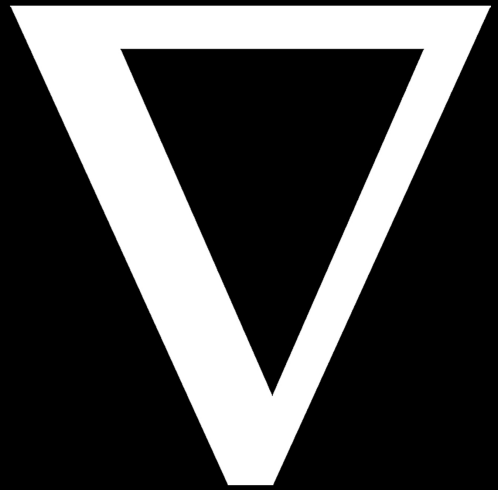
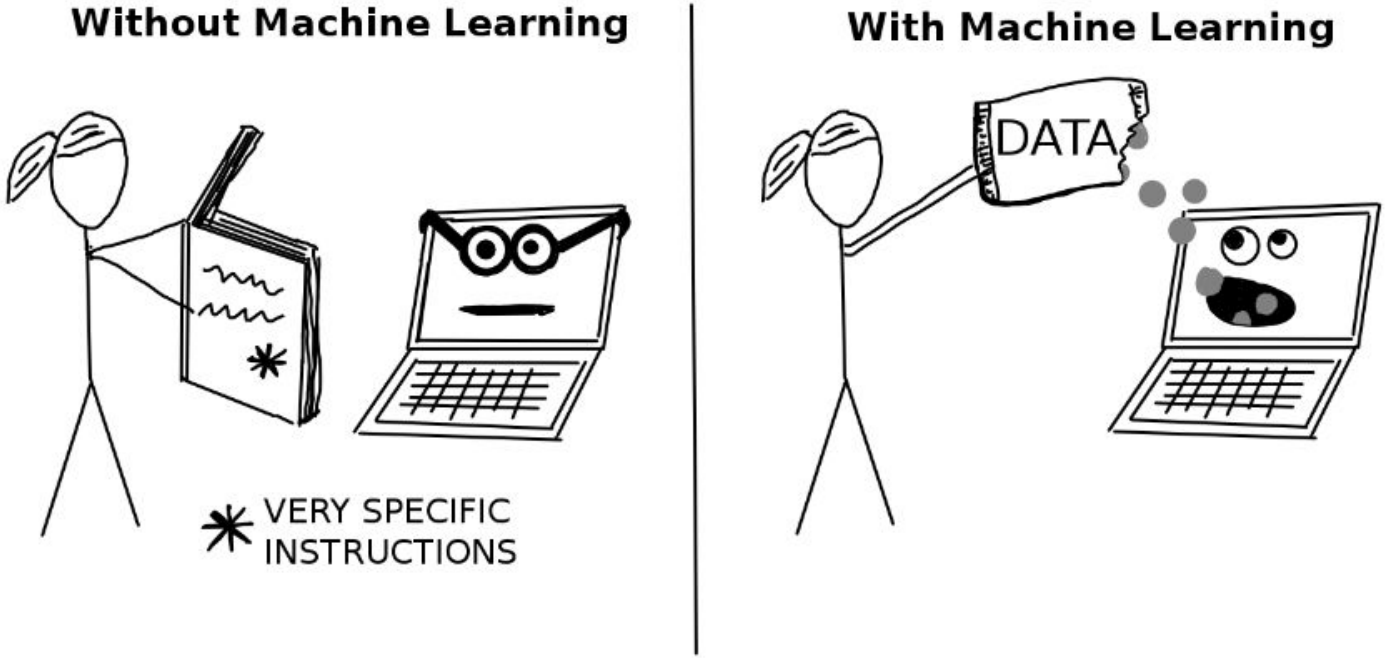


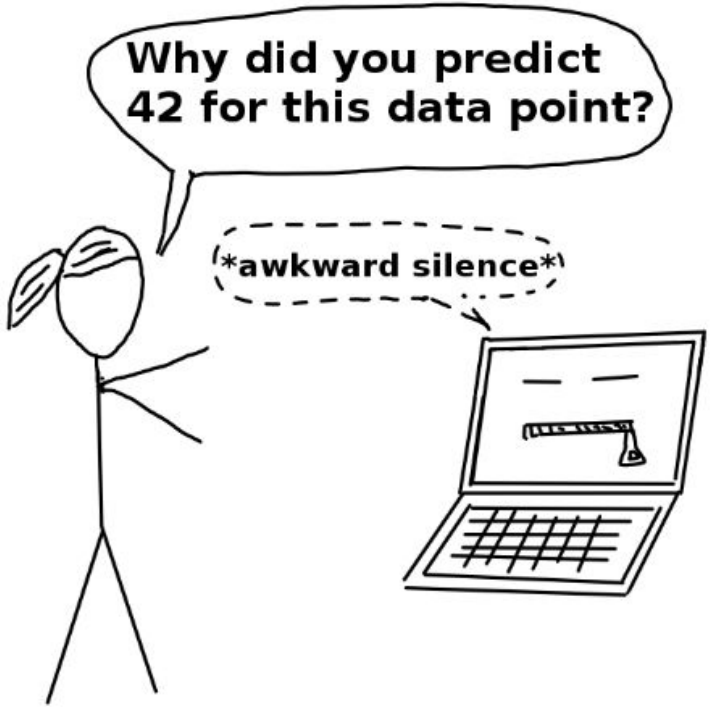
eXplainable Artificial Intelligence

CJ
@augustocj
06/03/2023

Machine Learning



Machine Learning



Por que interpretabilidade?

- Entendimento científico
- Segurança
- Ética
- Usuários afetados pelas decisões do modelo
- etc

Interpretabilidade - Definição

- Interpretabilidade é o grau com que um humano pode **entender** a causa de uma certa decisão. (Miller, 2017)
- Interpretabilidade é o grau com que um humano pode **consistentemente** predizer a resposta de um modelo. (Been et al, 2016)
- Machine Learning interpretável é um termo guarda-chuva para “extração de conhecimentos relevantes de um modelo preditivo relativo a correlações contidas nos dados ou aprendidas pelo modelo”.

Interpretabilidade - Escopo

- Transparência do algoritmo (Como o algoritmo cria o modelo? Quais relações são aprendidas?)
- Interpretabilidade global do modelo (Como o modelo treinado faz predições?)
- Interpretabilidade global do modelo em um nível modular (Como as partes do modelo afetam as predições?)

Interpretabilidade - Escopo

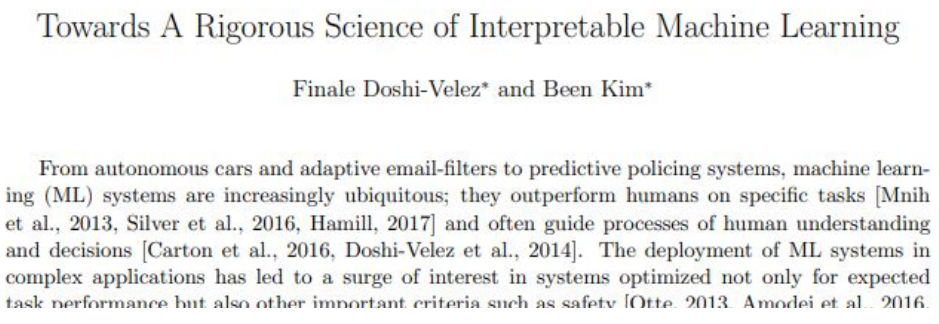
- Interpretabilidade local para uma única predição (Por que o modelo fez essa predição?)
- Interpretabilidade local para um grupo de predições (Por que o modelo fez essa predição para esse grupo de observações?)

Interpretabilidade - Taxonomia

- Intrinsic or post-hoc?
- Model-specific or model-agnostic?
- Local or global?

Interpretabilidade - Evaluation

- Doshi-Velez and Kim (2017) propuseram três níveis para a avaliação da interpretabilidade
  - Avaliação em nível de aplicação (tarefa real)
  - Avaliação em nível humano (tarefa simples)
  - Avaliação em nível de função (tarefa de intermediário ou proxy)



Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez* and Been Kim*

From autonomous cars and adaptive email-filters to predictive policing systems, machine learning (ML) systems are increasingly ubiquitous; they outperform humans on specific tasks [Mnih et al., 2013, Silver et al., 2016, Hamill, 2017] and often guide processes of human understanding and decisions [Carton et al., 2016, Doshi-Velez et al., 2014]. The deployment of ML systems in complex applications has led to a surge of interest in systems optimized not only for expected task performance but also other important criteria such as safety [Otte, 2013, Amodei et al., 2016

Modelos interpretáveis

- Existem modelos intrinsecamente interpretáveis
- Normalmente modelos mais estatísticos e/ou lineares

Regressão Linear

- Linearidade
- Homocedasticidade
- Independência
- Ausência de multicolinearidade

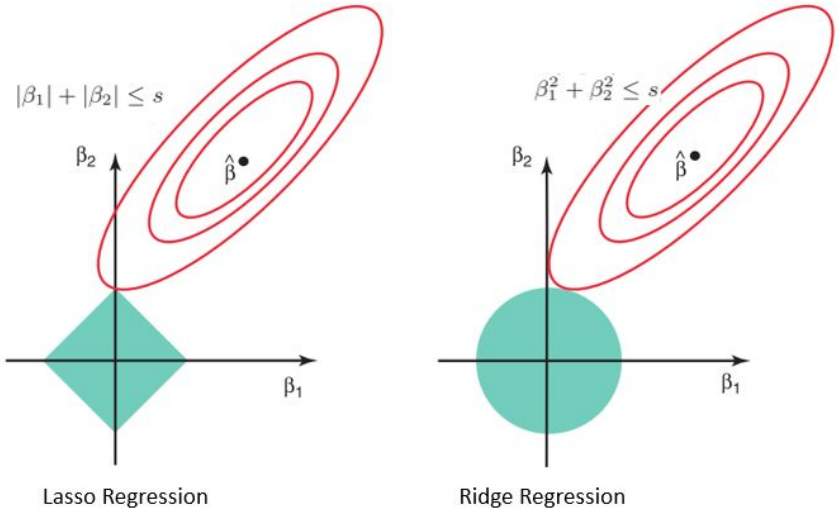
$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

Modelos Lineares Esparsos

- As penalidades podem levam a seleção de variáveis durante o treinamento
- Ridge
- Lasso

$$min_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-x_i^T\boldsymbol{\beta})^2+\lambda||\boldsymbol{\beta}||_1\right)$$

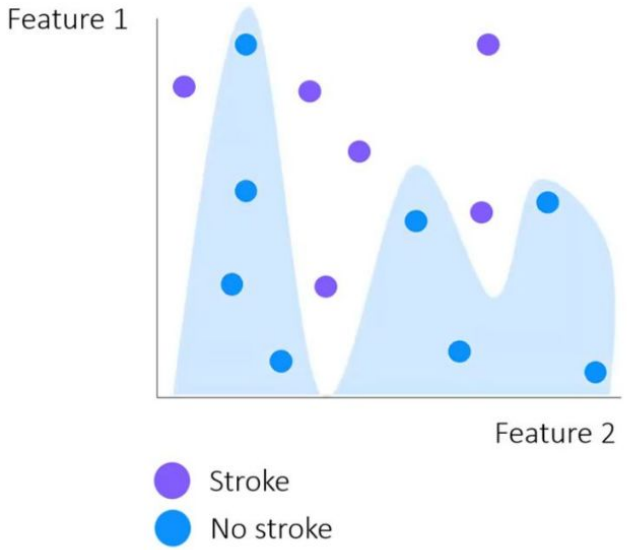


Outros

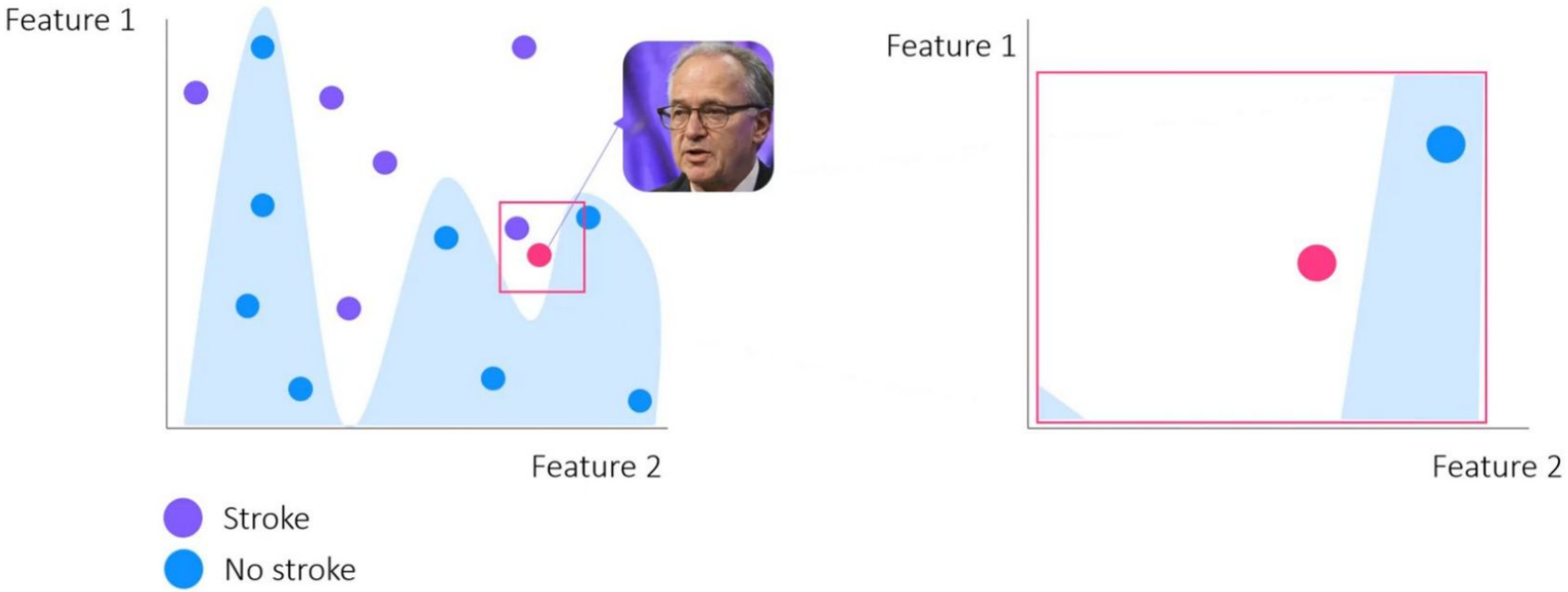
- Regressão Logística
- Modelo Lineares Generalizados, Modelos Aditivos Generalizados etc.
- Árvores de Decisão
- Regras de Decisão
- Naive Bayes
- KNN

dúvidas?

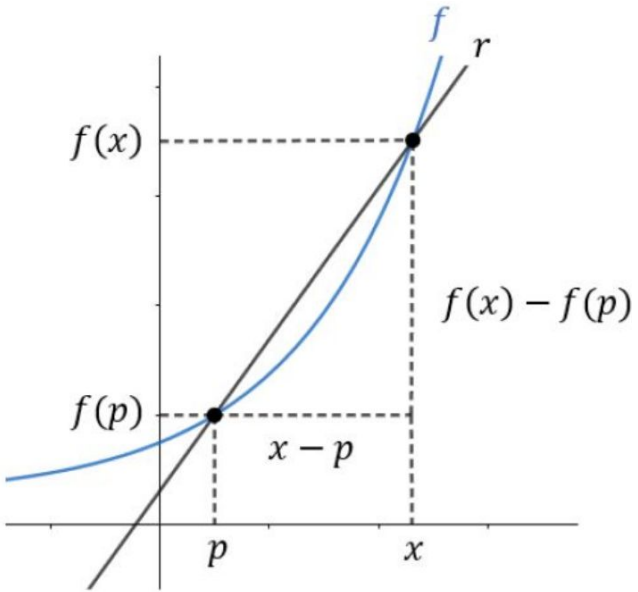
LIME - Local Interpretable Model-agnostic Explanations

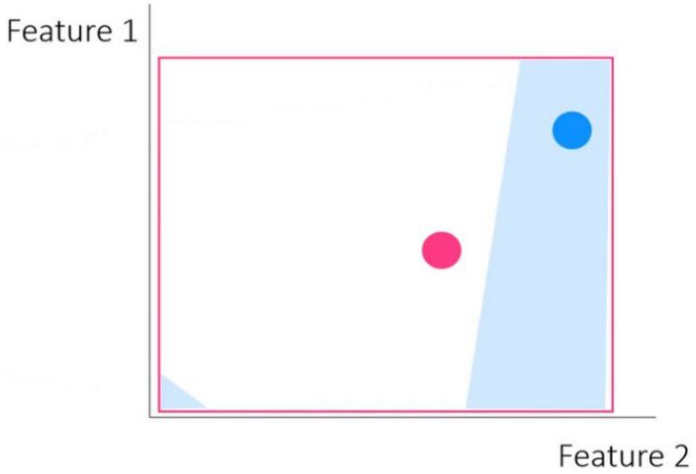


LIME

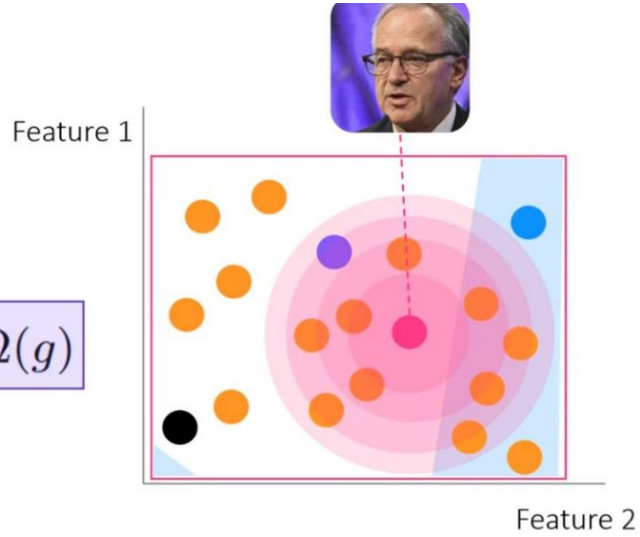


LIME





LIME



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

g = Sparse Linear Model

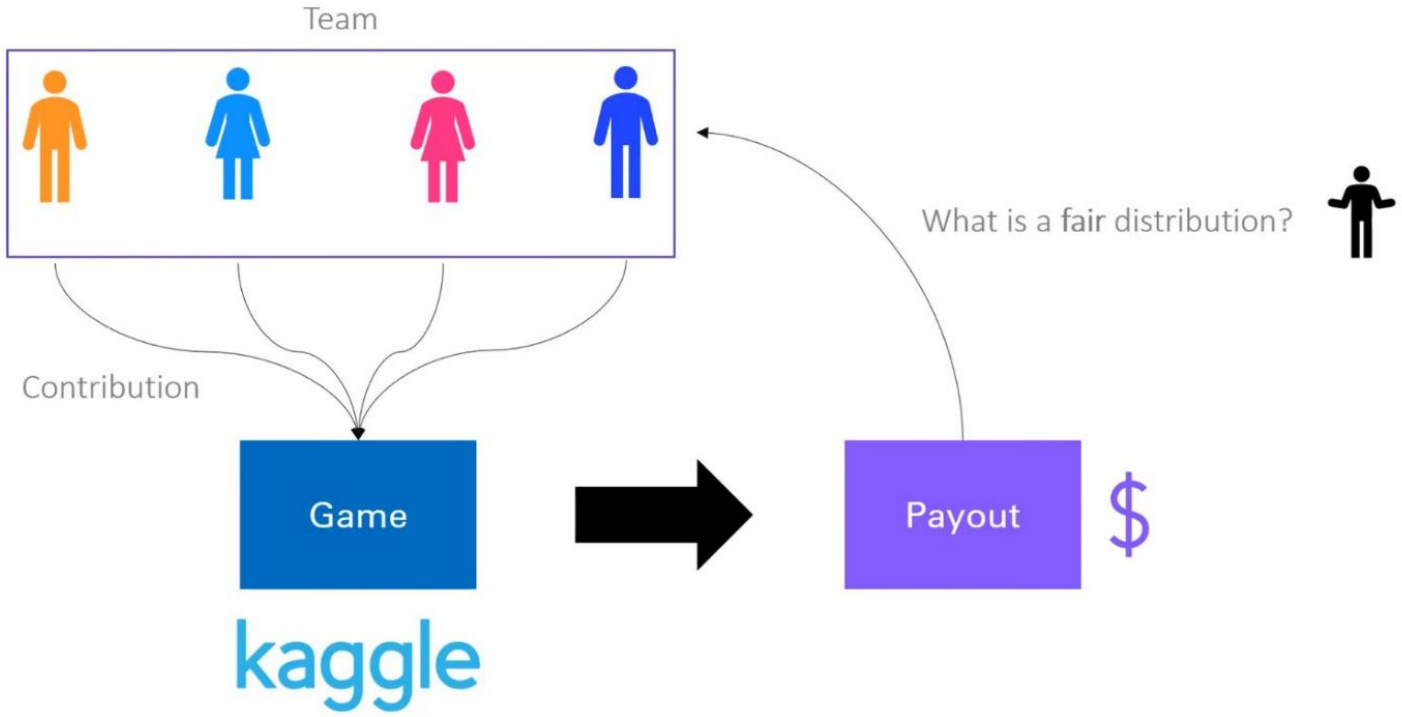
LIME

Exemplo

SHAP - SHapley Additive exPlanations

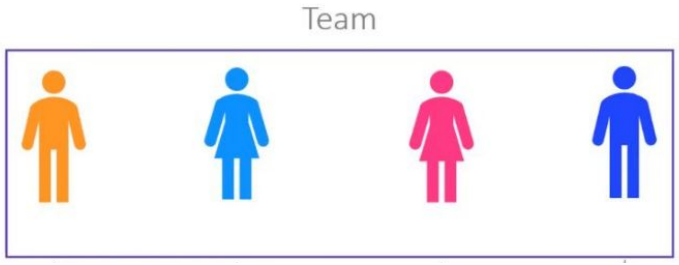
- Bastante utilizado
- Otimizado
- Dependendo do modelo, pode demorar um pouco

Teoria de Jogos Cooperativos

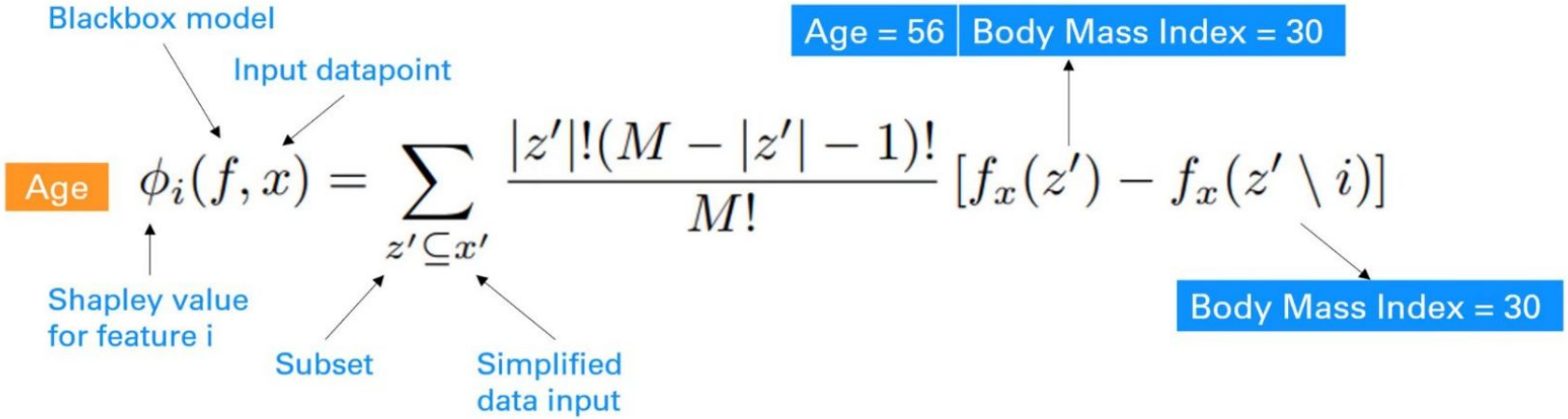


Shapley values

- Shapley values é uma medida que nos diz o quanto cada jogador contribuiu cooperativamente para a vitória
- Temos que considerar vários subsets

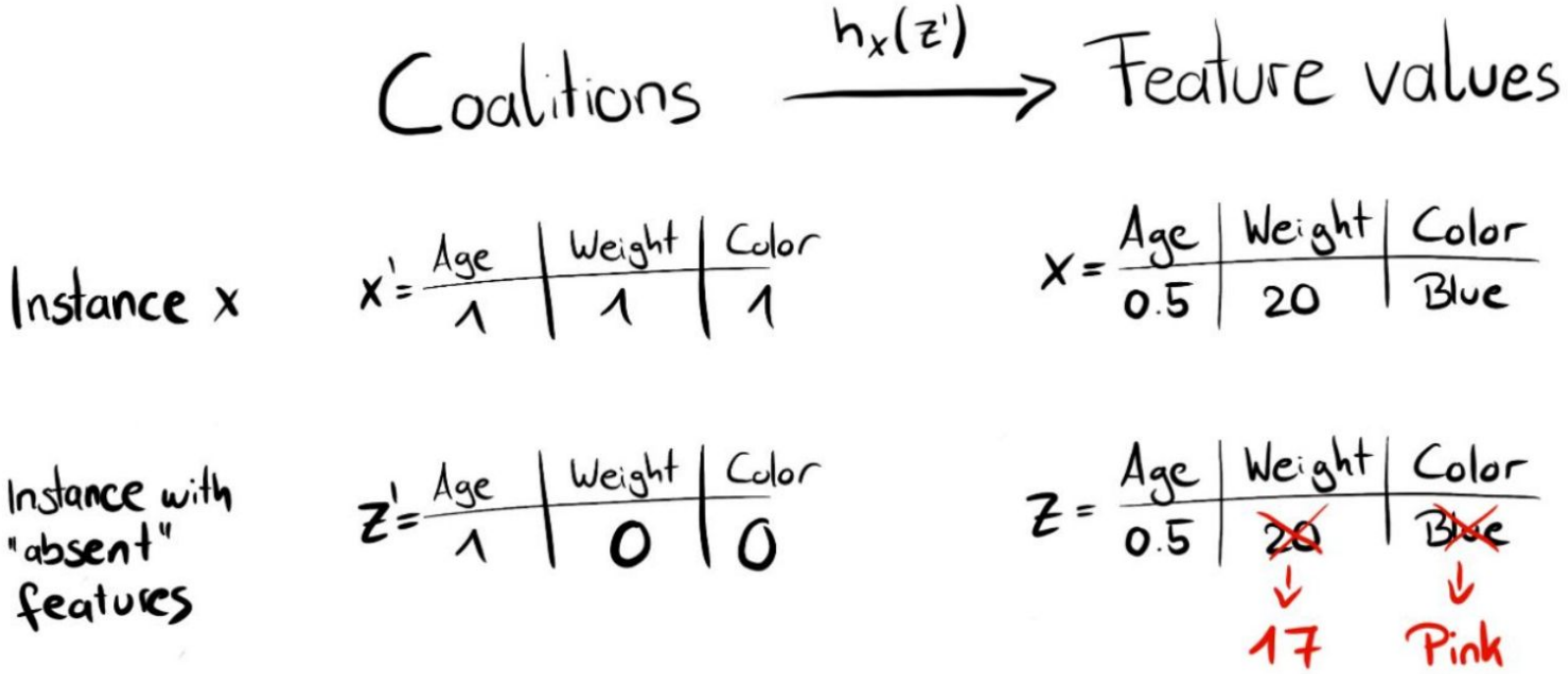


Calculando Shapley values



$x =$ | Age = 56 | Gender = F | Body Mass Index = 30 | Heart disease = yes | ... |

Aproximando os valores



KernelSHAP

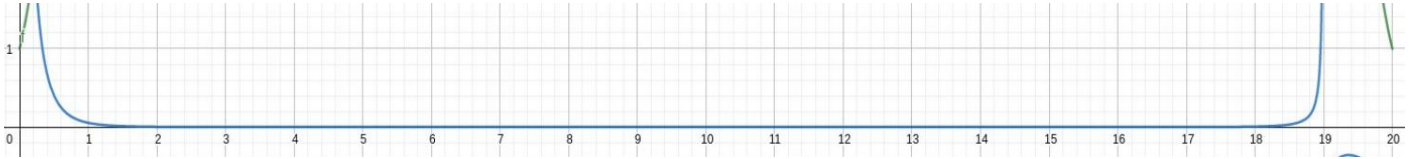
- Bem similar ao LIME, mas aqui não nos preocupamos em aproximar um modelo menos complexo, apenas ponderamos sobre a quantidade de informação que um subset tem.
- A intuição é que aprendemos mais sobre features individualmente se conseguimos medir o efeito delas em isolamento.

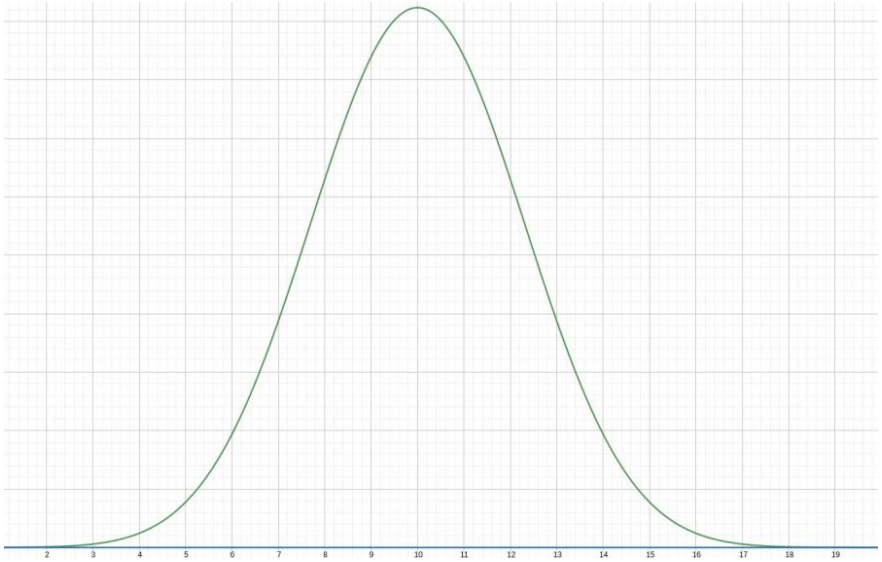
KernelSHAP

- A maior diferença do Lime está no peso atribuído nas instâncias da regressão linear!
  - LIME pesa as observações de acordo com a distância do “alvo”
  - SHAP pesa as observações de acordo com o peso de cooperação no Shapley values (queremos dar uma informação maior para grandes grupos e pequenos grupos.)

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

KernelSHAP





$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

$$f(x) = \frac{20!}{x!\,(20-x)!}$$

$a = f(1)$

$\rightarrow 20$

$d = f(10)$

$\rightarrow 184756$

$c = f(19)$

$\rightarrow 20$

Aproximando os valores

- Sample coalitions $z'_k \in \{0,1\}^M, \quad k \in \{1,\ldots,K\}$ (1 = feature present in coalition, 0 = feature absent).
- Get prediction for each z'_k by first converting z'_k to the original feature space and then applying model $\hat{f} : \hat{f}\left(h_x(z'_k)\right)$
- Compute the weight for each z'_k with the SHAP kernel.
- Fit weighted linear model.
- Return Shapley values ϕ_k, the coefficients from the linear model.

Aproximando os valores

- Os coeficientes dessa regressão linear são os shapley values

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z')$$

TreeSHAP

- Não é model-agnostic
- Usa esperança condicional em vez da esperança marginal
    - Isso pode dificultar a análise pois colunas que não influenciam na predição podem obter valores diferentes de 0
    - Correlação com outra coluna que tem impacto na predição
- Graças a propriedade aditiva de Shapley values, eles podem ser ponderados quando se tem um ensemble de árvores

$$E_{X_S|X_C}(\hat{f}(x)|x_S)$$

Vantagens

- SHAP tem uma fundamentação teórica sólida
- Temos explicações contrastivas que comprara a predição com a predição média
- SHAP conecta LIME com Shapley values
- SHAP tem uma implementação rápida para modelos baseados em árvores
- É legal para interpretações globais

Desvantagens

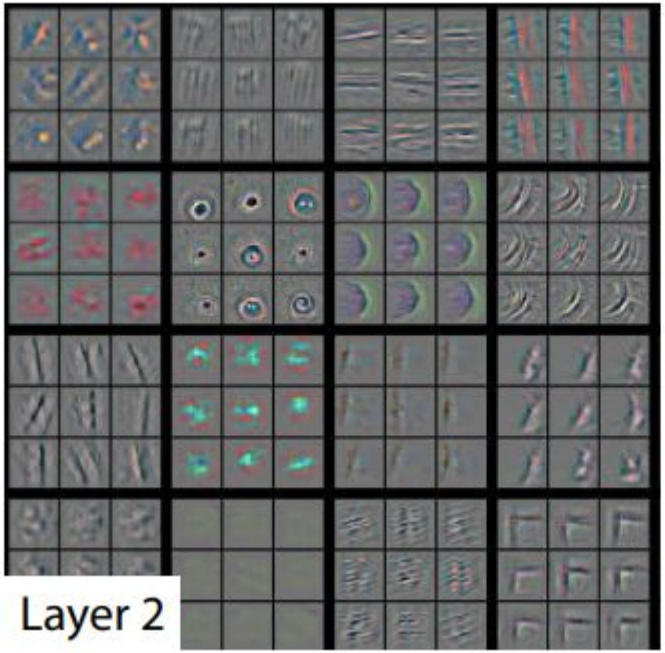
- KernelSHAP é lento
- KernelSHAP ignora feature dependence
- TreeSHAP pode produzir atribuições não intuitivas
- Shapley values podem ser mal interpretados
- Possível criar intencionalmente interpretações erradas com SHAP
- Não temos certeza absoluta da explicação

Visualizing and Understanding Convolutional Networks

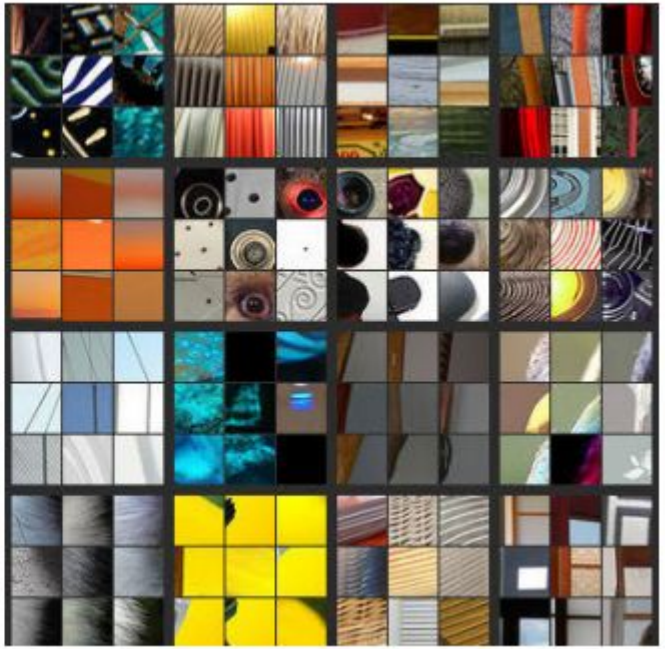
Matthew D. Zeiler and Rob Fergus

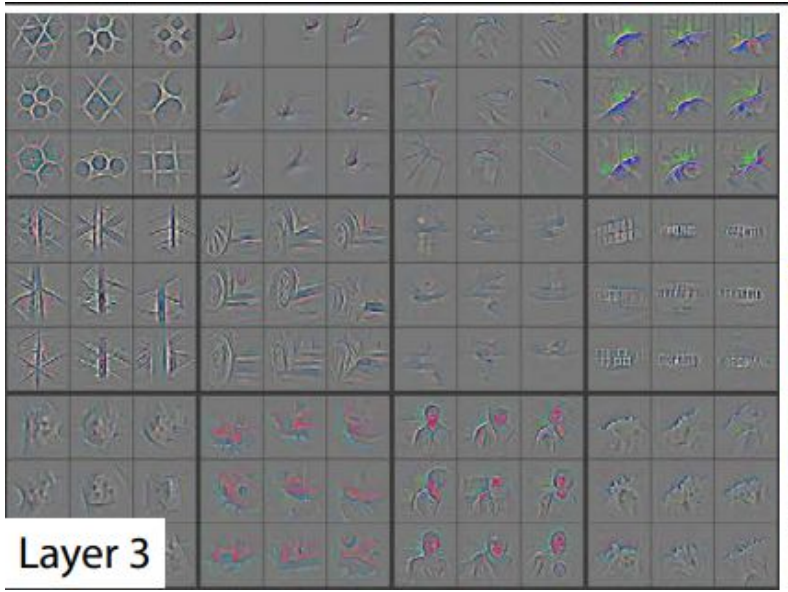
Dept. of Computer Science,
New York University, USA
{zeiler,fergus}@cs.nyu.edu

Layer 1

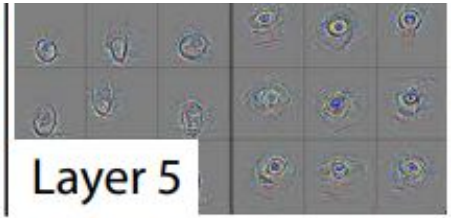


Layer 2




Layer 3


Layer 4


Layer 5

Grupo de XAI

- Leitura e discussão de papers na maioria dos encontros
- Votação de papers
- No final vamos fazer uma coisa, mas é surpresa

Como ler papers? Andrew Ng ensina

- Como são escritos? Exemplo
- Como devemos ler?
  - 1a vez: título, resumo, e figuras
  - 2a vez: introdução, conclusão, figuras, e dar uma olhada no resto (pular trabalhos relacionados)
  - 3a vez: ler tudo, mas pular a matemática
  - 4a vez: ler, mas pular as partes que não fazem sentido

Materiais complementares

- Série de vídeos do Explainable AI explained!
- Interpretable ml book
- “Why should I Trust You?” Explaining the predictions of Any Classifier (LIME paper)
- Explaining prediction models and individual predictions with feature contributions (KernelSHAP paper)
- A Unified Approach to Interpreting Model Predictions (SHAP paper)
- MeLIME: Meaningful Local Explanations for Machine Learning Models

Materiais complementares

- From local explanations to global understanding with explainable AI for trees (TreeSHAP)