Repositório

Tópicos Avançados em Inteligência Computacional 2

Atividade 02

[Explicabilidade (Visão Geral)]

descrição:

Leia o capítulo Interpretability do livro "Interpretable Machine Learning" (de Christoph Molnar) e responda às seguintes perguntas:

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

2. Qual é a diferença entre explicações intrínsicas e explicações post hoc?

3. O que é interpretabilidade local?

4. O que são métodos agnósticos para explicabilidade?

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

[tarefas a serem feitas]

1. ~~Pergunta 1.~~
2. ~~Pergunta 2.~~
3. ~~Pergunta 3.~~
4. ~~Pergunta 4.~~
5. ~~Pergunta 5.~~

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

R.:

1.1. Em cenários aonde existe um ambiente de baixo risco, significando que o erro não haverá serias consequencias, ou seja, não tem um impacto significante.

1.2. Outro cenário, é quando o problema é muito bem definido e estudado, aonde existe aplicações suficientemente estudadas e desenvolvidas ao longo do tempo capazes de justificar, sendo suficientemente para explicar e resolver a problematica.

Em geral, em situações aonde o foco principal é saber **oque** é predito, esta a frente do **porque** é predito.

2. Qual é a diferença entre explicações intrínsicas e explicações post hoc?

R.:

2.1 Explicações **intrínsicas** são referentes a explicações que podem ser facilmente obtidas e interpretadas apartir de modelos com estruturas mais simples como modelos lineares ou arvores de decisões. Um

exemplo são os coeficientes estimados para cada variavel.

2.2 Explicações **post hoc** são explicações obtidas apos o treinamento do modelo, em especial para modelos com estruturas mais complexas e não facilmente interpretavel. Um exemplo seria o método de *Permutation Feature Importance* e *Shapley Values*.

3. O que é interpretabilidade local?

R.:

Interpretabilidade local trata-se de uma forma de entender, mais agradavelmente, o porque um modelo realizou um determinado tipo de predição, podendo ser de uma visão para instâncias individuais ou para um grupo de instâncias.

3.1 Instância simples: o intuito é analisar para uma instância pontual e entender uma possivel dependência linearmente ou monotonamente da predição com um numero menor de atributos.

3.2 Grupo de instâncias: pode ser aplicador métodos globais, a nível modular ou utilizando da interpretabilidade local para instâncias simples. É possivel aplicar métodos globais selecionando um subgrupo de instâncias, considerando como se fosse um conjunto completo.

Explicações locais tendem a ser mais acuradas do que explicações globais.

4. O que são métodos agnósticos para explicabilidade?

R.:

Métodos agnósticos são métodos que podem ser utilizados por qualquer modelo de aprendizado de máquina (caixa preta), podendo ser aplicados também como métodos *post hoc*, ou seja, após o treinamento.

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

R.:

5.1 Acurácia tem como foco avaliar o desempenho do modelo em dados não vistos, ou seja, o poder de generalização do modelo para novos dados.

5.2 Fidelidade de explicações foca em tentar entender o quão bem a explicação aproxima-se das predições do modelo *black box*.

A diferença entre elas dar-se pois a acurácia tem como foco na precisão para previsões de dados novos e a fidelidade foca na capacidade das explicações em conseguir reproduzir previsões, com fidelidade, do modelo *black box*.

Acuracia e fidelidade são muito proxímas, caso o modelo possua uma alta acuracia e uma alta explicação, consequentemente a explicação possui uma alta acurácia.