I Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Avaliação da interpretabilidade de técnicas de aprendizado de máquina

Daniel de Oliveira Caires[1]
Fabrício Simeoni de Sousa [2]
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

1 Introdução

Atualmente técnicas de aprendizado de máquina estão sendo constantemente utilizadas em sistemas e aplicações que fazem parte do dia a dia das pessoas, seja para prover alguma informação relevante, apresentando sugestões, apoiar ou até mesmo para diretamente realizar decisões importantes para indivíduos ou corporações [3].

Com o peso das decisões baseadas nessas técnicas, vem também inúmeras preocupações sobre o seu funcionamento e que condições foram necessárias para levar aos resultados obtidos. Por esse motivo, a interpretabilidade dos métodos de aprendizado de máquina vem se tornando uma importante área de pesquisa [7]. A não explicação das decisões pode se tornar uma grande barreira para adoção destes métodos. Existem muitas preocupações quanto as decisões tomadas por modelos de aprendizado de máquina não apenas no que diz respeito a aspectos legais, mas também levando em consideração que o grande volume de dados coletados podem conter vários tipos de vieses, levando a decisões injustas e até mesmo erradas [4]. Modelos treinados com esses dados podem vir até a prover resultados que sejam preconceituosos ou discriminatórios [8].

A Figura 1 ilustra de maneira simplificada as etapas até se obter um modelo interpretável. Modelos mais interpretáveis são mais confiáveis, fáceis de depurar, ajuda a promover melhorias no processo de coleta de dados e geração de atributos além de dar melhor apoio e suporte a tomadas de decisão.

2 Aprendizado de máquina

O aprendizado de máquina em termos gerais pode ser definido pela aquisição de conhecimento, a partir de um conjunto de dados, para a realização de melhores predições. Diferente dos

[1]dielcaires@usp.br
[2]fsimeoni@icmc.usp.br

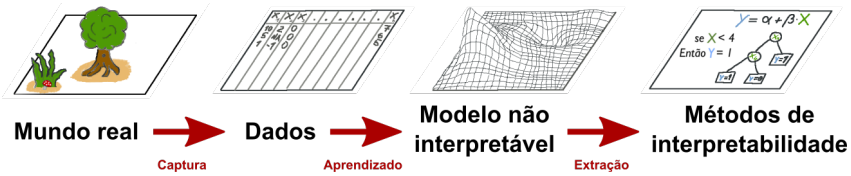

Figura 1: Representação das camadas para se obter modelo interpretável. Adaptada de [6]

algoritmos tradicionais, os algoritmos de aprendizado de máquina não possuem um conjunto de instruções pré-definidas. O que são providos, na verdade, são os dados e a partir deles o programa aprende os passos necessários para ponderar e estimar os parâmetros de acordo com a tarefa desejada [6]. Para isso normalmente é necessária uma etapa chamada de treinamento. Uma vez treinado, o modelo pode ser usado para fazer predições de novas instâncias de dados que forem inseridas. A Figura 2 apresenta de forma simplificada cada estágio de processo de aprendizado de máquina.

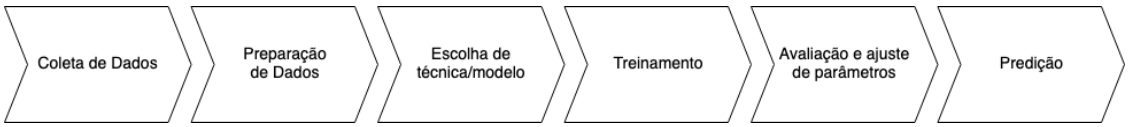

Figura 2: Ilustração mostrando as etapas de processo de aprendizado de máquina tradicional. Elaborada pelo Autor

Existem diferentes tipos de problemas que podem ser resolvidos por aprendizado de máquina. Casos onde o conjunto de dados históricos disponíveis para treinamento possui o atributo alvo e se deseja obter a informação desse atributo de interesse para novos dados, são chamados de problemas de aprendizado supervisionado. Para esses problemas, o objetivo do algoritmo é utilizar esses dados de entrada para aprender a mapear a saída desejada [1]. O que se obtém com esse aprendizado é um modelo, que uma vez definido pode ser utilizado em novas instancias de dados para fazer previsões.

Dentre os problemas que podem ser abordados com aprendizado supervisionado os mais comuns são a classificação, que é quando os rótulos possuem uma quantidade limitada de valores discretos, ou a regressão, que é quando possuem valores contínuos. Existem também problemas de aprendizado semi-supervisionado ou não supervisionados, que não serão o foco deste trabalho.

3 Interpretabilidade

Apesar de não possuir uma definição matemática, alguns autores propuseram definições. Interpretabilidade é grau o qual um humano consegue entender a causa que levou a uma determinada decisão [5]. No contexto de sistemas de aprendizado de máquina, é adicionada uma ênfase de prover explicações em termos entendíveis para seres humanos [2]. Apesar de não ser unânime entre os autores, diferenciam os termos interpretabilidade e explicabilidade, sendo esse segundo usado para se referir a explicações de predições individuais, afirmando ainda que modelos explicáveis são por consequência interpretáveis, mas o inverso não seria uma verdade [6].

Ao trabalhar com modelos de aprendizado de máquina não interpretáveis, o uso de métodos independentes (ou agnósticos) apresentam vantagens relevantes, sendo a flexibilidade a principal delas [7]. Muitas vezes os desenvolvedores utilizam mais de uma técnica de aprendizado para tratar um problema, e em sequência passam a avaliar algumas métricas para decidir com qual técnica seguir seu trabalho. Os métodos independentes são úteis nesse caso, pois podem ser aplicados em qualquer um desses modelos. A seguir serão apresentados os dois principais métodos estudados.

3.1 Gráfico de dependência parcial

O gráfico de dependência parcial mostra o efeito marginal de uma ou duas variáveis na predição de um modelo de aprendizado de máquina. Ele pode ser obtido simplesmente confrontando os valores de x, que seria o atributo cuja relação deseja ser analisada, com os valores dos respectivos $\hat{f}(x)$ que seria a variável resposta obtida com a previsão do aprendizado. Esse gráfico ajuda entender como o atributo influencia na variável resposta e se sua relação é linear, monotônica ou mais complexa, conforme ilustrado na Figura 3.

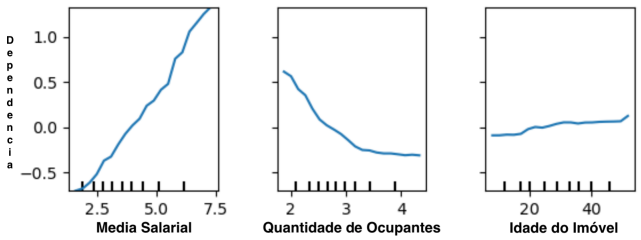

Figura 3: Exemplo de gráfico de dependência parcial de três variáveis relacionadas a mercado imobiliário.

3.2 *SHAP (SHapley Additive exPlanations)*

O SHAP é um método que prove explicação individual sobre predições, mostrando o impacto de cada atributo. É baseado nos *Shapley values*, uma técnica usada em teoria dos jogos que visa determinar quanto um jogador em um jogo colaborativo contribui para seu sucesso [6], mas no caso de uma predição ele funciona explicando o quanto cada variável contribuiu para o resultado final e se essa contribuição impactou positivamente ou negativamente como mostrado na Figura 4

4 Resultados esperados

Espera-se com o projeto alcançar maior profundidade na exploração de técnicas de interpretabilidade e aplicá-las em um estudo de caso real relacionado á areá de atuação do autor, afim de se propor uma metodologia mais consolidada para a interpretação de modelos de aprendizado de maquina que possa vir a ser usada em futuros projetos.

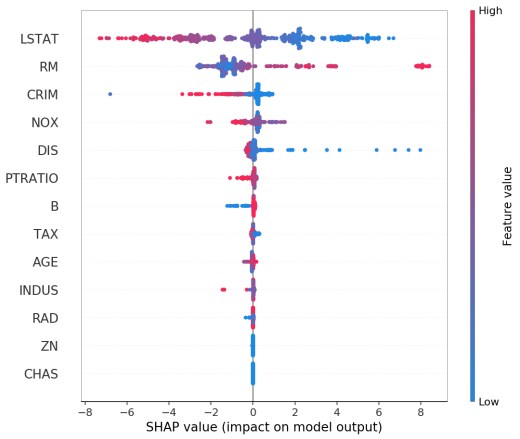

Figura 4: Exemplo de visualização gráfica da biblioteca SHAP, apresenta o efeito de cada atributo individualmente

Referências

[1] Alpaydin, *E.Introduction to Machine Learning. 2nd. ed. MIT press,* 2010.

[2] Doshi-Velez, F.; Kim, B. *Considerations for evaluation and generalization in interpreta-ble machine learning. Explainable and Interpretable Models in Computer Vision andMachine Learning. Springer,* 2018.

[3] Grgic-Hlaca, N.; Zafar, M. B.; Gummadi, K. P.; Weller, A. *The case for process fairness in learning: Feature selection for fair decision making. NIPS Symposium on Machine Learning and the Law.*, 2016.

[4] Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. *A survey of methods for explaining black box models. ACM computing surveys*, New York, 2018.

[5] Miller, T. *Explanation in artificial intelligence: Insights from the social sciences. Artificia lIntelligence, Elsevier,* 2019.

[6] Molnar, C. *Interpretable Machine Learning: A guide for making black box models explaina-ble.* 2019.

[7] Ribeiro, M. T.; Singh, S.; Guestrin, C. *Model-agnostic interpretability of machine learning* CoRR, 2016.

[8] Weller, A. *Transparency: motivations and challenges. Explainable AI: Interpreting, Explai-ning and Visualizing Deep Learning.* Springer, 2019.