passo de perturbação $\theta \in (0,1]$. Wiyatno & Xu [2018] sugerem usar $\theta = 1$ para bons resultados.
ϵ	Entrada	limite de perturbação $\epsilon \in (0,1]$
x'	Saída	amostra adversarial de x
Γ	Saída	pixels disponíveis para perturbação
η	Saída	pixels utilizados na perturbação que gerou x'
i	Saída	ultima iteração executada pelo Fase Adversarial Utilizando Ataque MJSMA que será o ponto de partida para a fase de reforço

Tabela 4.1. Detalhamento dos parâmetros do Algoritmo 1

Entrada: $x, y, t, f, I_{max}, \theta, \epsilon$
Saída: x', Γ, η, i

1 **início**
2 $x' \leftarrow x,\ \ i \leftarrow 0,\ \ \Gamma \leftarrow \{1, ..., n\},\ \ \eta \leftarrow \{0, 0, ...\}^n,$
3 **enquanto** $f(x') \neq t$ ***e*** $i < I_{max}$ ***e*** $|\Gamma| \geq 2$ **faça**
4 $\gamma \leftarrow 0$
5 **para todo** *par de pixels* $(p, q) \in \Gamma$ ***e*** *classe-alvo* t **faça**
6 $\alpha \leftarrow \sum_{k=p,q} \frac{\partial f(x')_{(t)}}{\partial x'_{(k)}}$
7 $\beta \leftarrow \sum_{k=p,q} \sum_{c \neq t} \frac{\partial f(x')_{(c)}}{\partial x'_{(k)}}$
8 **se** $-\alpha \cdot \beta > \gamma$ **então**
9 $(p^*, q^*) \leftarrow (p, q)$
10 $\gamma \leftarrow -\alpha \cdot \beta$
11 $\theta' \leftarrow sign(\alpha) \cdot \theta$
12 **fim**
13 **fim**
14 **se** $\gamma = 0$ **então**
15 $i \leftarrow i + 1$ **interrompa**
16 **fim**
17 $x'_{(p^*)}, x'_{(q^*)} \leftarrow Clip_\epsilon\{x'_{(p^*)} + \theta'\}, Clip_\epsilon\{x'_{(q^*)} + \theta'\}$
18 Remove p^* de Γ **se** $x'_{(p^*)} \notin (0, 1)$ **ou** $\eta_{(p^*)} = -\theta'$
19 Remove q^* de Γ **se** $x'_{(q^*)} \notin (0, 1)$ **ou** $\eta_{(q^*)} = -\theta'$
20 $\eta_{(p^*)}, \eta_{(q^*)} \leftarrow \theta'$
21 $i \leftarrow i + 1$
22 **fim**
23 **retorna** x', Γ, η, i
24 **fim**

Algoritmo 1: Fase adversarial das técnicas LARE-2M e LARE-MS

4.3.1.2 Fase de Reforço utilizando o MJSMA Adaptado

A fase de reforço da técnica LARE-2M pode ser descrita conforme o Algoritmo 2. As entradas desta etapa são as entradas e as saídas da anterior (Algoritmo 1, Tabela 4.1). As saídas desta etapa são a amostra adversarial reforçada x' e a lista de perturbações η responsáveis pela geração de x'.

Esta etapa da técnica LARE-2M consiste na flexibilização do ataque MJSMA enquanto a imagem adversarial x' ainda pertencer à classe-alvo t. As principais diferenças entre os algoritmos da fase adversarial (Algoritmo 1) e da fase de reforço (Algoritmo 2) estão nas linhas 4, 15 a 19 deste último algoritmo que visam assegurar que este reforço na amostra seja limitado a executar enquanto a amostra x' pertencer à classe-alvo t.

Entrada: amostra $x, y, t, f, I_{max}, \theta, \epsilon, x', i, \Gamma, \eta$
Saída: x', η
início
  enquanto $f(x') = t$ ***e*** $i < I_{max}$ ***e*** $|\Gamma| \geq 2$ **faça**
    $\gamma \leftarrow 0$
    para todo *par de pixels* $(p, q) \in \Gamma$ ***e*** *classe-alvo* t **faça**
      $\alpha \leftarrow \sum_{k=p,q} \frac{\partial f(x')_{(t)}}{\partial x'_{(k)}}$
      $\beta \leftarrow \sum_{k=p,q} \sum_{c \neq t} \frac{\partial f(x')_{(c)}}{\partial x'_{(k)}}$
      se $-\alpha \cdot \beta > \gamma$ **então**
        $(p^*, q^*), \gamma \leftarrow (p, q), -\alpha \cdot \beta$
        $\theta' \leftarrow sign(\alpha) \cdot \theta$
      fim
    fim
    se $\gamma = 0$ **então**
      interrompa
    fim
    $x'' \leftarrow x'$
    $x''_{(p^*)}, x''_{(q^*)} \leftarrow Clip_{\epsilon}\{x''_{(p^*)} + \theta'\}, Clip_{\epsilon}\{x''_{(q^*)} + \theta'\}$
    se $f(x'') \neq t$ **então**
      interrompa
    fim
    $x' \leftarrow x''$
    Remove p^* de Γ **se** $x'_{(p^*)} \notin (0, 1)$ **ou** $\eta_{(p^*)} = -\theta'$
    Remove q^* de Γ **se** $x'_{(q^*)} \notin (0, 1)$ **ou** $\eta_{(q^*)} = -\theta'$
    $\eta_{(p^*)}, \eta_{(q^*)} \leftarrow \theta'$
    $i \leftarrow i + 1$
  fim
  retorna x', η
fim

Algoritmo 2: Fase de reforço da técnica LARE-2M

4.3.1.3 Fase de Explicação

Após aplicar a Fase Adversarial Utilizando Ataque MJSMA e a Fase de Reforço utilizando o MJSMA Adaptado, a técnica LARE-2M tem informação suficiente para fornecer uma explicação. A partir de x pertencente à classe y e x' pertencente à classe-alvo t, ambas incluídas em um modelo f, é possível estabelecer uma explicação a partir das perturbações η que transformam x em x'. Estas perturbações indicam que atributos (neste caso pixels) são sensíveis nesta transição de classes e com que intensidade x pode ser transformado em x', uma amostra adversarial reforçada de forma a aumentar a aceitação, pelos usuários, de uma explicação nestes termos.

4.3.2 *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy*

A técnica *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy* (LARE-MS) é baseada no MJSMA e na exploração da margem ótima como usada no SVM. Como antes, esta técnica consiste de três fases: (a) adversarial, (b) reforço e (c) explicação. Em (a), aplica-se o algoritmo do ataque adversarial MJSMA. A amostra obtida é repassada a (b) que reforça o conhecimento em uma classe-alvo considerando a margem ótima de separação entre as classes. A última amostra obtida é então entregue para (c) que opera como na técnica LARE-2M.

4.3.2.1 Fase Adversarial Utilizando Ataque MJSMA

Para a técnica LARE-MS, a fase adversarial utiliza a mesma estratégia adotada por LARE-2M, isto é, a utilização do ataque adversarial MJSMA para obtenção de uma imagem adversarial x' de classe t, a partir de uma amostra inicial x de classe y, pertencente a um modelo f (Algoritmo 1).

4.3.2.2 Fase de Reforço utilizando o SVM Adaptado

A fase de reforço da técnica LARE-MS é descrita no Algoritmo 3. São entradas desta etapa as entradas e as saídas da anterior, adversarial (Algoritmo 1, Tabela 4.1), além das entradas adicionais: amostras de treino X, classes das amostras de treino Y, limite de paciência de execução P_{max} e margem tolerada μ. São esperadas as seguintes saídas: a amostra adversarial reforçada x' e a lista de perturbações η responsáveis pela geração de x'.

Esta etapa consiste na flexibilização do ataque MJSMA o mantendo enquanto a imagem adversarial x' não pertencer à classe inicial i e estiver dentro de uma margem de tolerância μ em uma janela de P_{max} iterações[3]. As principais diferenças entre os algoritmos da fase adversarial (Algoritmo 1) e da fase de reforço (Algoritmo 2) estão nas linhas 3 a 5 e 23 a 28 deste último algoritmo que visam assegurar que este reforço na amostra x' a mantenha com classe diferente da classe inicial y e esteja dentro da margem tolerada μ (definida pelos vetores de suporte como no SVM).

Neste algoritmo, na linha 3, é obtida a superfície de separação a partir do ponto de referência x'. Assim, dado x', são recuperadas da base de treino as K amostras da classe y mais próximas de x' bem como as K amostras da classe t mais próximas

[3]Como a superfície de separação local pode ter geometria muito irregular, a observação consistente de que houve uma transição de classes usando a janela de iterações aumenta a certeza de se ter uma imagem fora da margem ótima e, portanto, mais característica da classe destino.

de x'. A partir destas amostras, usando um kernel linear com $C = 100$, é obtida uma superfície de separação local entre as classes y e t (classificador Λ). Usando Λ, podemos verificar se uma amostra encontra-se dentro de uma margem tolerável (linha 23), bem como sua distância para a superfície de separação de Λ (linhas 4 e 26).

Entrada: $x, y, t, f, I_{max}, \theta, \epsilon, x', i, \Gamma, \eta, X, Y, P_{max}, \mu$
Saída: x', η

1 **início**
2 $\rho \leftarrow 0$
3 $\Lambda \leftarrow MargemLinearSuaveVizinhosProximos(X, Y, x', y, t, K)$
4 $\xi \leftarrow \mu{-}distância_{(x',\Lambda)}$
5 **enquanto** $f(x') \neq y$ ***e*** $i < I_{max}$ ***e*** $|\Gamma| \geq 2$ ***e*** $\rho < P_{max}$ **faça**
6 $\gamma \leftarrow 0$
7 **para todo** *par de pixels* $(p, q) \in \Gamma$ ***e*** *classe-alvo* t **faça**
8 $\alpha \leftarrow \sum_{k=p,q} \frac{\partial f(x')_{(t)}}{\partial x'_{(k)}}$
9 $\beta \leftarrow \sum_{k=p,q} \sum_{c \neq t} \frac{\partial f(x')_{(c)}}{\partial x'_{(k)}}$
10 **se** $-\alpha \cdot \beta > \gamma$ **então**
11 $(p^*, q^*), \gamma \leftarrow (p, q), -\alpha \cdot \beta$
12 $\theta' \leftarrow sign(\alpha) \cdot \theta$
13 **fim**
14 **fim**
15 **se** $\gamma = 0$ **então**
16 **interrompa**
17 **fim**
18 $x'_{(p^*)}, x'_{(q^*)} \leftarrow Clip_\epsilon\{x'_{(p^*)} + \theta'\}, Clip_\epsilon\{x'_{(q^*)} + \theta'\}$
19 Remove p^* de Γ **se** $x'_{(p^*)} \notin (0, 1)$ **ou** $\eta_{(p^*)} = -\theta'$
20 Remove q^* de Γ **se** $x'_{(q^*)} \notin (0, 1)$ **ou** $\eta_{(q^*)} = -\theta'$
21 $\eta_{(p^*)}, \eta_{(q^*)} \leftarrow \theta'$
22 $i \leftarrow i + 1$
23 **se** x' *não está dentro da margem tolerada* μ *em* Λ **então**
24 **interrompa**
25 **fim**
26 **se** $\mu{-}distância_{(x',\Lambda)} > \xi$ **então**
27 $\rho \leftarrow \rho + 1$
28 **fim**
29 **fim**
30 **retorna** x', η
31 **fim**

Algoritmo 3: Fase de reforço da técnica LARE-MS

4.3.2.3 Fase de Explicação

Após aplicar a Fase Adversarial Utilizando Ataque MJSMA e a Fase de Reforço utilizando o SVM Adaptado, a técnica LARE-MS tem informação suficiente para fornecer uma explicação, a exemplo da fase de explicação da técnica LARE-2M (Seção 4.3.1.3).

4.4 Considerações finais

Neste capítulo foram apresentadas duas técnicas de explicação: *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy* (LARE-2M) e *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy* (LARE-MS). Focadas em explicação de amostras, elas combinam conhecimentos das técnicas adversariais sobre transição de classes com técnicas de reforço de informação para indicar que características são mais afetadas por uma mudança de classe. A diferença entre as técnicas está na etapa de reforço. Enquanto LARE-2M relaxa a própria técnica adversarial para obter mais conhecimento sobre a classe-alvo, a LARE-MS utiliza o conceito de margem ótima do SVM para garantir um intervalo de confiança e um afastamento mínimo da superfície de separação.

Neste capítulo, exploramos aspectos importantes de métodos adversariais no contexto de explicação. Diferente da proposta adversarial de [Marino et al., 2018], não tentamos traduzir o aspecto mais subjetivo da percepção dos usuários sobre a razoabilidade da explicação em termos de um critério de otimização. Optamos por usar critérios mais relaxados que pretendemos avaliar diretamente em testes com usuários, no próximo capítulo.

Capítulo 5

Experimentos

Neste capítulo serão apresentados os experimentos de avaliação da técnica de explicação proposta neste trabalho a partir da seguinte organização: detalhamento dos Pré-requisitos para execução dos experimentos, descrição das técnicas que servem de Baselines, especificação do Protocolo Experimental e Metodologia dos experimentos, finalizando com a apresentação dos Resultados obtidos e Considerações Finais.

5.1 Pré-requisitos

Neste trabalho, foi utilizada a base de dados MINIST[1], contendo amostras de dígitos manuscritos de 0 a 9, distribuídas em 60000 amostras separadas para treino e 10000 amostras separadas para teste. Para classificar os dígitos desta base entre 0 a 9 usamos uma CNN com 2 camadas de convolução (4 e 10 filtros 5x5), 2 camadas de pooling, uma camada densa de 100 neurônios seguida de um camada classificadora softmax com dez neurônios. O modelo, com cerca de 17 mil parâmetros, foi treinado por 30 épocas, atingindo uma acurácia de 98.28% nas 10000 amostras de teste da MNIST. Embora este não seja o resultado de estado-da-arte para este problema, o modelo é adequado para nosso objetivo de avaliação de técnicas de explicação, dada a sua complexidade e natureza não transparente. Assim, este modelo será usado neste capítulo para avaliarmos técnicas de explicação local.

[1] http://yann.lecun.com/exdb/mnist/

5.2 Baselines

Neste trabalho foram usadas duas técnicas existentes na literatura para comparação com as técnicas propostas neste trabalho. A primeira foi proposta por Marino et al. [2018] e foi escolhida por utilizar uma estratégia adversarial para explicação de amostras. A segunda técnica é o *SHapley Additive exPlanations* (SHAP) [Lundberg & Lee, 2017] que apresenta uma técnica de explicação de importância de características bastante consistente, coesa e muito popular na área. Nossas técnicas propostas utilizam um ataque adversarial diferente de [Marino et al., 2018], além de possuir uma etapa complementar, de reforço de informação, antes de fornecer a explicação de uma amostra.

Como a técnica de [Marino et al., 2018] não está disponível publicamente, a implementamos seguindo a descrição feita no artigo publicado pelos autores. O SHAP, por sua vez, é disponível publicamente, devidamente documentado e de fácil uso. Nós utilizamos a última versão presente no GitHub[2].

5.3 Protocolo Experimental

Como a técnica de Marino et al., que de agora em diante chamamos *Marino2018*, é a mais aproximada das nossas na literatura, conduzimos as nossas avaliações adotando a mesma tarefa usada no trabalho de Marino et al.. Ou seja, avaliamos explicações dadas para amostras incorretamente classificadas por um modelo de aprendizagem de máquina, indicando que alterações seriam necessárias para reverter o erro. Estas avaliações foram feitas por usuários que tinham que selecionar a melhor explicação entre os métodos comparados, de maneira a obter o aspecto subjetivo da percepção dos usuários e sua receptividade às técnicas de explicação.

As figuras a seguir demonstram explicações fornecidas por cada uma das técnicas avaliadas neste trabalho para uma mesma amostra: (1) Marino2018 [Marino et al., 2018] (Figura 5.1), (2) *SHapley Additive exPlanations* (SHAP) [Lundberg & Lee, 2017] (Figura 5.2), (3) *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy* (LARE-2M) (Figura 5.3) e (4) *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy* (LARE-MS) (Figura 5.4). Nas figuras, a explicação procura justificar porque uma amostra foi classificada incorretamente como 5, indicando as alterações necessárias para se obter uma imagem que o modelo classificaria corretamente como 3.

[2] https://github.com/slundberg/shap

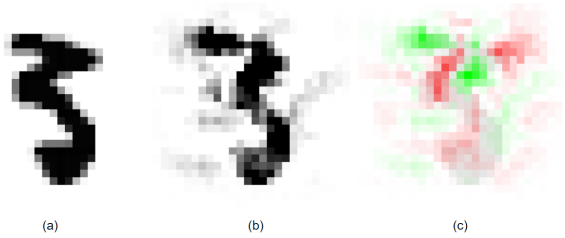

Figura 5.1. Explicação fornecida pela técnica *Marino2018* [Marino et al., 2018]. Para esta técnica, dada a amostra original incorretamente classificada como 5 (a), obtém-se uma imagem que seria classificada corretamente como 3 (b) a partir das alterações indicadas em (c). Nesta última imagem, o verde indica áreas de (a) que colaboram para obter a classe correta, enquanto o vermelho indica as áreas de (a) que menos impactaram na decisão pela classe correta.

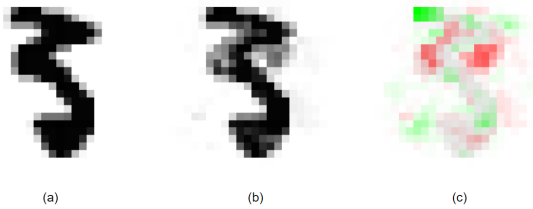

Figura 5.2. Explicação fornecida pela técnica SHAP [Lundberg & Lee, 2017]. Como antes, (a) indica a amostra original incorretamente classificada como 5; (b) é a imagem que seria classificada como 3 a partir das alterações indicadas em (c). Verde indica áreas que colaboram para decisão correta e vermelho, o oposto.

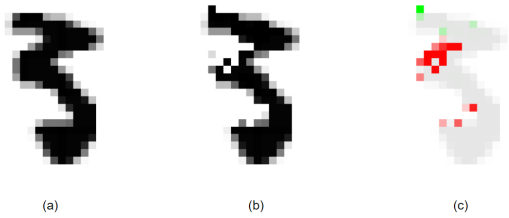

Figura 5.3. Explicação fornecida pela técnica LARE-2M. Como antes, (a) indica a amostra original incorretamente classificada como 5; (b) é a imagem que seria classificada como 3 a partir das alterações indicadas em (c). Verde indica áreas que colaboram para decisão correta e vermelho, o oposto.

Nos exemplos notamos que os métodos LARE-2M e LARE-MS produzem explicações mais sucintas. Ambos os métodos deixam claro que a parte central da imagem original é mais densa e se prolonga mais para esquerda do que se esperaria para um 3. Esta informação também está incluída nas explicações dos métodos SHAP e de Marino et al. [2018]. Contudo, estes dão mais informação, o que inclui algum ruído, especial-

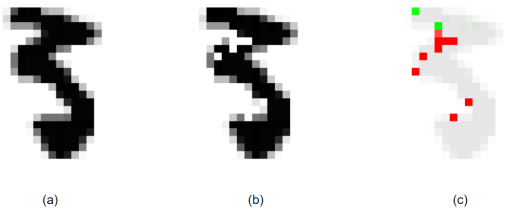

Figura 5.4. Explicação fornecida pela técnica LARE-MS. Como antes, (a) indica a amostra original incorretamente classificada como 5; (b) é a imagem que seria classificada como 3 a partir das alterações indicadas em (c). Verde indica áreas que colaboram para decisão correta e vermelho, o oposto.

mente no segundo método.

5.4 Metodologia

Para compararmos as técnicas estudadas, conduzimos quatro grupos de experimentos: (1) comparação entre técnicas de explicação adversariais; (2) desempate entre os nossos métodos propostos LARE-2M e LARE-MS; (3) comparação entre os baselines Marino2018 e SHAP [Lundberg & Lee, 2017]; e (4) comparação entre métodos propostos e o melhor baseline, o SHAP.

No primeiro grupo comparamos o LARE-2M, o LARE-MS e a técnica Marino2018. Para tanto, foram exibidos dígitos classificados incorretamente e três alternativas de sua versão corrigida fornecida por cada um dos três métodos comparados e dispostas de forma aleatória. A ideia foi usar o dígito gerado corrigido como uma síntese da importância dos atributos. Assim, para cada usuário foi solicitado que: (1) escolhessem a melhor representação entre as três, (2) as ordenassem da melhor para pior, com possibilidade de indicar empates e (3) as ordenassem da melhor para pior, *sem* a possibilidade de indicar empates.

Realizamos experimentos de ranking (melhor para pior) sem a possibilidade de indicação de empates como forma de forçar os usuários a decidirem por um método quando em dúvida. Como, para alguns casos, LARE-2M e LARE-MS geram a mesma amostra e, portanto, fornecem a mesma explicação (Figura 5.5), estes casos foram sempre considerados empates, independente do que foi indicado pelo usuário.

Ao contrário do primeiro grupo de experimentos, em todos os demais grupos uma única tarefa foi avaliada: o usuário tinha que indicar a melhor entre as representações comparadas.

Em particular, o segundo grupo de experimentos foi realizado como uma forma de

Figura 5.5. Exemplo de caso onde LARE-2M e LARE-MS geram a mesma imagem adversarial. Neste caso, (a) é a imagem original classificada incorretamente como 2 e (b) sinaliza as perturbações geradas para se obter uma imagem que seria classificada corretamente como 7.

desempatar as técnicas propostas, LARE-2M e LARE-MS, uma vez que os resultados do grupo anterior não foram conclusivos em relação a elas. No terceiro grupo de experimentos, foi avaliado o melhor entre os baselines, Marino2018 ou SHAP. A ideia era determinar como o método Marino2018 se compara com um complexo método de explicação não adversarial. Finalmente, o quarto grupo de experimentos visou comparar as técnicas propostas com o melhor baseline, o SHAP.

Para todos os experimentos, foram rodados testes estatísticos para garantir que as diferenças obtidas foram significativas. Para os testes envolvendo múltiplas comparações, usamos o teste ANOVA com correção de Bonferroni e o Teste de Friedman. Para os testes entre dois métodos usamos o teste T pareado de duas caudas. Os resultados dos testes estatísticos são reportados junto com os resultados obtidos.

5.5 Resultados

Nesta sessão apresentamos os resultados comparativos entre as seguintes técnicas de explicação: LARE-2M, LARE-MS, SHAP e Adversarial. As porcentagens demonstradas nos experimentos dizem respeito à receptividade destas técnicas de explicação pelo grupo de usuários que participou dos experimentos. Em suma, usuários recebiam explicação de várias técnicas para uma amostra e escolhiam a que melhor técnica de explicação (ou sinalizava um ranking entre as técnicas apresentadas) que lhes fazia entender uma determinada predição para de uma amostra; esta tarefa era distribuída entre várias questões ao longo dos experimentos. Desta forma, quando maior a pontuação de uma técnica, maior era a receptividade desta junto a usuários reais, implicando diretamente na comparação de explicabilidade entre as técnicas. Sendo assim, as melhores técnicas sinalizadas, indicam um melhor entendimento do processo decisório do modelo original, por parte do usuário através das melhores técnicas escolhidas.

5.5.1 O melhor entre os métodos adversariais

No primeiro grupo de experimentos, realizamos três sessões de experimentos. Na primeira, 33 usuários responderam 15 questões. Ao todo, foram 495 escolhas, além 33 empates observados automaticamente em casos em que os métodos forneceram a mesma solução, totalizando 528 avaliações. A tarefa neste experimento era escolher a melhor representação de um dígito gerado pelos métodos LARE-2M e LARE-MS e Marino2018 arranjados de forma aleatória em um formulário. A Tabela 5.1 apresenta os resultados das avaliações dos usuários enquanto a Tabela 5.2 apresenta o resultado dos testes estatísticos de confiabilidade. Como podemos observar, o método LARE-2M foi considerado o melhor em 56.25% dos casos. O método Marino2018 teve um desempenho bem aquém dos demais métodos. Os resultados foram considerados significativos tanto pelo teste ANOVA quanto pelo Teste de Friedman com $p < 0.0001$. Em termos de comparação de pares de métodos, o teste de Friedman foi inconclusivo quando comparando os métodos LARE-2M e LARE-MS.

Técnicas	*Resultado*	
	Qtde	*%*
LARE-2M	**297**	**56.25%**
LARE-MS	199	37.69%-
Marino2018	32	6.06%
TOTAL	528	

Tabela 5.1. Melhor Representação entre LARE-2M, LARE-MS e Marino2018

Teste Estatístico	*Geral*	*Método a Método*		
	p	*Adv. LARE-2M*	*Adv. LARE-MS*	*LARE-2M LARE-MS*
ANOVA - Bonferroni < 0.05	< 0.0001	< 0.05	< 0.05	< 0.05
Friedman	< 0.0001	< 0.05	< 0.05	ns

Tabela 5.2. Testes estatísticos para avaliar diferença de desempenho entre LARE-2M, LARE-MS e Marino2018

Na segunda sessão de experimentos, 17 usuários ranquearam 15 triplas de dígitos fornecidas pelos métodos LARE-2M, LARE-MS e Marino2018, construindo 255 rankings, sem possibilidade de indicar empates. Como antes, nos casos de mesma solução fornecida por LARE-2M e LARE-MS, um empate foi computado automaticamente (o que foi previamente informado aos usuários). A Tabela 5.3 apresenta os resultados das

avaliações dos usuários enquanto a Tabela 5.4 apresenta o resultado dos testes estatísticos de confiabilidade. Experimentos com ranking são melhores para avaliar como os usuários discernem o desempenho dos métodos, pois eles são obrigados a prestar mais atenção nas alternativas. Ao contrário do resultado anterior (cf. Tabela 5.1), o método LARE-MS foi melhor que o LARE-2M, embora os desempenhos tenham sido mais próximos. Da mesma forma, o método Marino2018 também teve desempenho melhor, embora ainda seja apontado como o pior entre os três. Como antes, as diferenças foram significativas com $p < 0.0001$ de acordo com ANOVA e Friedman. Desta vez, tanto o ANOVA quanto o teste de Friedman indicam empate quando comparados os métodos LARE-2M e LARE-MS diretamente.

Técnicas	*Distribuição do Ranking*						*Total*
	#1		*#2*		*#3*		
	Qtde	*%*	*Qtde*	*%*	*Qtde*	*%*	*Qtde*
LARE-2M	100	37.17%	**111**	**43.19%**	44	18.41%	255
LARE-MS	**107**	**39.78%**	99	38.52%	49	20.50%	255
Marino2018	62	23.05%	47	18.29%	**146**	**61.09%**	255
TOTAL	269		257		239		

Tabela 5.3. Ranking de LARE-2M, LARE-MS e Marino2018, com possibilidade de sinalização de empate, pelo usuário.

Teste Estatístico	*Geral*	*Método a Método*		
	p	*Adv. LARE-2M*	*Adv. LARE-MS*	*LARE-2M LARE-MS*
ANOVA - Bonferroni < 0.05	< 0.0001	< 0.05	< 0.05	> 0.05
Friedman	< 0.0001	< 0.05	< 0.05	ns

Tabela 5.4. Teste estatístico para avaliar diferença significativa no ranking de LARE-2M, LARE-MS e Marino2018, com possibilidade de sinalização de empate, pelo usuário.

Nesta sessão de experimentos notamos que os usuários tendem a optar pelo empate quando em dúvida. Além disso, à medida que vão avançando na sessão, mesmo sem saber quais exatamente são os métodos que estão produzindo as explicações, eles começam a observar o padrão de que dois métodos dão resultados mais similares e bem melhores que o terceiro. É possível que isso leve os usuários a adotar uma prática padrão de optar, sem muita consideração, pelo empate entre os dois melhores métodos.

Para desestimular este comportamento, realizamos uma terceira sessão de experimentos, similar à segunda, mas em que não era permitindo aos usuários indicar empate. Eles foram avisados, contudo, que nos casos de mesma solução fornecida por LARE-2M e LARE-MS, um empate seria computado automaticamente.

Assim, na terceira sessão de experimentos, 16 usuários ranquearam 15 triplas de dígitos fornecidas pelos métodos LARE-2M, LARE-MS e Marino2018, construindo 240 rankings, sem possibilidade de indicar empates. Como antes, a Tabela 5.5 apresenta os resultados das avaliações dos usuários enquanto a Tabela 5.6 apresenta o resultado dos testes estatísticos de confiabilidade. Os resultados obtidos confirmam aqueles da primeira sessão (cf. Tabela 5.1), mas fica claro que a diferença percebida em relação aos métodos LARE-2M e LARE-MS é pequena. Como antes, as diferenças foram significativas com $p < 0.0001$, mas tanto o ANOVA quanto o teste de Friedman falham em apontar como significativa a diferença entre os métodos LARE-2M e LARE-MS quando comparados diretamente entre si.

Técnicas	*Distribuição do Ranking*						
	#1		*#2*		*#3*		*Total*
	Qtde	*%*	*Qtde*	*%*	*Qtde*	*%*	*Qtde*
LARE-2M	**130**	**51.59%**	96	39.83%	14	6.17%	240
LARE-MS	109	43.25%	**113**	**46.89%**	18	7.93%	240
Marino2018	13	5.16%	32	13.28%	**195**	**85.90%**	240
TOTAL	252		241		227		

Tabela 5.5. Ranking de LARE-2M, LARE-MS e Adversarial

Teste Estatístico	*Geral*	*Método a Método*		
	p	*Adv. LARE-2M*	*Adv. LARE-MS*	*LARE-2M LARE-MS*
ANOVA - Bonferroni < 0.05	< 0.0001	< 0.05	< 0.05	> 0.05
Friedman	< 0.0001	< 0.05	< 0.05	ns

Tabela 5.6. Teste estatístico para avaliar diferença significativa no ranking de LARE-2M, LARE-MS e Adversarial

Como conclusão geral deste primeiro bloco de experimentos, observamos que os métodos propostos são melhores que o Adversarial, não ficando claro quem é o melhor entre LARE-2M e LARE-MS.

5.5.2 A melhor entre as técnicas propostas

Dado o resultado inconclusivo do primeiro bloco de experimentos em relação a LARE-2M e LARE-MS, conduzimos um segundo bloco com mais usuários para confirmar se havia diferença entre os métodos considerando um nível de confiança de 95%. Neste experimento, 46 usuários responderam a 15 questões, totalizando 690 escolhas. Considerando os 46 empates observados de forma automática para duas questões, ao fim obtivemos 736 avaliações. A tarefa neste experimento era escolher a melhor representação entre os dois algoritmos. Os resultados obtidos estão na Tabela 5.7. Ao observamos o número de questões por usuários, e levando em consideração o intervalo de confiança, podemos afirmar que os usuários preferem o método LARE-2M em 56% das ocasiões.

Técnicas	*Resultado*			
	Questões	*%*	*Questões por Usuário*	*%*
LARE-2M	401	54.48%	8.717 ± 0.463	56.16%
LARE-MS	335	45.52%	6.804 ± 0.458	43.84%
TOTAL	690			

Tabela 5.7. Comparação entre LARE-2M e LARE-MS. A coluna "Questões por usuário" é informada como a média mais ou menos intervalo de confiança considerando um nível de confiança de 95%.

5.5.3 O melhor entre os baselines

Neste bloco de experimentos, 46 usuários responderam 15 questões, totalizando 690 escolhas. A tarefa dada consistiu em determinar qual o melhor baseline entre os escolhidos. O resultado desta avaliação é apresentado na Tabela 5.8. O método SHAP, como esperado, foi melhor que o Adversarial em 65.65% dos casos. Este resultado foi observado significativo com $p = 0.02$ em Teste T pareado de duas caudas. O método SHAP consiste de uma madura técnica de explicação que combina ideias de várias outras na literatura, embora nenhuma baseada em ataques adversariais.

5.5.4 LARE-2M e LARE-MS versus SHAP

Finalmente, no último bloco de experimentos, comparamos nossas técnicas contra o melhor baseline, o SHAP. Foram realizadas duas sessões de experimentos, onde os usuários deviam escolher a melhor representação entre duas dispostas aleatoriamente, obtidas dos algoritmos (1) SHAP e LARE-2M na primeira sessão e (2) SHAP e LARE-MS na segunda sessão. Participaram da primeira sessão de experimento 46 usuários,

Técnicas	*Resultado*	
	Qtde	*%*
Adversarial	237	34.35%
SHAP	**453**	**65.65%**
TOTAL	690	

Tabela 5.8. Melhor Representação entre Adversarial e SHAP

respondendo 15 questões, para um total de 690 escolhas. Os resultados das avaliações dos usuários estão na Tabela 5.9. O método LARE-2M foi melhor que o SHAP em 67.68% dos casos. Este resultado foi observado significativo com $p = 0.002$ em Teste T pareado de duas caudas

Técnicas	*Resultado*	
	Qtde	*%*
LARE-2M	**467**	**67.68%**
SHAP	223	32.32%
TOTAL	690	

Tabela 5.9. Melhor Representação entre SHAP e LARE-2M

Finalmente, participaram da segunda sessão de experimento 33 usuários, respondendo 15 questões, para um total de 495 escolhas. Os resultados das avaliações dos usuários estão na Tabela 5.10. O método LARE-MS foi melhor que o SHAP em 74.34% dos casos. Este resultado foi observado significativo com $p < 0.0001$ em Teste T pareado de duas caudas

Técnicas	*Resultado*	
	Qtde	*%*
LARE-MS	**368**	**74.34%**
SHAP	127	25.66%
TOTAL	495	

Tabela 5.10. Melhor Representação entre SHAP e LARE-MS

5.6 Considerações Finais

Neste capítulo, avaliamos as técnicas propostas – *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy* (LARE-2M) e *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy* (LARE-MS) – comparando-as com dois baselines, o método adversarial proposto por Marino et al. [2018] e o SHAP Lundberg & Lee [2017]. Ambas foram significativamente melhores que os baselines de acordo com os usuários.

O desempenho fraco dos baselines pode ser explicado, em parte, pela percepção dos usuários de que eles fornecem explicações complexas para as instâncias de teste. Esta complexidade é comumente vista como ruído, já que pixels inseridos/removidos muitas vezes não parecem estar associados aos casos em observação. É importante frisar que estes métodos foram propostos para explicação geral de instâncias complexas, o que inclui imagens coloridas de cenas reais, descrições discretas de casos de invasão em redes, usuários com problemas de crédito etc. Uma comparação mais sistemática com estes métodos deve considerar estas outras bases de dados conceitualmente mais complexas.

Emboras as técnicas propostas tenham desempenho muito similar, os usuários consideram a *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy* (LARE-2M) levemente superior à *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy* (LARE-MS). Os resultados obtidos indicam que a nossa hipótese inicial de que um reforço de informação em relação a um ataque adversarial podem representar uma contribuição útil para o campo de explicação de imagens.

Capítulo 6

Conclusão

A complexidade crescente de modelos de aprendizagem de máquina está intrinsecamente associada à sua alta acurácia e baixa transparência, o que tem motivado o desenvolvimento de técnicas para explicar os processos decisórios internos desses algoritmos. Além disso, quanto maior a sensibilidade da área de aplicação, pelo impacto da utilização de algoritmos de aprendizagem de máquina, maior é a necessidade de transparência nos processos decisório destes algoritmos.

Como consequência, muitas técnicas de explicação tem sido propostas na literatura. Dependendo do propósito, estas são técnicas que buscam o entendimento genérico do modelo – Explicação Global (GBL) – ou de amostras específicas – Explicação Local (LCL). Dados seu objetivo e demanda, explicações orientadas a amostra têm sido muito estudadas, o que levou ao desenvolvimento de muitas estratégias gerais, como Máscara de Saliência (MDS), Explicador Baseado em Regras (EBR) ou Importância de Características (IMC).

Muitos destes métodos exploram a vizinhança das amostras analisadas para delimitar a superfície de separação mais próxima e isolar a área sobre a qual serão extraídos os artefatos de explicação de uma amostra inicial. Este conhecimento pode ser obtido através de técnicas de geração de amostras sintéticas ou através da utilização do conhecimento dos ataques adversariais sobre a transição de classes.

Neste trabalho foram propostos dois métodos de explicação de amostras, *Local Adversarial-Reinforcement-Based Explanation through MJSMA-MJSMA Strategy* (LARE-2M) e *Local Adversarial-Reinforcement-Based Explanation through MJSMA-SVM Strategy* (LARE-MS), baseados na utilização de uma técnica adversarial, seguida de uma etapa de reforço de informação. Nos baseamos na ideia de que uma imagem sintética obtida por meio de um ataque adversarial representa um esforço mínimo para a observação de uma transição de classes. Contudo, assumimos que a simples utilização

da técnica adversarial não é suficiente para a obtenção de uma justificativa adequada, do ponto de vista de um usuário, dadas as diferenças de objetivos entre um ataque e uma explicação. Desta forma, complementamos a utilização de técnicas adversariais com o uso de técnicas de reforço de informação para obter uma amostra sintética mais distante da superfície de separação e, portanto, mais representativa da classe-alvo. Com base nesta imagem, é possível indicar as características mais impactadas pela transição de classes sob a perspectiva do modelo e do usuário.

Estas técnicas foram comparadas com a técnica de explicação adversarial proposta por Marino et al. [2018] e com o *SHapley Additive exPlanations* (SHAP) [Lundberg & Lee, 2017], para justificar decisões de classificação incorreta feitas pelo modelo a ser explicado. Ambos os trabalhos fornecem explicações baseadas em Importância de Características (IMC) com o primeiro adotando apenas os conhecimentos de ataques adversariais para atingir seu objetivo.

Baseado em nossos experimentos, o método LARE-2M se mostrou levemente superior ao LARE-MS, com ambos significativamente melhores que as técnicas encontradas na literatura com objetivo parecido. Por exemplo, quando comparado ao SHAP, o método mais usado na área, as técnicas apresentaram ganhos de 67.68% e 74.34%. Os resultados dos experimentos indicam que os usuários parecem concordar que uma boa explicação baseada em uma técnica adversarial é aquela baseada em informação de pontos mais distantes da superfície de separação do que os que seriam mais adequados para um ataque.

6.1 Limitações

De um ponto de vista mais geral, algumas limitações deste trabalho estão relacionadas com limitações gerais da própria área de explicação. Esta é uma área ainda imatura que carece de unidade. Isto fica claro na falta de formalismos básicos, consenso sobre conceitos e uma teoria clara de avaliação. Este último aspecto é particularmente importante, pois afeta diretamente a confiabilidade nos modelos. Muitos trabalhos, por exemplo, requerem que avaliadores humanos sejam especialistas, o que dificulta mais ainda o desenvolvimento na área.

Considerando o nosso trabalho de forma mais específica, é importante notar que conclusões mais definitivas sobre os métodos propostos requerem contextualização detalhada e avaliação mais sistemática. Por exemplo, o fraco desempenho dos baselines em nossa avaliação pode estar relacionado ao fato de que eles foram propostos para explicação geral. Assim, para uma avaliação mais correta, uma vez definido o contexto

de imagens, deveríamos considerar outras coleções além da MNIST como, por exemplo, uma que envolvesse imagens coloridas de cenas reais. Isso permitiria entender melhor como a proposta realmente se compara a um método geral como o SHAP.

Outra limitação dos métodos propostos é como lidar com problemas onde instâncias têm características discretas. Estes são problemas muito comuns e para os quais há enorme demanda por explicadores. De fato, indo mais além, a própria contextualização que impusemos neste trabalho (lidar com imagens incorretamente classificadas) é importante, dada a enorme variedade de possíveis explicações e aplicações.

6.2 Trabalhos Futuros

Considerando as limitação mais gerais descritas anteriormente, uma linha de pesquisa importante nesta área envolve a compreensão melhor do fator humano, desde o estudo sobre a interferência e participação nos processos de explicação até os aspectos mais subjetivos de sua percepção. De fato, estes aspectos mais subjetivos afetam significativamente propriedades importantes de uma boa explicação como coerência, consistência e adequação. Embora tenhamos descrito estas propriedades como requisitos deste trabalho em nossos objetivos iniciais, não há muito consenso em como fazer estas avaliações.

Em termos mais específicos, é importante ampliar a avaliação realizada neste trabalho para mais coleções e tipos de aplicação. Possibilidades nesta linha incluem entender como aplicar as técnicas discutidas no contexto de informações textuais e tabulares e comparar novamente com a técnica proposta por Marino et al. [2018], que suporta este tipo de informação. Para coleções de cenas reais, um possível problema é como estabelecer a importância de regiões de maior ou menor granularidade. Nossas implementações removem ou inserem dois pixels por iteração. Uma modificação que talvez pudesse ser útil para imagens complexas seria definir regiões da imagem (objetos por exemplo) que deveriam ser tratadas como conceitos relativamente independentes (por exemplo, uma pessoa na foto de um parque).

Em termos dos métodos adversariais estudados, é importante analisar novas formas de se calcular saliência que não seja restrita apenas à classe incorreta e à classe alvo. Em termos de explicação, seria importante estimar a importância de uma dimensão também em relação a todas as demais classes. Um conceito pode ser mais útil para uma explicação se em vez de ser apenas significativo para as classes da transição, ele também é um tanto exclusivo para elas. A intuição é que uma explicação pode não ser tão boa se ela serve igualmente para explicar qualquer coisa.

Finalmente, pretendemos publicar nosso trabalho e os códigos das nossas técnicas para contribuir com trabalhos futuros interessados em nossas técnicas.

Referências Bibliográficas

Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B. Y. & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. Em *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 582:1----582:18, New York, NY, USA. ACM. Citado na página 46.

Adadi, A. & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138--52160. ISSN 2169-3536 VO - 6. Citado 2 vezes nas páginas 3 e 4.

Biran, O. & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable Artificial Intelligence (XAI)*, pp. 8--13. ISSN 13563289. Citado 2 vezes nas páginas 3 e 25.

Carlini, N. & Wagner, D. A. (2017). Towards evaluating the robustness of neural networks. Em *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39--57. IEEE Computer Society. Citado 4 vezes nas páginas 13, 15, 41 e 42.

Cohen, S. (2021). The evolution of machine learning: past, present, and future. Em *Artificial Intelligence and Deep Learning in Pathology*, pp. 1--12. Elsevier. Citado na página 10.

Dağlarli, E. (2020). Explainable artificial intelligence (xai) approaches and deep meta-learning models. Em Aceves-Fernandez, M. A., editor, *Advances and Applications in Deep Learning*, capítulo 5. IntechOpen, Rijeka. Citado na página 2.

Doshi-Velez, F. & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*. Citado 2 vezes nas páginas 25 e 27.

Fong, R. C. & Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. Em *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449--3457. Citado 6 vezes nas páginas 19, 20, 25, 34, 35 e 45.

Goodfellow, I. J.; Shlens, J. & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. Citado na página 13.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F. & Giannotti, F. (2018a). Local Rule-Based Explanations of Black Box Decision Systems. *CoRR*. Citado 8 vezes nas páginas 4, 20, 22, 24, 36, 37, 45 e 46.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F. & Pedreschi, D. (2018b). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):1--42. ISSN 3600300. Citado 8 vezes nas páginas 3, 4, 11, 17, 19, 20, 21 e 24.

Jia, Y.; Bailey, J.; Ramamohanarao, K.; Leckie, C. & Houle, M. E. (2019). Improving the Quality of Explanations with Local Embedding Perturbations. Em *25th ACM-SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19)*. Citado 3 vezes nas páginas 6, 20 e 22.

Khalifa, H.; Babiker, B. & Goebel, R. (2018). An introduction to deep visual explanation. Citado 3 vezes nas páginas 34, 36 e 45.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F. & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Em *Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018*. Citado 5 vezes nas páginas 20, 25, 32, 33 e 45.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S. & Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. *CoRR*, abs/1902.00006. Citado 3 vezes nas páginas 27, 28 e 29.

Laugel, T.; Renard, X.; Lesot, M.; Marsala, C. & Detyniecki, M. (2018). Defining locality for surrogates in post-hoc interpretablity. *CoRR*, abs/1806.07498. Citado na página 6.

Lundberg, S. M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Em Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S. & Garnett, R., editores, *Advances in Neural Information Processing Sys-*

tems 30, pp. 4765--4774. Curran Associates, Inc. Citado 11 vezes nas páginas 24, 25, 39, 40, 45, 46, 66, 67, 68, 75 e 77.

Marino, D. L.; Wickramasinghe, C. S. & Manic, M. (2018). An Adversarial Approach for Explainable AI in Intrusion Detection Systems. Em *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 3237--3243. Institute of Electrical and Electronics Engineers Inc. Citado 17 vezes nas páginas 25, 29, 40, 41, 42, 45, 46, 47, 49, 54, 58, 64, 66, 67, 75, 77 e 78.

Moosavi-Dezfooli, S.-M.; Fawzi, A. & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. Em *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Citado 2 vezes nas páginas 13 e 14.

Okajima, Y. & Sadamasa, K. (2019). Deep Neural Networks Constrained by Decision Rules. Em *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*. Citado 4 vezes nas páginas 20, 21, 33 e 45.

Oyallon, E. (2017). Building a regular decision boundary with deep networks. Em *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Citado na página 12.

Pang, T.; Du, C.; Dong, Y. & Zhu, J. (2018). Towards Robust Detection of Adversarial Examples. Em *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*, pp. 4584--459. Citado na página 22.

Papernot, N.; McDaniel, P. D.; Jha, S.; Fredrikson, M.; Celik, Z. B. & Swami, A. (2016). The limitations of deep learning in adversarial settings. Em *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pp. 372--387. IEEE. Citado 4 vezes nas páginas 13, 15, 17 e 59.

Pastor, E. & Baralis, E. (2019). Explaining black box models by means of local rules. Em *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pp. 510--517. Citado 4 vezes nas páginas 24, 42, 45 e 49.

Rasouli, P. & Yu, I. C. (2020). EXPLAN: Explaining Black-box Classifiers using Adaptive Neighborhood Generation. Em *Proceedings of the International Joint Conference on Neural Networks*. Institute of Electrical and Electronics Engineers Inc. Citado 5 vezes nas páginas 21, 24, 37, 38 e 45.

Ribeiro, M. T.; Singh, S. & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Em *Proceedings of the 22Nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 1135--1144, New York, NY, USA. ACM. Citado 12 vezes nas páginas 18, 19, 20, 21, 22, 24, 25, 38, 39, 40, 45 e 49.

Ribeiro, M. T.; Singh, S. & Guestrin, C. (2018a). Anchors: High-Precision Model-Agnostic Explanations. *Thirty-Second AAAI Conference on Artificial Intelligence*. Citado 9 vezes nas páginas 3, 20, 22, 24, 36, 37, 39, 45 e 49.

Ribeiro, M. T.; Singh, S. & Guestrin, C. (2018b). *Semantically Equivalent Adversarial Rules for Debugging NLP models*. Association for Computational Linguistics (ACL). Citado na página 21.

Ruan, W.; Yi, X. & Huang, X. (2020). Adversarial Robustness of Deep Learning: Theory, Algorithms, and Applications. Relatório técnico, 20th IEEE International Conference on Data Mining (ICDM). Citado na página 12.

Samek, W.; Wiegand, T. & Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296. Citado 5 vezes nas páginas 20, 25, 34, 35 e 45.

Schneider, J. & Handali, J. (2019). Personalized explanation in machine learning. Citado 2 vezes nas páginas 19 e 20.

Schuessler, M. & Weiß, P. (2019). Minimalistic Explanations: Capturing the Essence of Decisions. Em *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pp. LBW2810:1----LBW2810:6, New York, NY, USA. ACM. Citado 2 vezes nas páginas 3 e 4.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D. & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*. Citado na página 24.

Shrikumar, A.; Greenside, P. & Kundaje, A. (2017). Learning important features through propagating activation differences. Citado 2 vezes nas páginas 24 e 40.

Wiyatno, R. & Xu, A. (2018). Maximal jacobian-based saliency map attack. *CoRR*, abs/1808.07945. Citado 3 vezes nas páginas 16, 18 e 59.

Yang, F.; Du, M. & Hu, X. (2019a). Evaluating explanation without ground truth in interpretable machine learning. *CoRR*, abs/1907.06831. Citado na página 11.

Yang, F.; Du, M. & Hu, X. (2019b). Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. *arXiv*. Citado na página 23.

Yuan, X.; He, P.; Zhu, Q. & Li, X. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*. ISSN 2162-237X. Citado na página 12.