EPI5717: Machine learning para predições em saúde

Aula 15

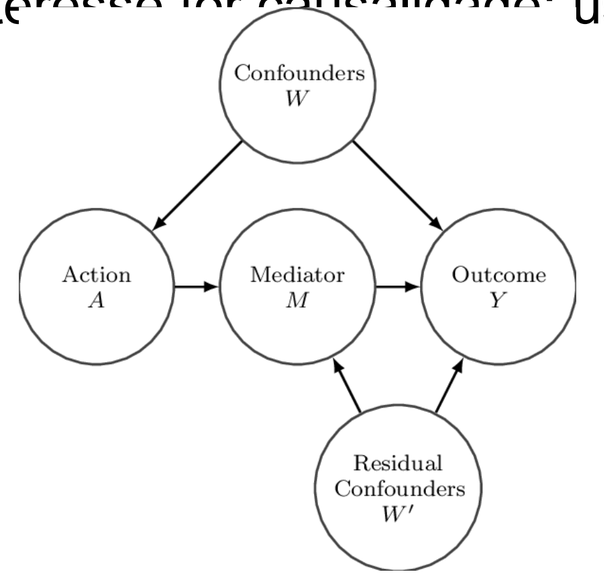
Prof. Dr. Alexandre Chiavegatto Filho

A BUSCA POR ALGORITMOS INTERPRETÁVEIS

Muito se fala sobre necessidade de “interpretação” dos algoritmos.

- O objetivo de uma predição **não é compreender a causalidade de um fenômeno**, é predizê-lo!

- Se o interesse for causalidade: usar métodos causais.



A BUSCA POR ALGORITMOS INTERPRETÁVEIS

Entretanto é sim possível interpretação da importância das variáveis em algoritmos preditivos.

- Importante: não confundir com causalidade.

- Exemplo: na predição de risco de uma pessoa ir a óbito talvez seja interessante incluir o fato de ela ter sido internada em UTI recentemente.

- O fato de ela ter ido para a UTI **não é a causa** de ela ir a óbito no futuro, é apenas um preditor (ninguém cogita impedir idas à UTI para diminuir óbito).

A BUSCA POR ALGORITMOS INTERPRETÁVEIS

Conclusão: *interpretação em machine learning não é causalidade, mas sim uma forma de entender como o algoritmo realizou a predição.*

A BUSCA POR ALGORITMOS INTERPRETÁVEIS

Por que interpretação:

- Identificar presença de preditores indesejáveis (vazamento de dados).

- Auxiliar na adesão pelos profissionais (médicos aceitarão melhor se entenderem como toma a decisão).

- Garantir maior robustez (se o algoritmo de carro sem condutor está identificando motos pela roda, cuidado com motos com bolsas laterais).

- Identificar preconceitos (pode ser que o algoritmo esteja usando raça para rejeitar empréstimos bancários).

A BUSCA POR ALGORITMOS INTERPRETÁVEIS

Por que **não** ter interpretação:

- Manipular o sistema:

- Imaginem se forem divulgados como um algoritmo estabelece prioridades para receber cirurgia, e uma das variáveis for morar no Butantã (bairro menos poluído de SP). As pessoas vão começar a dizer que moram no Butantã para manipular o algoritmo.

A BUSCA POR ALGORITMOS INTERPRETÁVEIS

Possibilidades para interpretação:

- Interpretação intrínseca: utilizar algoritmos interpretáveis (regressão linear/logística ou árvores simples de decisão).

- Interpretação extrínseca: utilizar técnicas que permitem retirar interpretação de algoritmos complexos após o treinamento.

IMPORTÂNCIA DE VARIÁVEIS PREDITORAS

Solução mais comum

- Análise da mudança do erro de predição ao **permutar valores** da variável.
  - Variável é importante para predição se erro aumenta.
  - Se o modelo não utiliza essa variável o erro não muda.

- Outras soluções utilizando estratégias um pouco mais complexas (LIME, Shapley).

INTERPRETABLE MACHINE LEARNING - A BRIEF HISTORY, STATE-OF-THE-ART AND CHALLENGES

Authors: Christoph Molnar, Giuseppe Casalicchio, Bernd Bischl.

Resumo

- Uma breve história do machine learning interpretável
- Estratégias principais: observar diretamente os componentes, analisar o efeito de pertubações nos preditores e analisar aproximações locais
- Desafios
- Aceitação da interpretabilidade pela comunidade científica

Uma breve história do machine learning interpretável

- Interpretabilidade é um fator decisivo
- Podem ser usados para entender ou justificar as predições
- Muitas pesquisas nos últimos 2 anos
- Modelos de ML geralmente têm uma abordagem não-linear e não-paramétrica (controle de hiperparâmetros com CV)
- Resultado: modelos com bom desempenho, mas menos "interpretáveis"

Uma breve história de machine learning interpretável

SVM, Árvores de decisão

1974

Redes Neurais

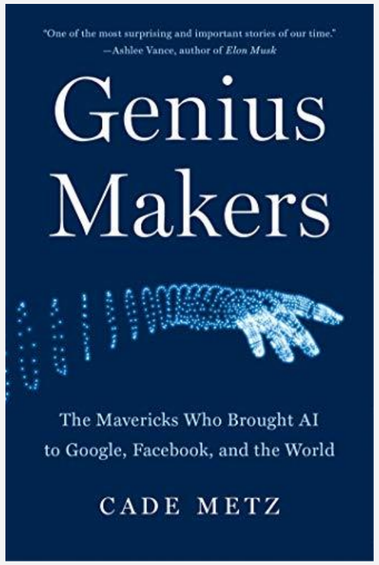
1960

Algoritmos de boosting

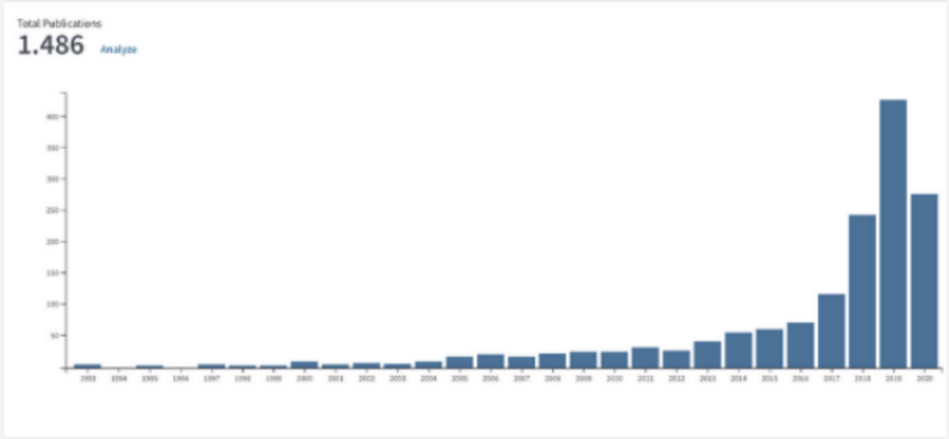
1990

Redes Neurais Profundas

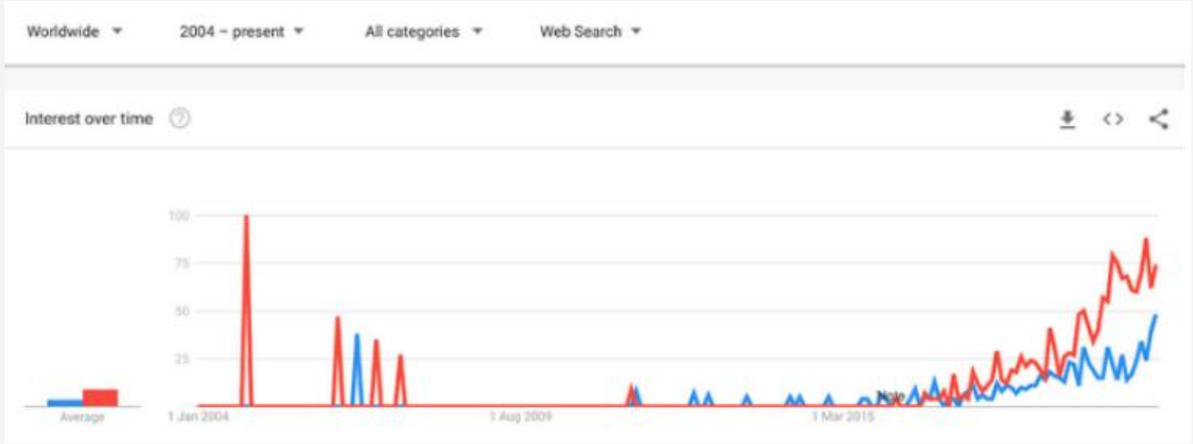
2010



Uma breve história do machine learning interpretável



Número de artigos com as palavras-chave (Web of Science): "Interpretable Machine Learning" ou "Explainable AI"



Número de artigos com as palavras-chave (Google search trends): "Interpretable Machine Learning" ou "Explainable AI"

O que temos hoje...

- Amadurecimento em termos de métodos e definição de interpretabilidade
- Há melhor compreensão dos pontos fracos dos métodos
- Softwares de código aberto para implementação de métodos
- Regulamentações – LGPD, Marco da Inteligência Artificial (PL 21/2020)
- Necessidade de confiabilidade, transparência e justiça
- Existem startups que focam em interpretabilidade de ML

Métodos para machine learning interpretável

- Análise de componentes do modelo
- Sensibilidade do modelo a perturbações
- Modelos substitutos/aproximação

Fig. 2. Some IML approaches work by assigning meaning to individual model components (left), some by analyzing the model predictions for perturbations of the data (right). The surrogate approach, a mixture of the two other approaches, approximates the ML model using (perturbed) data and then analyzes the components of the interpretable surrogate model.

Analisando componentes de modelos interpretáveis

- O modelo precisa ser decomposto em partes (interpretáveis)
  - Sem a necessidade da compreensão total do modelo
- Análise está ligada a estrutura do modelo (estrutura e parâmetros aprendidos)
- Modelo de Regressão Linear e árvores de decisão são considerados interpretáveis
- Em cenários de alta dimensão pode chegar a um ponto não mais interpretável (mesmo em modelos lineares – interações de variáveis)

Analisando componentes de modelos mais complexos

- Rede neural convolucional profunda (CNN) – imagens que mapeiam os componentes
- Random Forest – analisar a estrutura da árvore para quantificar importância das variáveis
  - A desvantagem de análise de componentes é estar ligada a esses modelos específicos
- Não funciona bem para as abordagens atuais de ML, onde se trabalha com diversos modelos de ML

Sensibilidade do modelo a pertubações

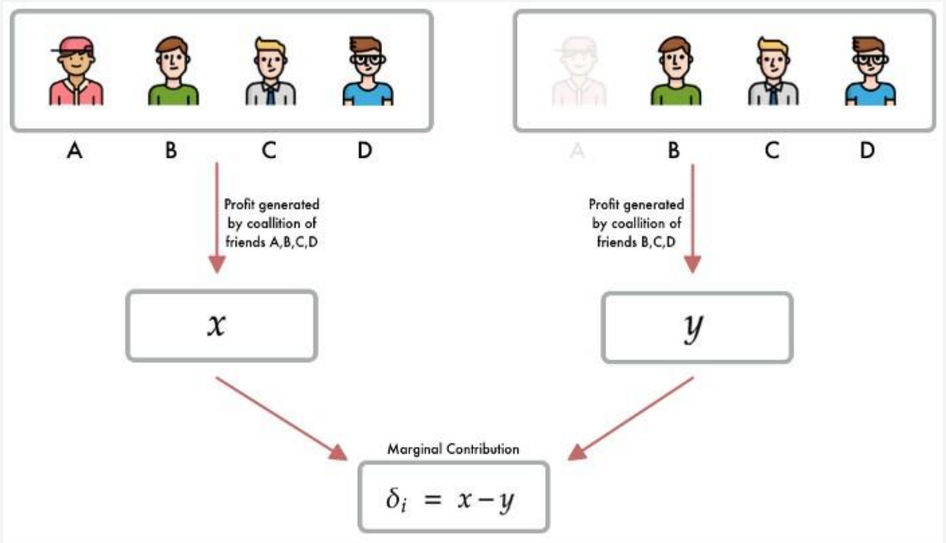
- Em sua maioria agnósticos em relação ao modelo
- Manipular os preditores e analisar como a predição muda
- Dividem-se entre locais e globais
- Locais: explicam predições individuais dos modelos
- Mais populares:
  - Explicações contrafactuais: cenários hipotéticos/filosofia
  - Shapley Values: colaboração/teoria dos jogos

Explicando predições individuais

- SHAP (SHapley Additive exPlanations)

- Baseado na teoria dos jogos
  - O objetivo é identificar a contribuição individual de cada jogador.
  - O valor de Shapley nos diz quanto cada variável afasta a predição da predição média.

- O valor de Shapley é a contribuição média marginal de cada variável ao longo de todas as coalizões possíveis.



https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

Explicando predições individuais

- SHAP (SHapley Additive exPlanations)

- Criar o conjunto de todas as combinações de valores existentes de variáveis (chamadas coalizões).

- Calcular a predição média de cada modelo.

- Para cada coalizão, calcular a diferença entre a predição do modelo com valor de F de outra observação e a predição média.

- Para cada coalizão, calcular a diferença entre a predição do modelo com esse mesmo valor de F e a predição média.

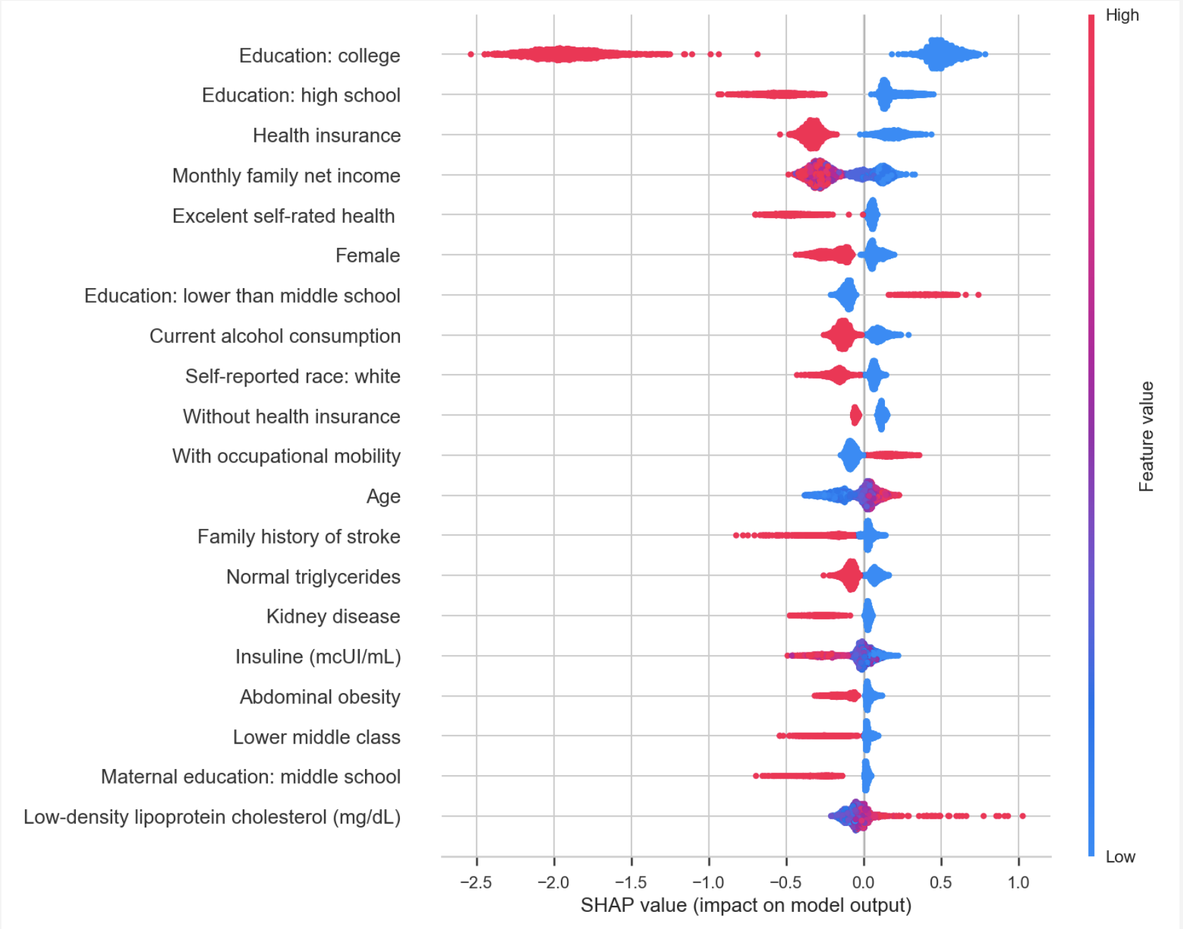
- Para cada coalizão, calcular quanto F mudou a predição do modelo da média (ou seja, passo 4 – passo 3) – essa é a contribuição marginal de F.

- Valor de Shapley = a média de todos os valores calculados no passo 5 (ou seja, a média das contribuições marginais de F)

https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

Explicando predições individuais

- SHAP (SHapley Additive exPlanations)



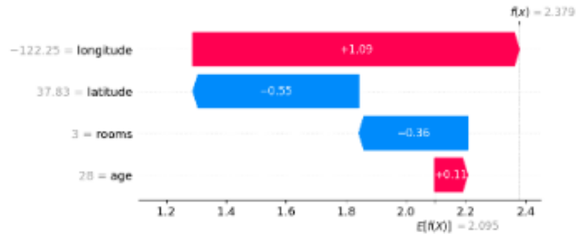
SHAP Plots For Tabular Data – Interpretation Cheat Sheet

Quick Guide

- Explain single prediction ⇒ waterfall or force plot
- Overall effects + importance ⇒ beeswarm plot
- Feature effect in detail ⇒ scatter plot

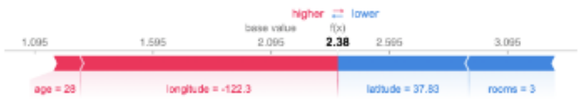
Example: Explain model that predicts house value (in $100k) from age, longitude, latitude an no. of rooms.

Waterfall Plot (Single)



- Visualizes Shapley values $\phi_{i,j}$ as arrows that either increase or decrease prediction $f(x_i)$ compared to expected prediction $E[f(x)]$.
- *Example*: The predicted value ($238k) of house #3 is larger than the average $209.5k. Longitude of -122.3 and, to lesser extent, age of 28 increase prediction; latitude of 37.83 and rooms=3 decrease predicted value, but don't cancel out the positive effects.

Force Plot (Single)

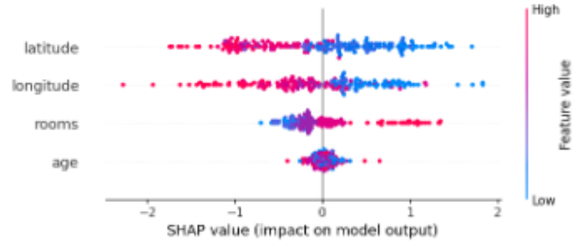


Rearrangement of arrows from waterfall plot.

Bar Plot (Single)

Also a rearrangement of arrows from waterfall plot. But no space left for the plot here. Sorry.

Summary / Beeswarm Plot (Multi)

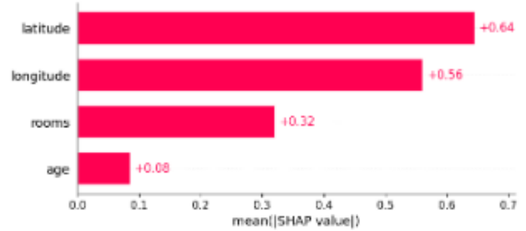


Each point is a Shapley value $\phi_{i,j}$. Color indicates the value of the feature. *Example*: latitude and longitude have the highest spread of Shapley values, which makes them the most important features; the lower the latitude/longitude, the higher the Shapley value and therefore the predicted house value.

Force Plot (Multi)

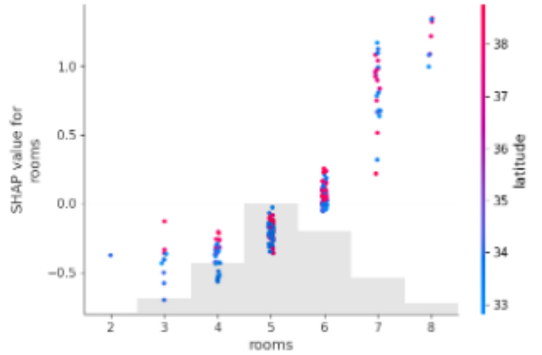
Individual force plots vertically put together. Difficult to interpret; better use summary plot.

Importance / Bar Plot (Multi)



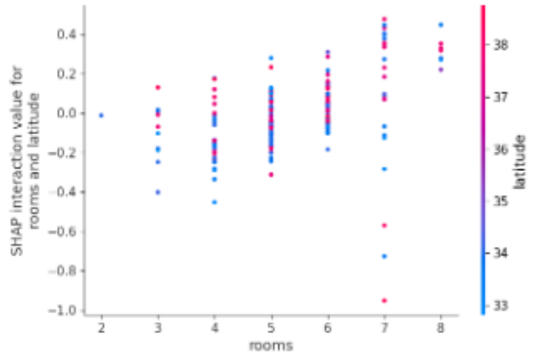
Summary plot reduced to importance: mean of $|\phi_j|$

Dependence / Scatter Plot (Multi)



Plot shows how Shapley values $\phi_{i,j}$ change with increasing value of feature j. 1 dot per data instance. Color based on the feature k that interacts most with j. *Example*: The higher the number of rooms the higher the associated Shapley value and therefore the predicted house value.

Interaction / Scatter Plot (Multi)



Plot shows how *interaction Shapley values* $\phi_{i,jk}$ change with increasing feature j (x-axis); color based on feature k. Interaction = combined effect on prediction after accounting for individual effects. *Example*: More rooms means higher house value due to interactions with latitude.

Explicando o modelo de comportamento global

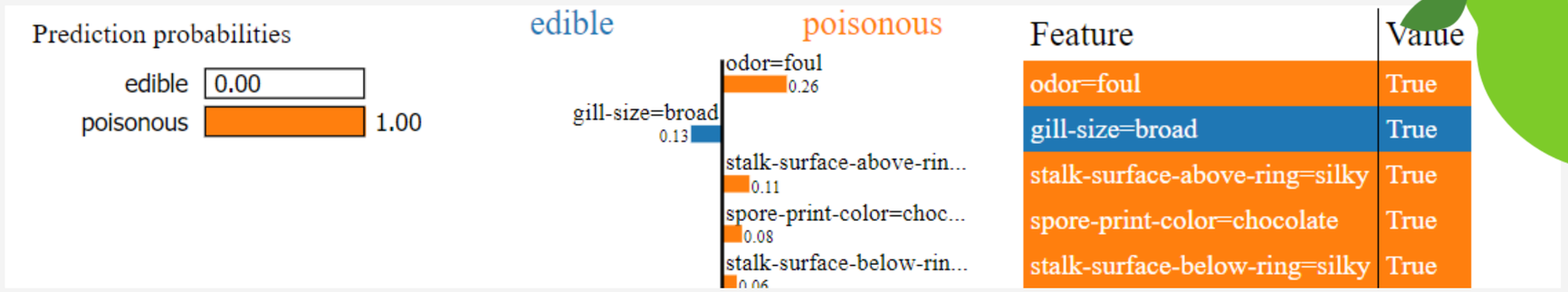
- Comportamento esperado médio em um banco de dados
- Importância da variável é o seu efeito total na predição
- Algumas medidas de importância utilizam a exclusão de variáveis do treino e subsequente retreino do modelo
- Efeito da variáveis é como uma mudança na variável afeta a predição do desfecho

Modelos substitutos

- Projetados para copiar o comportamento do modelo original de ML
- Interpretação baseada na análise dos componentes do modelo substitute (que é interpretável)
- Ex.: regressão linear com as mesmas variáveis.
- Frequentemente são usados métodos para extrair regras de decisão (ex.: pesos na NN)

Modelos substitutos

- LIME é um método substituto local
  - Gera um novo banco de dados com valores alterados das variáveis para ver predição.
  - Modelo interpretável (e.g. regressão linear) a partir desses resultados.



Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Desafios

Incerteza estatística e inferência

- Muitos métodos fornecem explicações sem quantificar a incerteza
- Os modelos estão sujeitos à incerteza
- Precisamos fazer suposições estruturais ou distributivas
- Essas suposições precisam ser testadas

Desafios

Interpretação causal

- Um modelo deve refletir a estrutura causal para ter uma interpretação causal
- Desempenho preditivo x causalidade

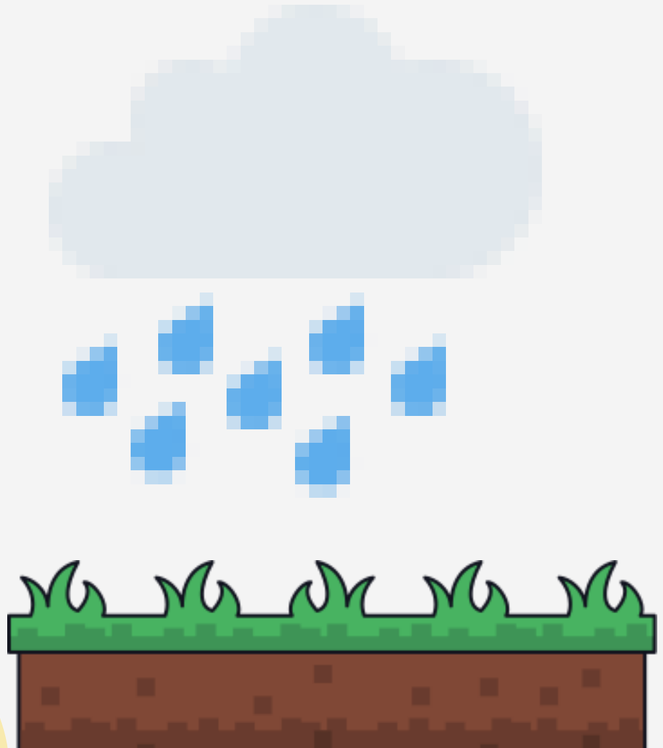
Exemplo: O tempo hoje causa o tempo de amanhã.

Mas...

Desafios

continuando...

Tempo Hoje



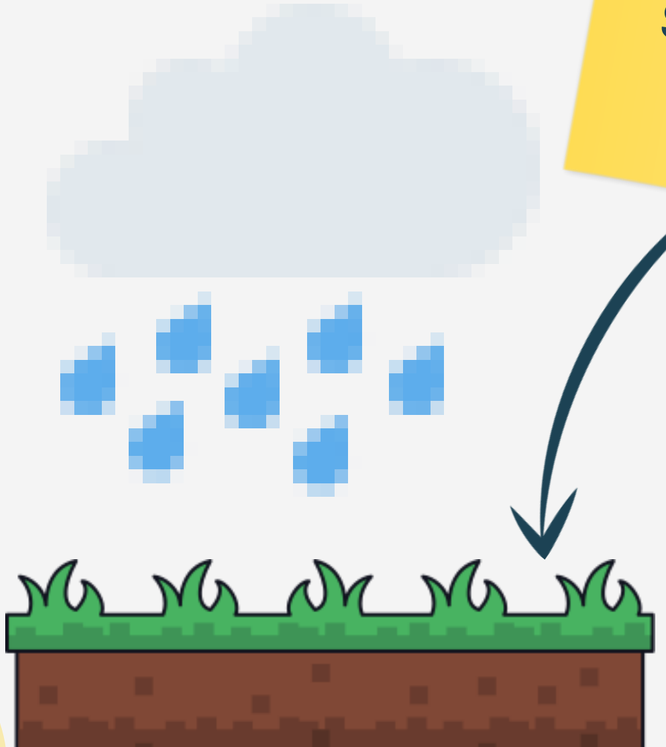
Tempo Amanhã

Desafios

continuando...

Variável disponível: Solo úmido

Tempo Hoje

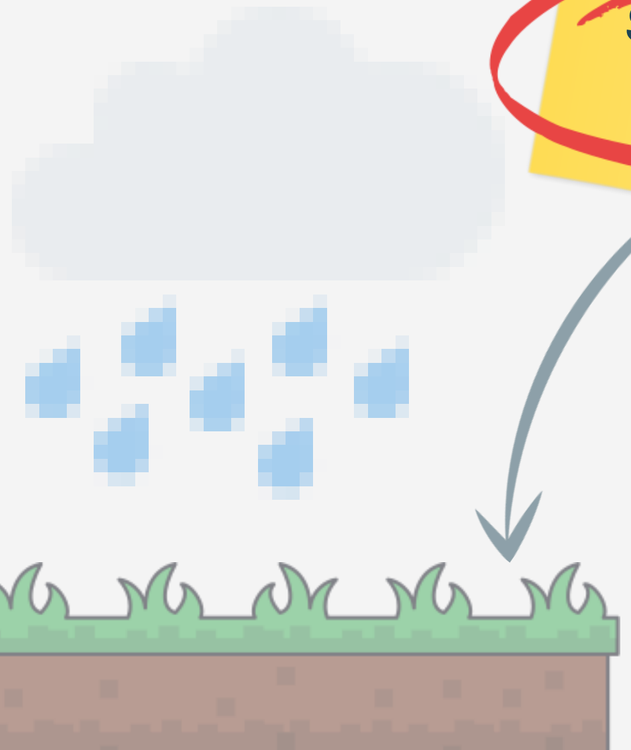


Tempo Amanhã

Desafios

Solo úmido
não causa um dia
seguinte chuvoso...

Tempo Hoje



Variável disponível:
Solo úmido

Tempo Amanhã

Desafios

Dependência de variáveis

- Traz problemas em relação a atribuição e extrapolação
- Atribuição de importância - difícil quando as variáveis são correlacionadas
- Muitos métodos baseados em sensibilidade permutam variáveis
- No caso de alta correlação com outras variáveis esse efeito se perde
  - Não existe efeito independente na prática
- Extrapolação pode gerar interpretações enganosas

Desafios

Definição de interpretabilidade

- Falta de definição é um problema
- Não existe uma forma direta de quantificar o quão interpretável é um modelo (ao contrário de performance preditiva) – não se sabe o “valor real”
- Duas principais formas de avaliar interpretabilidade:
  - Métricas matematicamente quantificáveis, e
  - Conversar com especialistas na área

Próximos desafios

- Necessidade de uma visão dinâmica da interpretação de ML
  - Da coleta dos dados até o seu uso final
- Como explicar predições para indivíduos com backgrounds diferentes
- Necessidade do campo da interpretabilidade se estender para outros domínios, como psicologia e sociologia

Referência:

Molnar, C., Casalicchio, G., & Bischl, B. Interpretable machine learning–a brief history, state-of-the-art and challenges. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 417-431). Springer, Cham. (2020).