Interpretabilidade em Aprendizado de Máquina

Tibor Zequini Boglár

Resumo A utilização de modelos de aprendizado de máquina nos últimos anos vêm crescendo drasticamente, abrangendo um amplo espectro de atuação, com aplicações que vão desde a esfera acadêmica até a esfera industrial. Apesar dos modelos serem capazes de executar tarefas complexas atingindo grande acurácia, diversas vezes são subutilizados em áreas de risco devido a falta de transparência na obtenção dos resultados finais. Nesse relatório são abordadas conceitos sobre interpretabilidade, a dificuldade de criação de modelos interpretáveis, e como o raciocínio probabilístico está imerso nessa problemática.

Ciências Moleculares
Relatório de Iniciação à Pesquisa
Universidade de São Paulo
Junho, 2019

Conteúdo

Lista de Figuras

Capítulo 1

Objetivos

1.1 Problema de Pesquisa

Atualmente, estudos voltados à interpretabilidade em modelos de inteligência artificial estão ganhando cada vez mais protagonismo, notavelmente pelo fato da inteligência artificial estar progressivamente se destacando na indústria e na academia. O aumento do poder computacional trata-se do facilitador na construção de modelos de elevado desempenho, e permite que o estado da arte da resolução de problemas complexos avance nas mais diversas áreas de aplicação.

A propulsão da adoção de sistemas especialistas inteligentes em áreas de risco se dá justamente devido às elevadas taxas de acuracidade atingidas, e um campo que vem adotando essas novas tecnologias trata-se da medicina, com aplicações que abrangem desde o diagnóstico de doenças, até a descoberta de novos fármacos [1].

Como a utilização de sistemas especialistas em áreas de grande risco trata-se de um tema sensível, passível de problemas éticos, políticos e que pode envolver riscos à vida, a comunidade científica começou a compreender a importância de se entender o raciocínio empregado pelos modelos criados, e não apenas suas saídas, e portanto inúmeros esforços estão sendo feitos na área que estuda a interpretabilidade de modelos, atualmente denominada por Explainable Artificial Intelligence (XAI).

Apesar disso, ainda não se chegou a uma definição exata do que se trata interpretabilidade, ou explicabilidade. No artigo "The Mythos of Model Interpretability"[2], os autores fazem uma proposta de formalização do termo interpretabilidade, apresentando propriedades de modelos denominados interpretáveis, tal como ser *transparente* e garantir *explicações post-hoc*, propriedades as quais garantiriam um maior entendimento acerca das resoluções propostas pelos modelos.

Os modelos comumente ditos interpretáveis tratam-se de modelos de regressão linear, árvores de classificação e regras de decisão, por diversos fatores, mas sobretudo devido a simplicidade. Contudo, é dito que existe um compromisso entre a performance de um modelo e a sua interpretabilidade [3], e dessa maneira, os pesquisadores devotam esforços para interpretar modelos caixa preta, buscando uma solução a esse compromisso.

Nas próximas seções, serão discutidos os modelos ditos interpretáveis, as propriedades desses modelos, e algumas abordagens utilizadas para superar o entrave do compromisso entre interpretabilidade e performance, mencionando uma abordagem de explicações locais agnósticas ao modelo [4].

Capítulo 2

Atividades Realizadas

2.1 Estudo de Modelos

Uma maneira de se familiar com a interpretabilidade de modelos, trata-se da exploração de exemplos de modelos ditos interpretáveis, e o entendimento dos mecanismos que os fazem interpretáveis. Neste capítulo, introduz-se os modelos mais comumente explorados, como a regressão linear e árvores de classificação. Posteriormente, serão feitas conexões dos modelos explorados com o estudo de interpretabilidade, também mencionando brevemente as redes Bayesianas e a análise LS/LN, uma vez que ambas são técnicas de raciocínio que sugerem causalidade entre evidências e hipóteses, que podem facilitar o entendimento do problema a ser resolvido.

2.1.1 Regressão Linear

O modelo de regressão linear tem como objetivo a criação de um hiperplano que é capaz de representar dados lineares. Isto é, supondo um conjunto de dados D_N de N observações,

$$D_N = \{(X_i, Y_i) : X_i, Y_i \in \mathbb{R}, i = 1, ..., N\} \tag{2.1}$$

o objetivo é encontrar os parâmetros w da função que representa o hiperplano, que pode ser encontrado através da minimização de uma função de perda arbitrária $L(y, \hat{y})$, e comumente utiliza-se a função de perda dada por

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.2}$$

denominada por resíduos quadráticos. Ao minimizar a função de perda, obtém-se os parâmetros w, e pode-se construir o hiperplano desejado através da função a seguir.

$$\hat{f}(x_i) = \hat{y}_i = w^T x_i \tag{2.3}$$

A derivação do problema $\hat{w}= \text{argmin}_w \ L(y, \hat{y})$ pode ser encontrado no capítulo 3.2 do livro "The Elements of Statistical Learning"[5].

Um exemplo ilustrativo de regressão linear pode ser visto na figura 2.1. Devido à simplicidade do modelo, nota-se facilmente como a saída dependerá de cada uma das variáveis, pois o aumento em uma unidade da variável X_i fará com que a saída aumente em w_i unidades, ou seja, a interpretabilidade aqui se trata de um conceito mais intuitivo e visual, que pode ser facilmente perdido quando se trabalha em dimensões muito grandes.

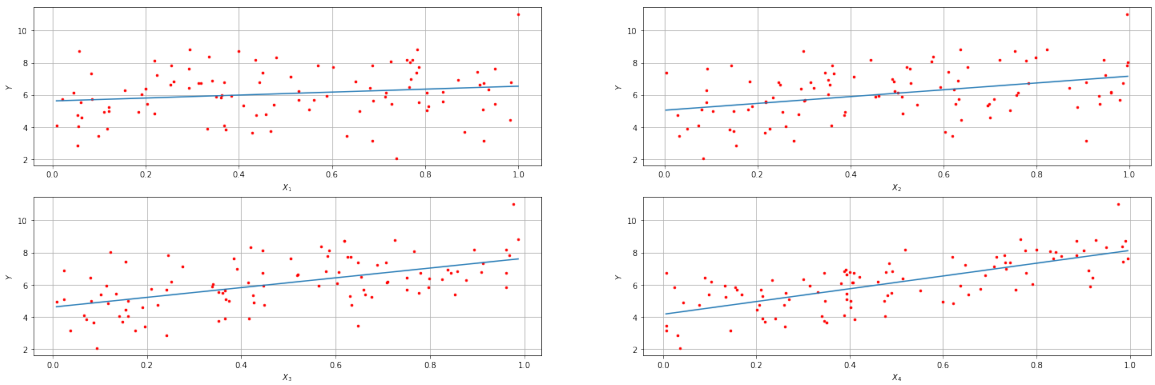

Figura 2.1: Regressão Linear

2.1.2 Árvores de Classificação

Ao contrário da regressão linear, as árvores de classificação são úteis para classificar dados, e representam funções capazes de receber um vetor de atributos em sua entrada, retornando uma decisão como saída. Os nós tratam-se de um atributo X_i, que terão regras definidas em suas arestas, as quais se ligarão a novos nós, até atingir as folhas da árvore. Nas folhas dessa árvore, encontram-se as decisões tomadas pela árvore de classificação, permitindo que as regras tomadas pela decisão sejam possíveis de visualizar, como mostra a figura 2.2, que representaria a decisão de esperar ou não em um restaurante.

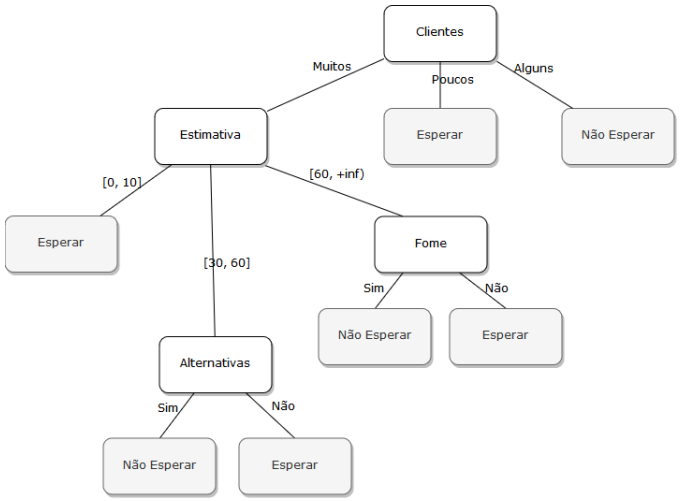

Figura 2.2: Árvore de Classificação

No entanto, apesar de árvores serem tratadas como interpretáveis, essa afirmação só é verdade para árvores não muito complexas, em que o número de nós e regras permite que a visualização não se torne confusa e atrapalhe o entendimento do modelo.

Árvores de classificação encontram-se na categoria de algoritmos de aprendizado supervisionado, em que a árvore é construída a partir de um conjunto de treinamento. Formalmente, seja $X = \{X_1, ..., X_p\}$ um conjunto de atributos, e x_{ij} o valor da i-ésima observação do preditor j, define-se o conjunto de treino como $D_{Treino} = \{(x_1, y_1), ...(x_n, y_n)\}$. A árvore de classificação que se deseja construir,

portanto, minimizará $P\left[\hat{y}(x) \neq y(x)\right]$, $\forall (x,y) \in D_{Treino}$, em que $\hat{y} \in \hat{Y}$ é a classificação feita pela árvore.

Para induzir a construção de uma árvore através de exemplos, adota-se uma estratégia gulosa de divisão e conquista, testando sempre o atributo de maior importância, cuja importância será calculada através de alguma métrica, comumente se tratando do *ganho de informação*. Recursivamente, repete-se este processo até expandir toda a árvore, exaurindo as opções de atributos e observações a serem selecionadas.

O ganho de informação se baseia na diminuição da entropia depois que um conjunto de dados é dividido em um atributo, formalmente,

$$\text{Ganho de Informação}(X_i) = \text{Entropia}(S) - \text{Entropia}(X_i), \text{ em que } S = \bigcup_{i=1}^{p} X_i \quad (2.4)$$

$$\text{e Entropia}(X) = \sum_{i=1}^{p} -\frac{|X_i|}{|X|} \log_2 \left(\frac{|X_i|}{|X|}\right).$$

2.1.3 Modelos Bayesianos

Existem três principais abordagens para se modelar problemas probabilísticos, a denominada abordagem clássica, a abordagem frequencialista (ou frequentista) e a abordagem bayesiana. Entre elas, não se pode dizer que existe uma abordagem superior a outra em termos absolutos, pois a utilização de cada qual irá depender do problema a ser modelado.

A abordagem clássica irá enxergar as probabilidades como sendo,

$$P(\text{Evento}) = \frac{\text{Eventos bem sucedidos}}{\text{Eventos mal sucedidos + Eventos bem sucedidos}} \quad (2.5)$$

sendo o exemplo mais trivial o jogo de moedas, pois a probabilidade da face de uma moeda não viesada ser cara será $P(Cara) = \frac{P(Cara)}{P(Cara)+P(Coroa)}$.

Na abordagem clássica, utilizada geralmente em experimentos teóricos, supõe-se condições experimentais perfeitas, como o não viés de uma moeda, e portanto os resultados são determinísticos e possuem fórmulas exatas.

A segunda abordagem, amplamente utilizada, trata-se da abordagem frequencialista. Nesta abordagem, as probabilidades são verificadas através de inúmeros ensaios de um mesmo experimento, traçando-se o comportamento das amostras. Pode-se utilizar a abordagem frequencialista em simulações computacionais, por exemplo, e é fácil notar que as soluções serão sempre aproximadas, ao contrário da abordagem clássica. Como essa abordagem é dependente de amostra, as probabilidades podem ser descritas como,

$$P(\text{Evento}) = \lim_{N \to \infty} \frac{f(\text{Evento})}{N} \quad (2.6)$$

Em que $f(\text{Evento})$ é a frequência observada de determinado evento. Nota-se, portanto, que essa abordagem é inadequada quando é impossível realizar eventos múltiplas vezes, apesar de ser muito difundida nesses casos.

Por fim, têm-se a abordagem Bayesiana, também conhecida como abordagem subjetivista. Nesta abordagem, as probabilidades não necessariamente se baseiam

na repetição de ensaios, baseando-se no conhecimento à priori do domínio do problema. As probabilidades neste caso podem ser escritas através da função de verossimilhança, de acordo com a equação 2.7.

$$\overbrace{P(H|E)}^{\text{À posteriori}} = \underbrace{P(E|H)}_{Verossimilhança} \times \frac{\overbrace{P(H)}^{\text{À priori}}}{\underbrace{P(E)}_{Normalização}} \qquad \text{H: Hipótese, E: Evidência.} \tag{2.7}$$

Nota-se da equação 2.7 que a probabilidade *à priori* é proporcional a probabilidade à posteriori, isto é, $P(H|E) \propto P(E|H)$. Essa relação possibilita que sejam feitas inferências sobre a causalidade de hipóteses e evidências, característica de suma importância no campo de interpretabilidade.

A seguir, enuncia-se um exemplo de problema de causalidade que pode ser resolvido utilizando-se a abordagem Bayesiana: Considere que um médico descobriu que seu paciente está com câncer de pele, contudo, o médico não conhece os hábitos de seu paciente. As causas do câncer de pele podem ser "fumar", "tomar sol"e "ter predisposição genética". Esse médico quer descobrir se seu paciente é fumante, e ele possui os seguintes dados,

$$P(Fumar) = 5\% \qquad P(TomarSol) = 40\% \qquad P(Cancer) = 2\% \tag{2.8}$$

e ele acredita que a probabilidade de ter câncer dado que um paciente é fumante é de cerca de 30%, portanto,

$$P(Fumar|Cancer) = P(Cancer|Fumar) \times \frac{P(Fumar)}{P(Cancer)} = 30\% \times \frac{5\%}{2\%} = 75\% \tag{2.9}$$

A fórmula da Bayes fornecida pela equação 2.7 pode ser utilizada para atualizar as probabilidades *à posteriori* sempre que novas informações do problema modelado aparecem. Sendo assim, é natural que se construam redes utilizando essas relações, e dessa extensão surgerem as redes Bayesianas, que resolvem problemas iguais ao resolvido acima pela equação 2.9, mas com múltiplas evidências e hipóteses.

Abaixo, ilustra-se a rede Bayesiana utilizando um conjunto de dados sobre a percepção política brasileira, em que se conhece todos os conjuntos de relações condicionais para se fazer inferências sobre o PIB brasileiro, e é possível realizar diversas simulações para verificar suas respectivas influências na saída. Apesar disso, redes Bayesianas muito complexas e com muitos nós recaem no mesmo problema da árvore de classificação, conforme a complexidade aumenta, seja considerando o aumento do número de nós ou de arestas, a noção de causalidade e interpretabilidade é esvaziada.

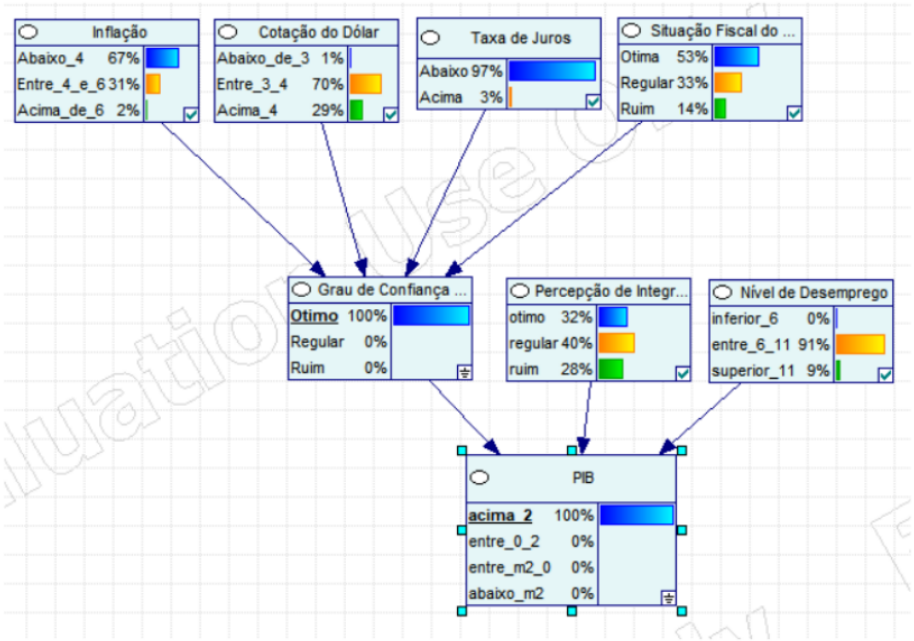

Figura 2.3: Redes Bayesianas

2.2 Propriedades em Interpretabilidade

Após mencionar alguns modelos que podem ser considerados interpretáveis, explora-se aqui alguns conceitos importantes em interpretabilidade, os quais são apenas direções a serem tomadas, uma vez que não existe um critério objetivo acerca do tema.

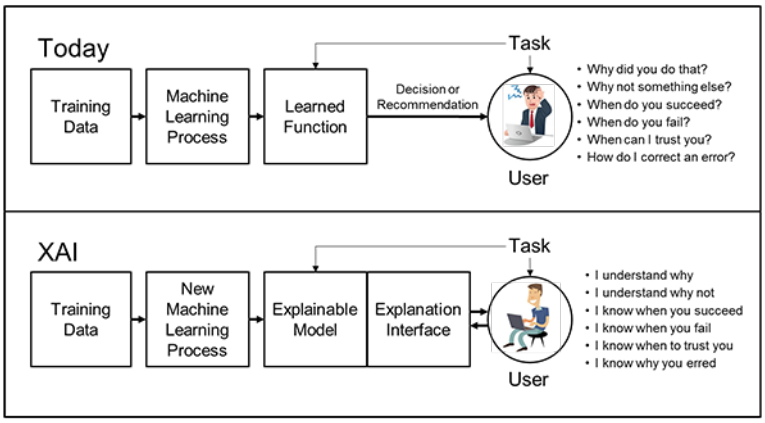

Figura 2.4: Conceito de XAI

Na figura 2.4, têm-se uma representação do fluxo de sistemas especialistas, no retângulo superior ilustra-se a utilização desses sistemas atualmente, em que um modelo caixa-preta recebe entradas do problema e devolve uma saída, deixando o usuário final sem entendimentos acerca do raciocínio do sistema. No retângulo de baixo, ilustra-se o fluxo de um sistema especialista interpretável, que permite o usuário final entender o raciocínio do sistema.

Contudo, não são claras as propriedades de modelos capazes de responder às perguntas "Como você chegou a essa conclusão?", "Por que não outro resultado?",

"Por que o sistema falhou?". Algumas das propriedades desejadas em modelos interpretáveis foram exploradas no artigo "The Mythos of Model Interpretability"[2], e a seguir serão citadas algumas dessas propriedades.

A primeira propriedade trata-se da *transparência*, que nesse contexto se opõe ao conceito de "caixa-preta". A transparência pode ser obtida através da facilidade de simulação desse modelo em novos contextos, através de seus parâmetros e também através de seu algoritmo. A regressão linear, por exemplo, apesar de não ser transparente em seu algoritmo (pois não há extração de informação útil na minimização de uma função perda), é transparente ao ser simulada, pois sabe-se exatamente o comportamento de sua saída ao fazer ajustes nas entradas do modelo.

A segunda propriedade trata-se da *capacidade de fornecer explicações post-hoc*, propriedade que se distancia das noções intrínsecas ao modelo, como seus parâmetros e algoritmo, e se aproxima mais aos aspectos textuais, visuais e explicativos fornecidos por este. Um exemplo de explicação *post-hoc* seria a utilização de modelos Bayesianos aliadas à análise LS/LN, pois a saída do modelo seria capaz de elucidar como a presença ou ausência das evidências e hipóteses influenciou a decisão a ser tomada.

Figura 2.5: Explicação post-hoc. A primeira imagem é a entrada de um classificador, o qual classificou na imagem a presença de um cachorro e de uma guitarra elétrica. As três últimas imagens são as explicações post-hoc da entrada.

Outro exemplo de explicação *post-hoc* seria a representação visual da saída, a figura 2.5 mostra uma técnica de interpretabilidade agnóstica a modelos (independente do modelo utilizado, consegue-se uma explicação *post-hoc*), que é capaz de inferir quais as características são levadas em consideração ao se produzir uma saída. Esse tipo de explicação satisfaz o segundo fluxo da imagem 2.4, e auxilia que os usuários confiem mais em sistemas especialistas. Essa técnica é demonstrada no artigo "'Why Should I Trust You?' Explaining the Predictions of Any Classifier"[4].

Apesar de diversos modelos possuírem essas propriedades, é comum que modelos interpretáveis possuam menor acuracidade, e esse aspecto da área XAI é denominado de compromisso entre acuracidade e interpretabilidade. A adoção de modelos caixa-preta se dá justamente devido a capacidade desses modelos serem paralelizados e conseguirem criar uma infinidade de relações complexas. A figura 2.6 ilustra o conceito do compromisso mencionado, e traça uma curva que pode ser atingida conforme as áreas XAI e inteligência artificial evoluem.

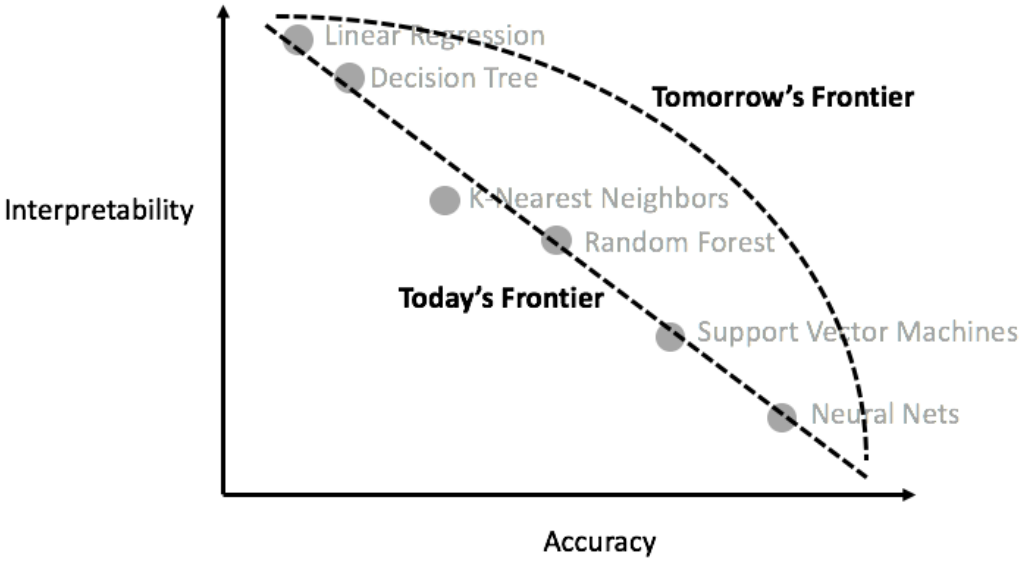

Figura 2.6: Compromisso entre acurácia e interpretabilidade

2.3 Bibliotecas sobre Interpretabilidade

As atividades nesse relatório referem-se majoritariamente à biblioteca denominada "Lime: Explaining the prediction of any machine learning classifier" [4], que tem por objetivo auxiliar a interpretação de quaisquer modelos de classificação, dependendo apenas das entradas e saídas.

2.3.1 Representações de Dados Interpretáveis

O autor menciona a importância de distinguir as representações de dados interpretáveis e atributos, uma vez que a representação de dados interpretáveis dependerá de uma análise subjetiva do usuário utilizador do modelo. Por exemplo, um vetor binário indicando a ausência ou presença de uma palavra, em um classificador de texto, poderia ser considerado uma representação interpretável, como mostra a figura 2.7.

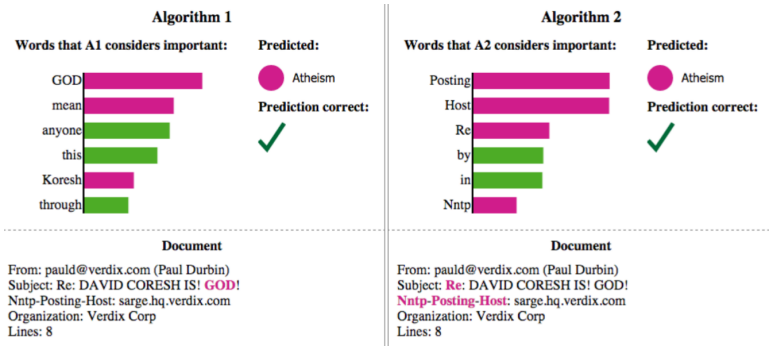

Figura 2.7: Explicações das predições de cada um dos classificadores, A1 e A2, respectivamente, os quais determinam se o documento é sobre "Ateísmo"ou "Cristianismo". As barras em verde representam palavras que contribuem mais a "Cristianismo", e as em rosa contribuem mais a "Ateísmo". O algoritmo A2 considera atributos como "Posting"e "Host"importantes.

Outro exemplo seria um vetor indicando a presença ou ausência de um superpixel em um classificador de imagem, que também possibilitaria a interpretação deste modelo, como visto na figura 2.5. Ou seja, não há definição sobre qual representação é interpretável ou não, e portanto cada representação dependerá do problema. Por fim, denota-se por $x \in \mathbb{R}^d$ a instância original do problema, e a representação interpretável é denotada por $x' \in \{0,1\}^{d'}$, que se trata de um vetor binário indicador de ausência ou presença de determinado elemento dos dados.

2.3.2 Criação do Modelo Explicativo

A biblioteca tem por objetivo criar um modelo explicativo $g \in G$, em que G é o conjunto de modelos *potencialmente* interpretáveis, isto é, modelos como os mencionados nas seções 2.1.1 e 2.1.2. Esses modelos são ditos *potencialmente* interpretáveis pelo fato de poderem ser extrinsecamente complexo, sejam os modelos de árvore havendo inúmeros nós e arestas, ou modelos de regressão linear contendo um elevado número de atributos. Considerando essas complexidades existentes, define-se uma função de complexidade, a qual,

$$\Omega : G \longrightarrow \mathbb{R} \tag{2.10}$$

um de regressão linear g, munido de p atributos, pode ter sua complexidade descrita por $\Omega(g) = p$. portanto, é possível desdobramentos mais abstratos para a função Ω e também para o conjunto de sua imagem. Aqui, devido a utilização dos modelos 2.1.1 e 2.1.2, a utilização do corpo dos reais se trata da escolha mais natural a ser feita.

Define-se agora o modelo de classificação f, o qual está sendo explicado por g, em que,

$$f : \mathbb{R}^d \longrightarrow [0,1] \tag{2.11}$$

e portanto $f(x)$ é a probabilidade atribuída pelo classificador f de uma instância x pertencer a alguma classe. Como o LIME utiliza a ideia de amostrar instâncias aleatoriamente, define-se uma função de proximidade $\pi_x(z)$, a qual induz uma distância entre as instâncias x e z.

Para dar completude ao modelo de explicação, necessita-se de uma função de confiança local acerca do classificador f, dada pelo modelo explicativo g. Daí, define-se $\mathcal{L}(f, g, \pi_x)$ como sendo a função de desconfiança na vizinhança de x. Por fim, para dar explicações à instância x, deve-se encontrar o modelo explicativo g, o qual pode ser obtido através da resolução do seguinte problema de otimização,

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2.12}$$

No artigo original [4], o autor define a função explicativa como sendo $g(z') = w_g \cdot z'$, e portanto G é o conjunto dos modelos lineares. No entanto, como mencionado anteriormente, outros arcabouços de modelos podem ser explorados. Outro ponto interessante a ser mencionado, trata-se da amostragem das instâncias z', que são feitas de acordo com uma distribuição uniforme e que são pesadas de acordo com a função π_x.

Contudo, a discussão sobre como aperfeiçoar a amostragem das instâncias z' é central no problema da interpretabilidade local, na figura 2.8, é exibida um exemplo

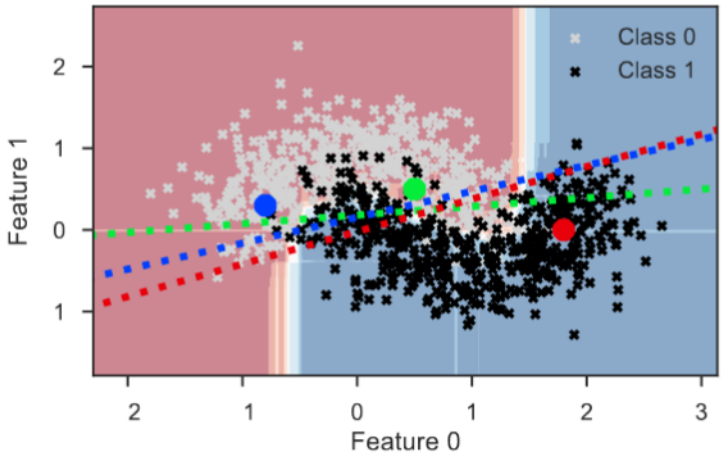

Figura 2.8: Aproximações lineares (linhas pontilhadas) geradas pelo LIME para 3 predições diferentes (círculos coloridos). O fundo colorido representa a fronteira de decisão de um classificador f.

de explicações locais que não condizem com a fronteira de decisão gerada por um classificador f, uma vez que os modelos lineares deveriam estar posicionados mais verticalmente em relação à figura. Esse tipo de fenômeno faz com que as explicações geradas por g não proporcione um bom grau de confiança ao usuário. No artigo [6] é feita uma proposta de amostragem das instâncias z', a qual é feita utilizando uma distribuição uniforme em uma hiperesfera, ao invés de uma distribuição uniforme linear.

Retornando à solução proposta no artigo [4], explicações denominadas linearmente esparsas podem ser feitas utilizando,

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \exp\left[-\frac{D(x,z)^2}{\sigma^2}\right] (f(z) - g(z'))^2 \tag{2.13}$$

em que $\mathcal{Z}$ é o conjunto das instâncias amostradas. Dessa forma, é possível gerar explicações como as apresentadas nas figuras 2.5 e 2.7, auxiliando a tomada de decisão do usuário ao utilizar modelos de classificação que outrora seriam difíceis de interpretar.

Por fim, apesar desse relatório focar majoritariamente na biblioteca denominada LIME, a qual compreendeu grande parte dos estudos do estudante, também menciona-se a biblioteca SHAP Values, que também se apoia na ideia de utilizar explicações locais para garantir interpretabilidade da saída de modelos de aprendizado de máquina. No entanto, os estudos de SHAP ficarão como uma próxima etapa de estudos do estudante, uma vez que esta biblioteca é capaz de agregar diversos conceitos de interpretabilidade.

Capítulo 3

Conclusão

Diversos conceitos de interpretabilidade em modelos de inteligência artificial ainda não se encontram bem definidos, e há um consenso na comunidade acadêmica de que é necessário explorar as dificuldades enfrentadas na interpretação de sistemas especialistas. Apesar da indefinição e ambiguidade do termo "interpretabilidade", diversas propriedades desejadas em modelos interpretáveis, como o fornecimento de explicações *post-hoc* e transparência de raciocínio, conseguem ser enxergadas como um caminho a ser seguido na área de XAI.

Alguns modelos previamente mencionados, como a regressão linear, árvores de classificação e redes Bayesianas, possuem intrinsecamente alguma das propriedades mencionadas, especialmente as redes Bayesianas, as quais não são tão difundidas como modelos de redes neurais devido ao seu desempenho inferior em tarefas de alta complexidade, recaindo ao compromisso entre acuracidade e interpretabilidade.

Atualmente, as pesquisas e descobertas em XAI ainda são difusas, mas como apresentado no gráfico da figura 2.6, esperam-se grandes avanços na acuracidade de modelos ditos interpretáveis, e também maior interpretabilidade em modelos de alta acurácia, sendo a técnica *post-hoc* "Local Interpretable Model-Agnostic Explanations"utilizada no artigo [4] um exemplo de como interpretar modelos caixa-preta minimizando a perda de acurácia.

Bibliografia

[1] Nature. Ascent of machine learning in medicine. Editorial Nature Magazine, 2019. URL https://doi.org/10.1038/s41563-019-0360-1.

[2] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. URL http://arxiv.org/abs/1606.03490.

[3] Defense Advanced Research Projects Agency. Broad agency announcement, explainable artificial intelligence (xai). URL https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf. DARPA-BAA-16-53 (DARPA, 2016).

[4] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[6] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. *CoRR*, abs/1806.07498, 2018. URL http://arxiv.org/abs/1806.07498.