Atividade 3 - Visão Geral de Explicabilidade

September 2, 2024

Tópicos Avançados em IC 2 - 2024.2

Universidade Federal de Pernambuco (UFPE)

Aluna: Bruna Zamith Santos

Professor: Dr. Ricardo Prudêncio

Data: 02/09/2024

P.S: As respostas a seguir são baseadas no livro "Interpretable Machine Learning", Christoph Molnar.

0.1 Em que situações, prover explicações de modelos não é importante?

Em alguns casos, não é relevante prover explicações de modelo, pois saber apenas a previsão do modelo é suficiente. Como por exemplo:

1. **Nenhum impacto significativo**: Alguns modelos podem não requerer explicações pois são usados em um ambiente de baixo risco - ou seja, erros não terão consequência graves (e.g., um sistema de recomendação de vídeos/músicas);
2. **O problema/método já é amplamente estudado e avaliado**: Algumas aplicações foram tão bem estudadas que os problemas do modelo foram resolvidos com a experiência prática. Por exemplo, o problema de reconhecimento óptico de caracteres.
3. **Risco de manipulação do sistema**: Em alguns casos, a interpretabilidade pode gerar riscos de usuários mal intencionados manipularem indevidamente o sistema. Nesses casos, a interpretabilidade não é desejável. Modelos de crédito são um exemplo.

0.2 Qual é a diferença entre explicações intrínsicas e explicações post hoc?

Métodos que geram explicações intrínsecas referem-se a modelos que são considerados interpretáveis devido à sua estrutura, como por exemplo árvores de decisão e modelos lineares (modelos "white-box").

Já os métodos que geram explicações post-hoc atingem à interepretabilidade depois que o modelo já foi treinado (modelos "black-box"). Exemplos são o SHAP, permutation feature importance, e outros.

0.3 O que é interpretabilidade local?

Interpretabilidade local refere-se à capacidade de entender quais fatores levaram um modelo de Machine Learning (ML) a prever um resultado específico para uma instância ou grupo de instâncias. Em outras palavras, a partir de uma instância ou grupo de instâncias, são geradas explicações que tornam os resultados compreensíveis para seres humanos. Isso contrasta com a interpretabilidade global, que busca explicações para o modelo como um todo, sem focar em instâncias específicas.

Um exemplo de interpretabilidade local ocorre quando um cliente de banco deseja saber por que seu pedido de crédito foi negado por um modelo de ML, ou quando um paciente quer entender por que um modelo de ML o diagnosticou com uma determinada doença. Um exemplo de método de interpretabilidade local é o LIME (Local-Intepretable Model agnostic Explanations).

0.4 O que são métodos agnósticos para explicabilidade?

São métodos que independem do algoritmo usado para treinar o modelo. Ou seja, podem gerar explicações para qualquer modelo de ML - isso porque eles tratam modelos como "black-boxes" e, assim, não assumem quaisquer parâmetros ou estruturas intrínsecas.

0.5 Considerando as propriedades de explicações individuais, qual a diferença entre acurácia e fidelidade de explicações?

- **Acurácia**: Refere-se à capacidade de uma explicação prever dados que ainda não foram vistos. Em outras palavras, mede o quão bem a explicação pode ser usada para prever novos dados, especialmente se ela estiver sendo usada no lugar do modelo de ML.
- **Fidelidade**: Avalia quão bem a explicação se aproxima das previsões feitas pelo modelo de ML. A fidelidade é crucial, pois uma explicação com baixa fidelidade não serve para explicar adequadamente o modelo.

Embora acurácia e fidelidade estejam relacionadas, a fidelidade é mais focada em quão bem a explicação reflete as decisões do modelo, enquanto a acurácia se concentra na previsão de dados novos. Além disso, algumas explicações podem ter fidelidade apenas local, ou seja, aproximam bem as previsões do modelo apenas para um subconjunto dos dados ou para uma instância específica.

Portanto, a principal diferença entre acurácia e fidelidade é o foco de cada uma. A acurácia é sobre a capacidade preditiva da explicação, enquanto a fidelidade é sobre o quão bem a explicação replica as previsões do modelo original. Uma explicação com alta fidelidade garante que estamos entendendo corretamente o que o modelo faz, mesmo que essa explicação não tenha necessariamente uma alta acurácia preditiva em dados novos.