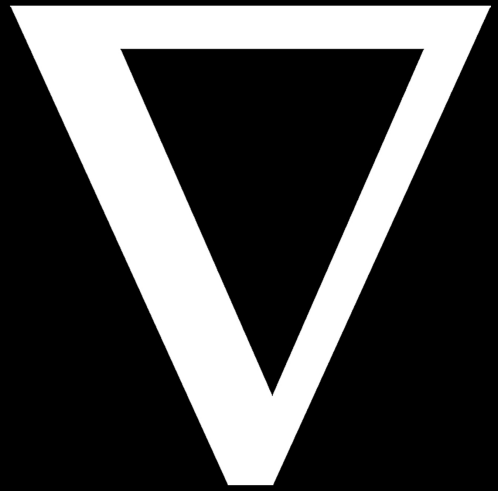


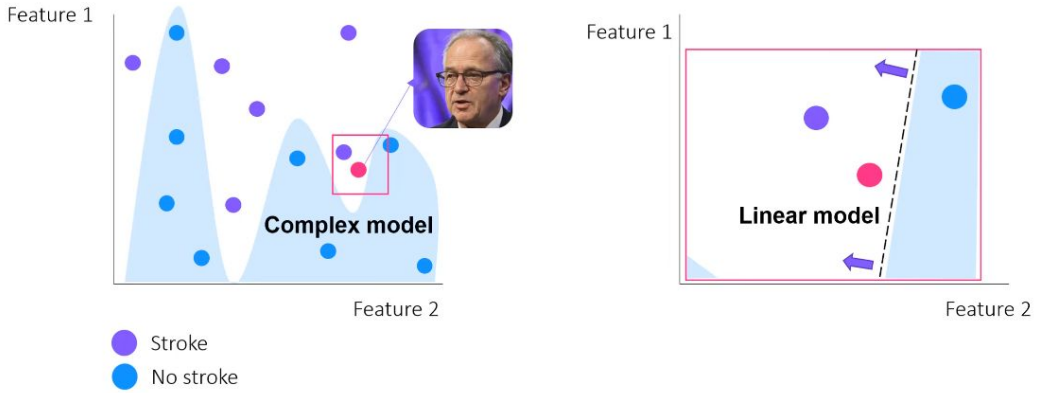
LIME

Local Interpretable Model-Agnostic Explanations

Lucas Greff Meneses
@greffao
10/04/2024

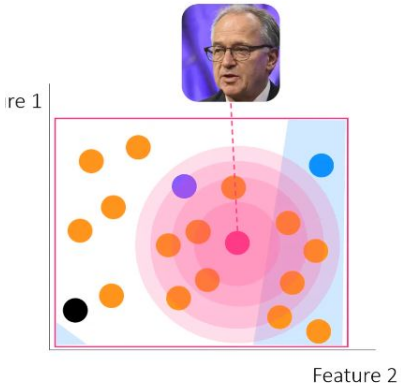
Local Surrogate Model

- LIME é basicamente uma implementação de um modelo substituto local. O objetivo é criar um modelo interpretável que explique uma determinada previsão do modelo não interpretável.



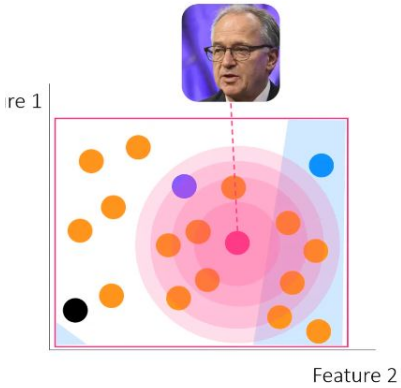
Local Surrogate Model

- LIME gera um novo dataset alterando as features da predição que se deseja entender e passando esses novos pontos pelo modelo não interpretável.



Local Surrogate Model

- Neste novo dataset, o método treina um novo modelo interpretável, de forma que os pontos mais próximos da instância que queremos entender tenham peso maior e usando as previsões do modelo não interpretável como label.



Matematicamente

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(x)$$

x - É a instância que queremos explicar

g - É um modelo interpretável dentro de um conjunto G de possíveis modelos interpretáveis (ex: regressão linear, árvore de decisão)

f - É o modelo não interpretável ($black\text{-}box$) (ex: rede neural, decision forest)

π_x - Função de proximidade que pondera a importância de cada novo ponto de acordo com a sua proximidade em relação à instância x

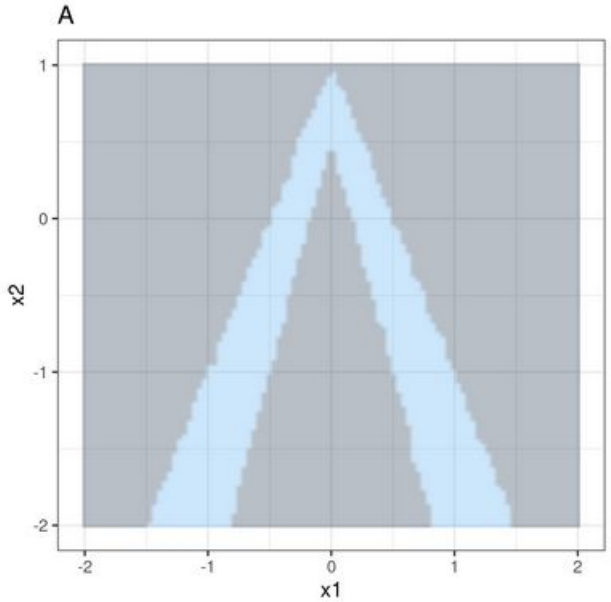
L - Função erro do modelo interpretável (ex: erro quadrático médio)

Ω - Função que mede a complexidade do modelo g (ex: quantidade de features numa regressão linear, profundidade da árvode de decisão)

Como gerar o novo dataset?

Como gerar o novo dataset para treinar o modelo g?

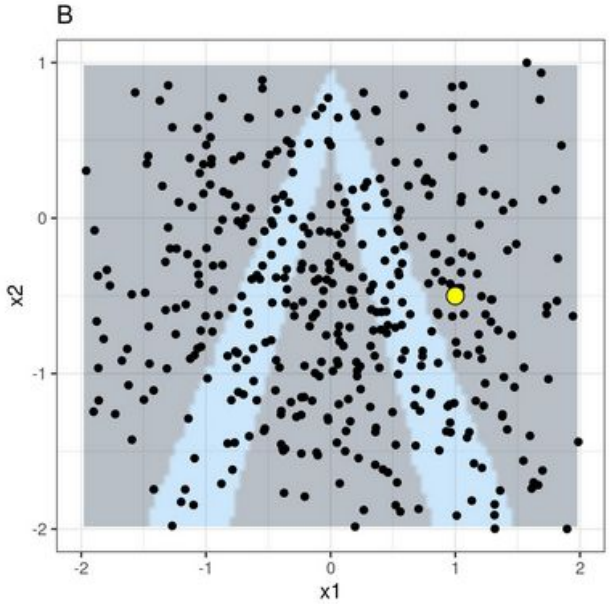
Dados estruturados



- Imagine que um algoritmo de classificação, com base nas features x1 e x2, previu as classes azul e preta de acordo com o gráfico

Como gerar o novo dataset para treinar o modelo g?

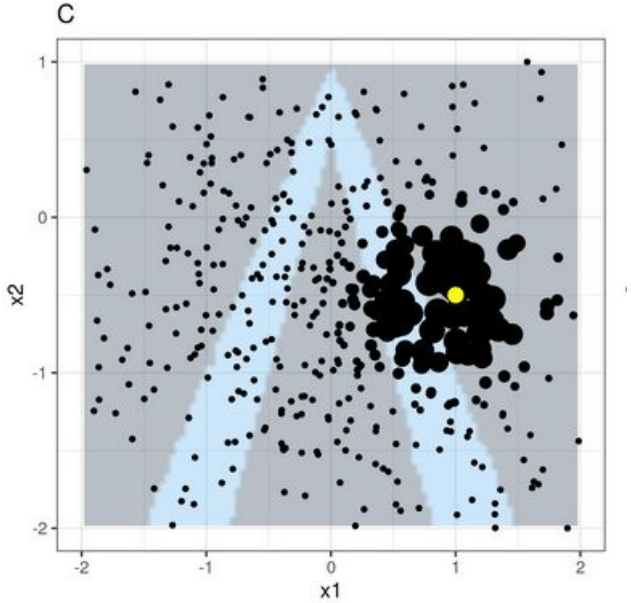
Dados estruturados



- Para entender porque o ponto amarelo foi classificado como preto, LIME gera novos pontos seguindo uma distribuição normal em relação ao centro de massa do dataset.

Como gerar o novo dataset para treinar o modelo g?

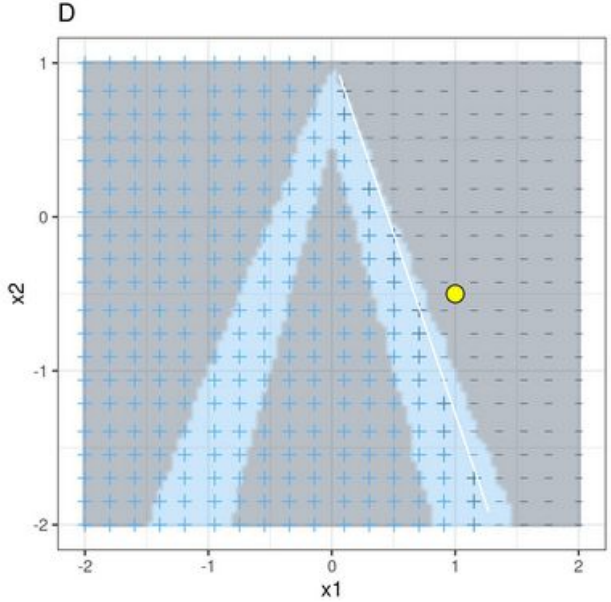
Dados estruturados



- Pontos perto da instância amarela recebem um peso maior, de acordo com a função pi.

Como gerar o novo dataset para treinar o modelo g?

Dados estruturados



- Um modelo interpretável é treinado a partir daqueles pontos.

Como gerar o novo dataset para treinar o modelo g?

Dados estruturados - Alguns problemas

- Os novos pontos são gerados a partir de uma distribuição normal, ignorando uma possível correlação entre as variáveis.
- Por exemplo, numa possível tarefa de classificação de casas, x1 pode ser o tamanho em m^2 e x2, o número de quartos. Ao criar novos pontos usando uma distribuição normal, uma casa com 20 m^2 e 7 quartos pode ser amostrada, o que é altamente improvável na realidade.

Como gerar o novo dataset para treinar o modelo g?

Dados estruturados - Alguns problemas

- Além disso, na implementação do LIME, a função pi é uma *exponential smoothing kernel* e o tamanho do *kernel* é 0.75 *(raiz da quantidade de colunas).

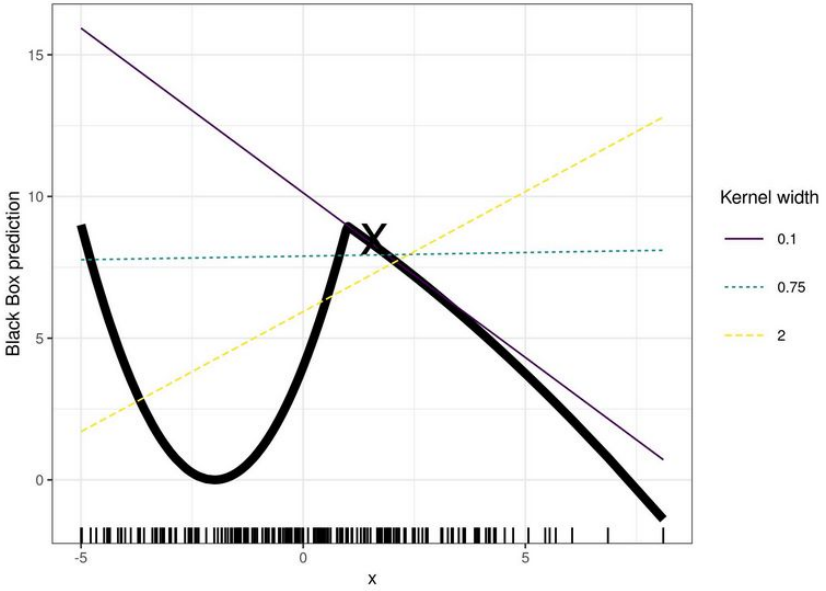
```
if kernel_width is None:
    kernel_width = np.sqrt(training_data.shape[1]) * .75
kernel_width = float(kernel_width)

if kernel is None:
    def kernel(d, kernel_width):
        return np.sqrt(np.exp(-(d ** 2) / kernel_width ** 2))
```

$$\pi_x(d, \text{kernel\_width}) = \sqrt{e^{-\frac{d^2}{\text{kernel\_width}^2}}}$$

Como gerar o novo dataset para treinar o modelo g?

Dados estruturados - Alguns problemas



- O problema é que não sabemos qual é a melhor função *kernel* e qual o melhor tamanho de kernel (kernel_width).

Como gerar o novo dataset para treinar o modelo g?

Dados não estruturados - texto

- Para criar variações dos dados textuais, o LIME gera novos textos removendo palavras do texto original.
- O dataset é feito de forma que cada palavra é uma feature e o seu valor binário indica a presença ou ausência da palavra no novo texto.

Como gerar o novo dataset para treinar o modelo g?

Dados não estruturados - texto

	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channel! ;)	1

For	Christmas	Song	visit	my	channel!	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

- prob é a probabilidade da nova sentença gerada ser spam.
- weight é a proximidade da nova sentença com a sentença original.

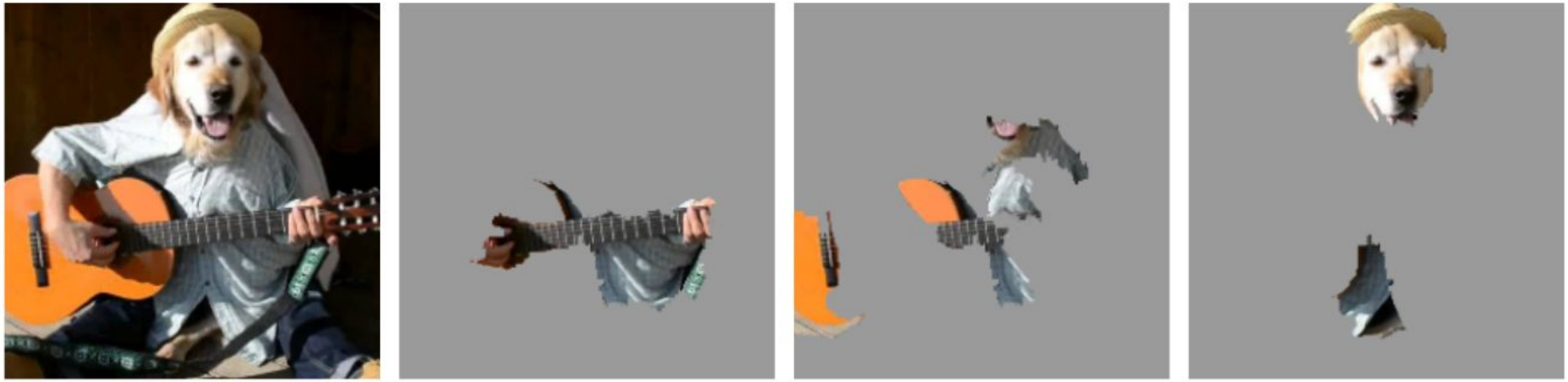
Como gerar o novo dataset para treinar o modelo g?

Dados não estruturados - imagens

- Para criar novas imagens, a imagem original é segmentada em superpixels que são ativados ou desativados a fim de criar o novo dataset.
- Superpixels são pixels interconectados com cores semelhantes.

Como gerar o novo dataset para treinar o modelo g?

Dados não estruturados - imagens



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Aplicações

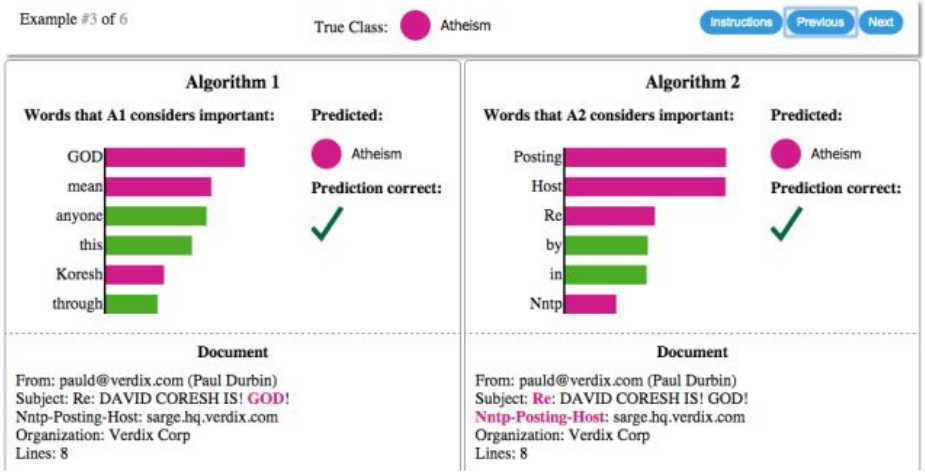
Aplicações

Identificação de data leakages

- Data leakages refere-se a uma situação em que informações externas não intencionais são incluídas no conjunto de treinamento do modelo.
- O artigo se refere à um caso em que o ID do paciente estava sendo usado pelo modelo e era uma variável considerada importante pelo modelo em certos casos.

Aplicações

Interpretação do modelo



- É possível usar o LIME para verificar se a análise do modelo faz sentido.
- Neste caso, o modelo deveria dizer se o texto não era sobre ateísmo” ou “cristianismo”. O algoritmo 2 obteve melhor desempenho.

Vantagens e Desvantagens

Vantagens

- É um método independente de modelo.
- Funciona para diversos tipos de dados.
- O modelo substituto pode usar features que não foram usadas pelo modelo original.

Desvantagens

- Não há uma definição exata do tamanho do kernel e nem da função que devemos usar.
- Usar uma distribuição normal para gerar novos dados ignora correlação entre features.
- Instabilidade nas explicações: mesmo em pontos muito próximos, dependendo do processo de geração do novo dataset, é possível gerar explicações totalmente diferentes, o que compromete a confiabilidade no método.
- As explicações podem ser manipuladas, a fim de esconder vieses.
- O usuário precisa escolher entre fidelidade e simplicidade do modelo interpretável.

Fim