

Teste de Software Escola de Artes Ciências e Humanidades

Professor Mestre: Adilson Lopes Khouri

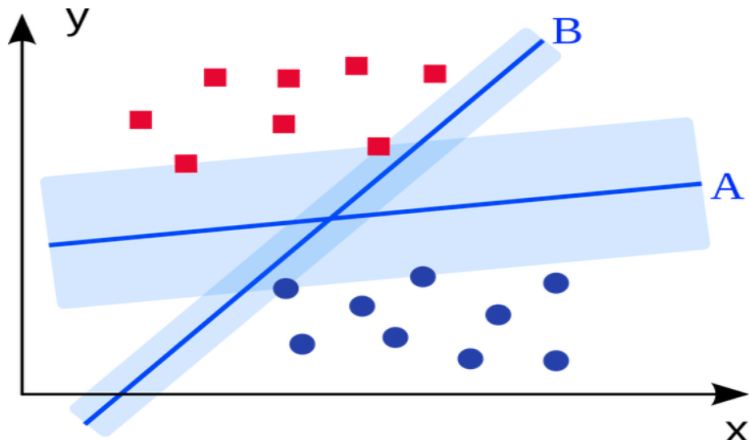
30 de junho de 2019

Sumário

- 1 Contextualização
- 2 Lime
- 3 Shap
- 4 Manifold
- 5 Metamorphic
- 6 Agradecimentos
- 7 Contato

Contextualização

- 1 Classificadores definem uma superfície de decisão para classificar dados (em até n-classes)



Contextualização

- 1 Os classificadores escolhem uma das retas azuis possíveis (há infinitas possibilidades) para ser a fronteira de decisão.
- 2 O problema acima é utópico, tipicamente as classes não ficam separadas perfeitamente.

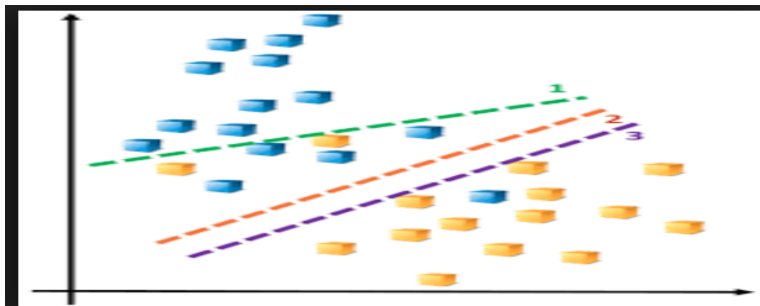


Figura: Exemplo real de fronteira de decisão

Contextualização

- 1 Para definir a fronteira é usado um algoritmo de otimização (que minimiza o erro de classificação)
- 2 **É considerado normal ter erros de classificação dado a natureza não determinística dos algoritmos de Machine Learning**
- 3 Quando ocorre um erro de classificação em algoritmos de machine learning podem ter várias razões (não mutuamente exclusivas):
 - Dados de treino insuficientes/viesados
 - Arquitetura do algoritmo mal planejada
 - O algoritmo aprendeu a função errada
 - Bug no código fonte
 - Variação na distribuição das principais variáveis usadas pelo modelo
 - Bug em outros sistemas que alimentam os dados do modelo

Contextualização

- 1 Além dos problemas citados há modelos caixa preta (não é possível entender como o modelo toma decisões[de onde foi criado o score?])
- 2 Regressão logística e CART são exemplos de algoritmos interpretáveis (é possível entender como foi tomada a decisão)

Contextualização

- 1 Dados todos esses problemas citados há linhas de pesquisa para cada um deles.
- 2 Os artigos LIME, SHAP e manifold tentam resolver o problema de quais variáveis são mais relevantes para modelos caixa preta.
- 3 O último artigo trata sobre identificação de bugs em código fonte de algoritmos de Machine Learning

Artigo 01

- 1 “Why Should I Trust You?” Explaining the Predictions of Any Classifier

Lime - overview

- 1 É agnóstico (não assume nada sobre o modelo, dessa forma, funciona para qualquer modelo)
- 2 Realiza uma interpretação Local do score gerado pelo modelo.
- 3 Um ponto falho é que a interpretação foi avaliada usando modelos pequenos (até 10 features) interpretáveis (explicar por que é falho)

Lime - funcionamento

- 1 Escolhe uma instância que deseja interpretar
- 2 Seleciona outras instâncias próximas a ela (usando uma métrica de distância D)
- 3 Usando a amostra executa uma logística para aprender, localmente, quais as variáveis mais importantes

Lime - fronteira de decisão

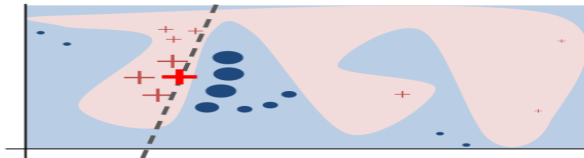


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Figura: LIME fronteira de decisão

Artigo 03

1 Consistent Individualized Feature Attribution for Tree Ensembles

Shap - overview

- 1 Não é agnóstico (funciona apenas para ensemble de árvore de decisão)
- 2 Realiza uma interpretação Local do score gerado pelo modelo.
- 3 Citam o problema de inconsistência (score alto para variáveis que não são importantes e o caso contrário)
- 4 Um ponto falho é que o SHAP foi avaliado perguntado para algumas pessoas (os autores não citam esse número) quais features eram mais relevantes.
- 5 Os autores não citam dados sobre o modelo como: número de features, parâmetros ótimos e nem qual modelo foi usado.

Shap - funcionamento

- 1 Realiza uma perturbação do dado de entrada, roda um modelo aditivo sobre o dado original e sobre o perturbado.
- 2 Interpreta o modelo aditivo para dizer o quanto a variável foi explicativa para aquela decisão
- 3 A diferença em relação ao LIME é que o modelo aditivo encontra o ϕ (equivalente ao beta da regressão) com probabilidade condicional, a logística usa uma técnica de otimização (e.g. Least square)

Artigo 02

- 1 Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models

Manifold - overview: TODO colocar figuras aqui

- 1 É um framework para ajudar o cientista a localizar falhas no modelo, para tal, os autores automatizam tarefas típicas para depurar modelos.
- 2 Para tal, os autores propõe criar uma “matriz de confusão” que relaciona modelos e classes, com essa matriz é possível avaliar visualmente onde os modelos não concordam.
- 3 Após selecionar a célula (tipicamente será escolhida a células de diferença Q1, Q3) é exibida uma comparação da distribuição de variáveis das instâncias daquela célula.
- 4 No proximo gráfico é exibida a distribuição das features, das instâncias selecionadas, e das instâncias pertencentes a cada classe.
- 5 A idéia é que as features com mesma ditribuição (na amostra e em todos os dados) são as mais relevantes para aquela decisão tomada.

Manifold - funcionamento

- 1 Realiza uma perturbação do dado de entrada, roda uma logística sobre o dado original e sobre o perturbado.
- 2 Interpreta a logística para dizer o quanto a variável foi explicativa para aquela decisão

Artigo 04

- 1 Identifying Implementation Bugs in Machine Learning Based Image Classifiers using Metamorphic Testing

Metamorphic - overview

- 1 Uso do conceito de teste metamórfico para testar algoritmos de machine learning.
- 2 Solução não agnóstica dado que há necessidade de definir relação metamórfica para cada novo modelo
- 3 Os autores conseguem validar a solução criando bugs artificiais em algoritmos conhecidos de machine learning (como SVM) e aplicando a solução proposta.
- 4 Não há uso de bugs reais para avaliar o algoritmo

Metamorphic - funcionamento

- 1 São definidos relações metamórficas para cada um dos dois algoritmos de machine learning
- 2 Os autores alteram o fonte dos algoritmos para chumbar uma semente aleatória fixa para todas as execuções
- 3 As relações metamórficas não devem causar alteração no output e função de perda no decorrer do treino
- 4 Com essas condições asseguradas, se ocorrer uma alteração na função de perda ou output há indício de bug

Fim!

Agradeço a atenção

Contato

- E-mail: *adilson.khour.usp@gmail.com*
- Phone: +55119444 – 26191
- Link LinkedIn
- Link Curriculum Lattes
- GitHub pessoal