O aumento do poder preditivo dos algoritmos de aprendizado vem acompanhado de maiores dificuldades de interpretação, pois devido a maior complexidade de tais algoritmos os vínculos entre as *features* e o valor predito é mais obscuro. Isto representa um desafio, uma vez que a interpretabilidade é especialmente importante em finanças, para estabelecer a confiança do modelo nas características corretas. A interpretabilidade também representa um problema prático que esbarra em questões éticas e regulamentares, cujos desdobramentos já começam a ser abordadas, por exemplo, pela Comissão Europeia (2020).

O conceito de AI interpretável (eXplainable AI) surge para permitir a interpretabilidade das técnicas de aprendizado de máquina, aumentando o conforto e a confiança das diversas instâncias de decisão das instituições nesses modelos. Entre as técnicas de interpretabilidade, de ML, a abordagem LIME e valor-SHAP têm sido mais abertamente reconhecidos como o estado da arte.

LIME

A técnica denominada modelo agnóstico interpretado localmente (LIME) procura aproximar um modelo de caixa-preta por um modelo local interpretável em cada previsão individual.

Ribeiro *et al*. (2016) sugerem que essa abordagem é aplicável em qualquer algoritmo de classificação, uma vez que o LIME é independente do classificador original. O LIME procura ajustar um modelo local usando amostra de dados semelhantes à observação que está sendo explicada, fornecendo explicações para a previsão relativa a cada observação. As explicações fornecidas pelo LIME para cada observação X são obtidas da seguinte forma:

Onde, G é a classe de modelos potencialmente interpretáveis, tais como modelos lineares e árvores de decisão: .

é a explicação do modelo.

: Medida de proximidade de uma instancia *z* de *x*.

: Medida de complexidade da explicação de .

O objetivo é minimizar a função de perda *L* localmente, sem lançar mão de pressupostos a respeito de *f*, assegurando a natureza agnóstica de *L*. Por sua vez, *L* é a medida agnóstica de quão infiel *g* é na aproximação de *f* na localidade definida por .

SHAP

Lundberg e Lee (2017) se inspiraram em conceitos de teoria dos jogos ao propor a metodologia SHAP de interpretação de modelos de aprendizado de máquina. A variabilidade das predições é dividida entre as covariáveis disponíveis, e a contribuição de cada variável explicativa para cada previsão de ponto pode ser avaliada independentemente do modelo subjacente (Joseph, 2019).

A metodologia SHAP (SHApley Additive exPlanation) procura expressar as previsões como se fossem combinações lineares de variáveis binárias, as quais descrevem se determinada covariável está ou não presente no modelo. O SHAP aproxima cada previsão por uma função linear das variáveis binárias , sendo as variáveis binarias e as quantidades de explicação, conforme:

onde M é o número de variáveis explicativas.

Somente modelos aditivos que satisfazem as propriedades de precisão local, *missingness* e consistência podem ser obtidos ao atribuir para cada variável um efeito , isto é, o valor de Shapley (Scott et al., 2018):

onde *f* é a explicação do modelo, são as variáveis disponíveis e são as variáveis selecionadas. A expressão demonstrada, para cada predição individual, o desvio dos valores de Shapley de sua média, ou seja, a contribuição da i-ésima variável.

Intuitivamente, os valores de Shapley são um modelo explicativo que se aproxima localmente o modelo original. Sempre que *x* for diferente de zero, o valor de Shapley também será (propriedade *missingness*), e se para m modelo diferente a contribuição de uma variável for maior, o valor de Shapley correspondente também será (propriedade de consistência).

Ambos LIME e SHAP obtêm os parâmetros para contribuição de cada *feature* em cada observação (explicação local). Entretanto, eles diferem no algoritmo que leva a tal resultado. A fim de ver qual abordagem é melhor para detectar a contribuição das variáveis no nível local, tentamos uma abordagem não supervisionada e verificamos se é possível agrupar observações empregando uma matriz de dissimilaridade construída sobre pesos LIME e valores SHAP, empregando a distância euclidiana padronizada como o base para agrupamento.

Resumo do Capítulo 9.6 de <https://christophm.github.io/interpretable-ml-book/shap.html>

Referencias:

Lundberg, M. Scott & Lee, Su-In. A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems 30 (NIPS 2017), p. 4766–4775, 2017. Disponível em: https://arxiv.org/abs/1705.07874

Joseph, A. Shapley Regressions: A Framework for Statistical Inference on Machine Learning Models, 2019. Disponível em: https://www.kcl.ac.uk/business/assets/pdf/dafm-working-papers/2019-papers/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.pdf.

Ribeiro, M., Singh, S., Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. Disponivel em: https://doi.org/10.1145/2939672.2939

Scott, L., Gabriel, E., Su-In, L. Consistent Individualized Feature Attribution for Tree Ensembles, 2018. Disponível em: https://arxiv.org/pdf/1802.03888.pdf.