

Introdução à Classificação com Regressão Logística

Felipe Maia Polo

Resumo:

1. O que é Aprendizado Supervisionado e Classificação?
2. Regressão Logística;
3. Regularização;
4. Classificação;
5. Métricas de avaliação;

Aprendizado Supervisionado

- O Aprendizado Supervisionado talvez seja a primeira coisa que pensamos quando tocamos no assunto Machine Learning;
- Temos dados a respeito de uma variável de interesse Y e um vetor de features $X=(X_1,...,X_d)$;
- Gostaríamos de criar um modelo estatístico para prever valores Y quando conhecemos apenas $X=(X_1,...,X_d)$ para dados nunca vistos;
- Dentro do Aprendizado Supervisionado ainda temos dois tipos de tarefa, que são Regressão e Classificação;

Exemplo de Classificação Binária

- Nesse caso, Y é uma variável categórica que diz um determinado tumor é benigno ou maligno. Ex: X=informações sobre tumor e Y=maligno, benigno.

id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion
1000025	5	1	1	1
1002945	5	4	4	5
1015425	3	1	1	1
1016277	6	8	8	1
1017023	4	1	1	3
1017122	8	10	10	8



Maligno ou Benigno?

Exemplo de Classificação Binária

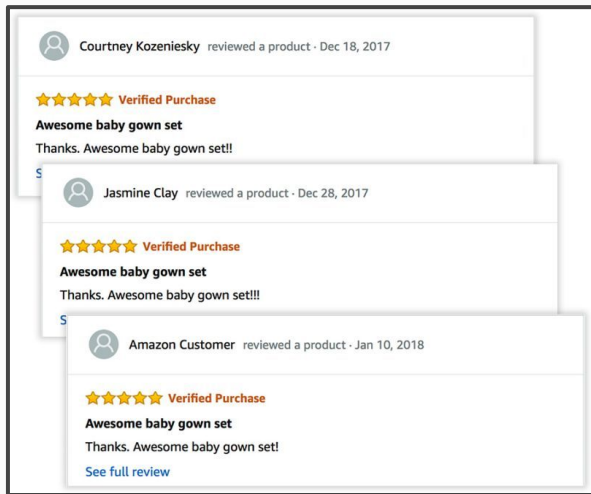
- Nesse caso, Y é uma variável categórica que diz se a imagem é de um doggo ou de um muffin. Ex: X =pixels e Y =chihuahua, muffin.



***Chihuahua
ou Muffin?***

Exemplo de Classificação Binária

- Nesse caso, Y é uma variável categórica que diz se um texto contém um sentimento positivo ou negativo. Ex: X =texto e Y =positivo, negativo.



*Sentimento
positivo ou
negativo?*

Procedimento para treinamentos de modelos

- Suponha que você tem uma base de dados

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$$

- Nós vamos dividir essa base de dados em uma parte para treino (~80%) e outra para teste (~20%):

$$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{train}}} \quad \mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=n_{\text{train}}+1}^n$$

- Toda a parte de estimação/treinamento de modelos de Machine Learning deve ser feita com os dados de treinamento. Usaremos a base de teste somente no fim.

Regressão Logística (Teoria)

- Y é uma variável aleatória que assume valores no conjunto $\{0,1\}$ ou $\{-,+\}$;
- $X=(X_1,\dots,X_d)$ é um vetor aleatório de features;

Nosso objetivo principal é estimar a função $f(x)=P(Y=1|X=x)$ que é a probabilidade condicional de que Y seja da classe 1 dado que observamos as características x .

Qual a diferença entre $P(Y=1)$ e $P(Y=1|X=x)$?

- $P(Y=1)$ é a fração da população de interesse em que verificamos o evento $\{Y=1\}$;
- $P(Y=1|X=x)$ é a fração da população que, dado que observamos $\{X=x\}$, podemos verificar $\{Y=1\}$.

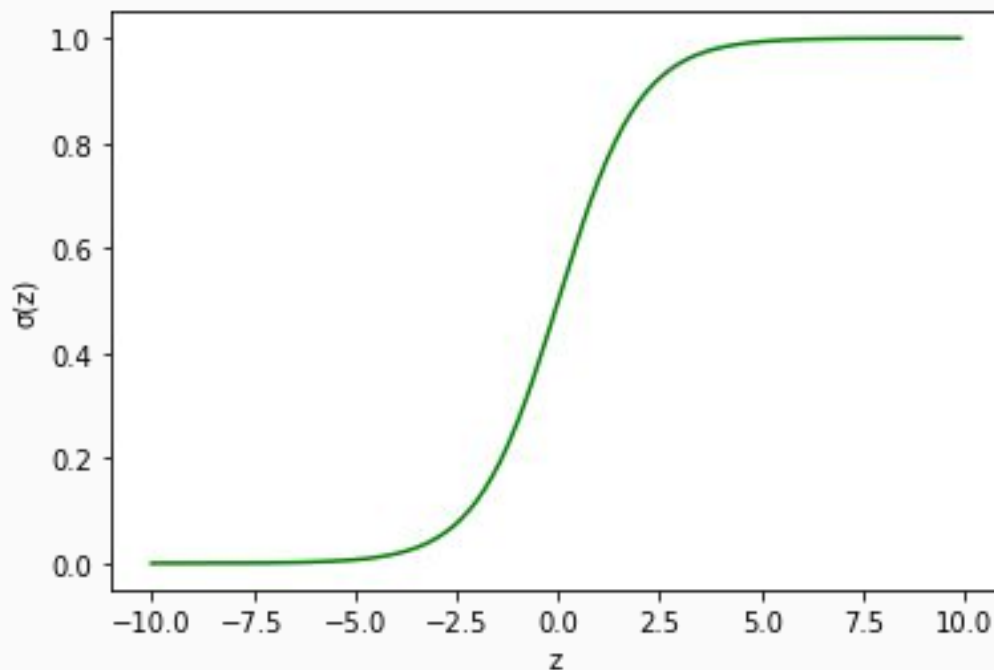
Regressão Logística (Teoria)

1. Dado que observamos essas medidas para o tumor, qual a probabilidade de ser benigno?
2. Dado que observamos esse conjunto de pixels, qual a probabilidade de ser chihuahua?
3. Dado que observamos esse conjunto ordenado de palavras, qual a probabilidade do comentário ser positivo?
4. Dado que o usuário tem esse padrão de comportamento dentro do meu site, qual a probabilidade de ele gastar mais de R\$1000?

Regressão Logística (Teoria)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Função Logística ou “Sigmoid”





Regressão Logística (Teoria)

Modelo:

$$\mathbb{P}(Y = 1|X = x) = \sigma(b + w_1x_1 + \dots + w_dx_d)$$

Intercepto ou “bias”



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$


Função Logística ou “Sigmoid”

Regressão Logística (Treinamento)

- O processo de estimação envolve a minimização do erro cometido pelo modelo na base de treinamento. Ou seja, os parâmetros estimados são aqueles valores que retornam o menor erro (“entropia cruzada” ou “log loss”) dentro dos nossos dados;
- Quanto menor o erro, melhor o ajuste do modelo aos dados;
- Depois de estimar os parâmetros:

$$\hat{\mathbb{P}}(Y = 1 | X = x) = \sigma(\hat{b} + \hat{w}_1 x_1 + \dots + \hat{w}_d x_d)$$

Regressão Logística (Treinamento - Exemplo)

Dados:

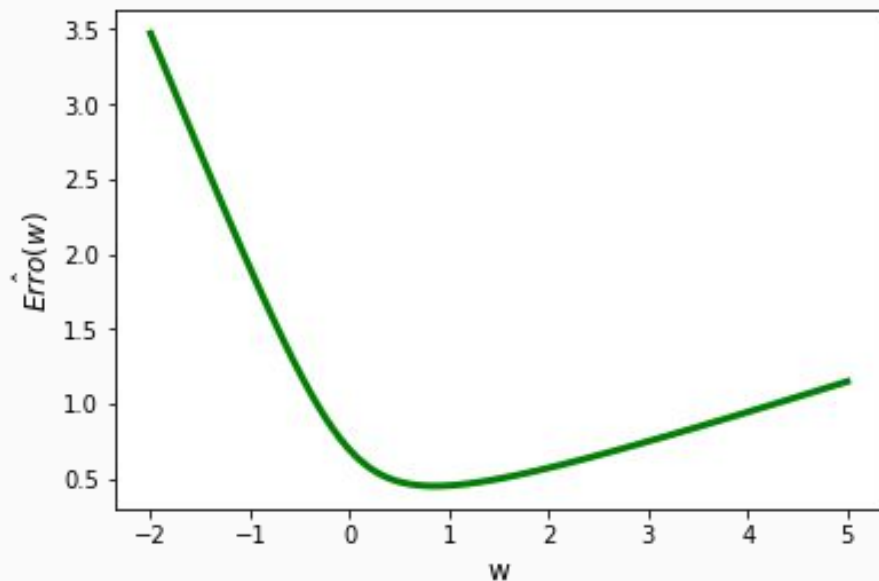
$$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{train}}}$$

Modelo:

$$\mathbb{P}(Y = 1 | X = x) = \sigma(w \cdot x)$$

Erro (empírico):

$$\hat{\text{ERRO}}(w)$$



Regressão Logística (Treinamento - Exemplo)

Dados:

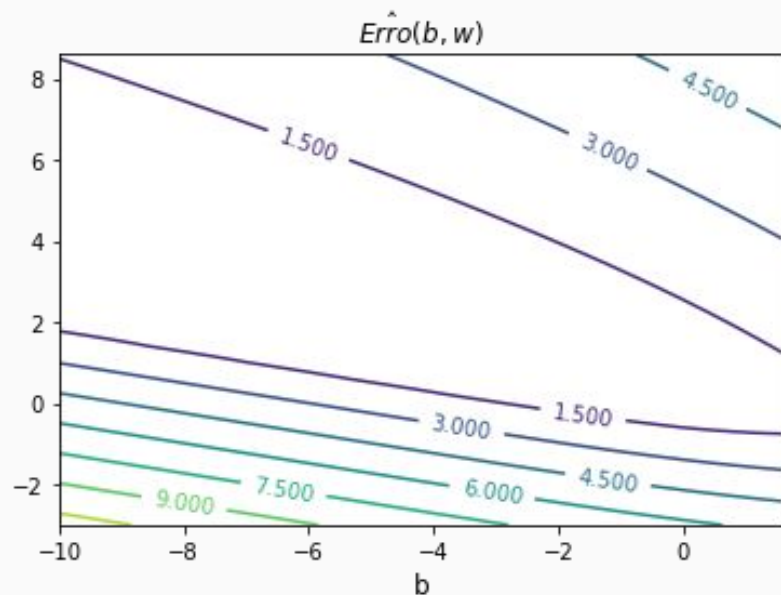
$$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{train}}}$$

Modelo:

$$\mathbb{P}(Y = 1 | X = x) = \sigma(b + w \cdot x)$$

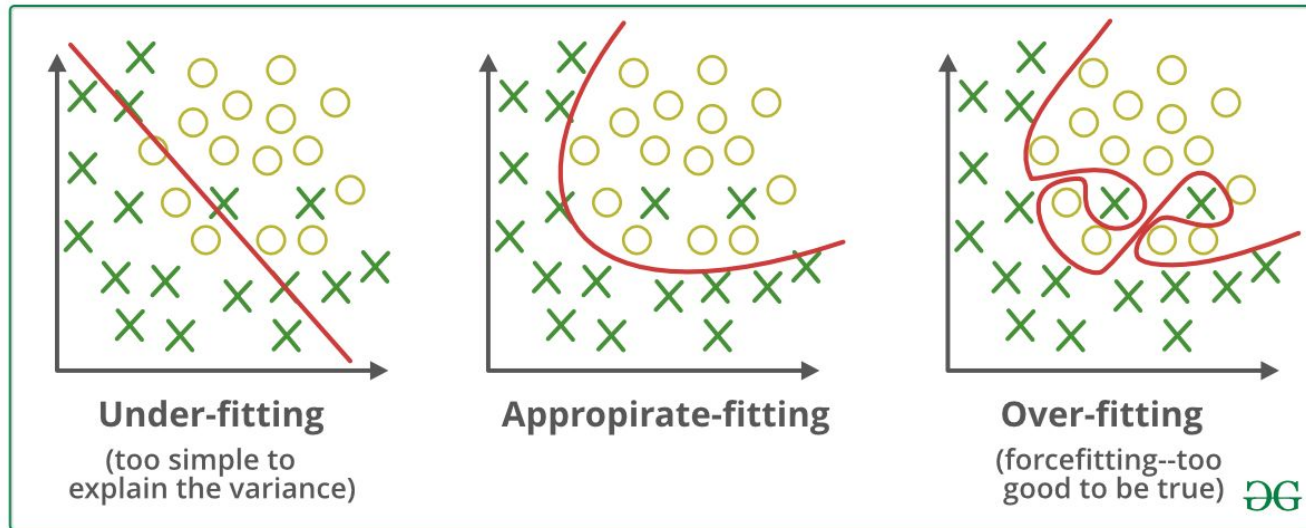
Erro (empírico):

$$\hat{\text{ERRO}}(b, w)$$



Notebook

Underfitting/Overfitting



Underfitting/Overfitting

Como me livrar do Underfitting?

- Aumentar a complexidade do modelo;

Como me livrar do Overfitting?

- Aumentar o número de amostras;
- Diminuir complexidade do modelo → **Regularização!**

Regularização (ou penalização)

Regularização pode ser dada pela penalização da complexidade do modelo;

- Aqui veremos a regularização do tipo “L1”, em que penalizamos os pesos separadamente:

$$ERRO'(b, w_1, w_2) = ERRO(b, w_1, w_2) + \frac{|b| + |w_1| + |w_2|}{C}, C > 0$$

- C é um hiperparâmetro e deve ser escolhido por validação;
- Esse tipo de penalização zera os pesos de variáveis irrelevantes;
- Funciona muito bem quando temos muitas variáveis e poucas amostras;

Validação Cruzada para a escolha de C

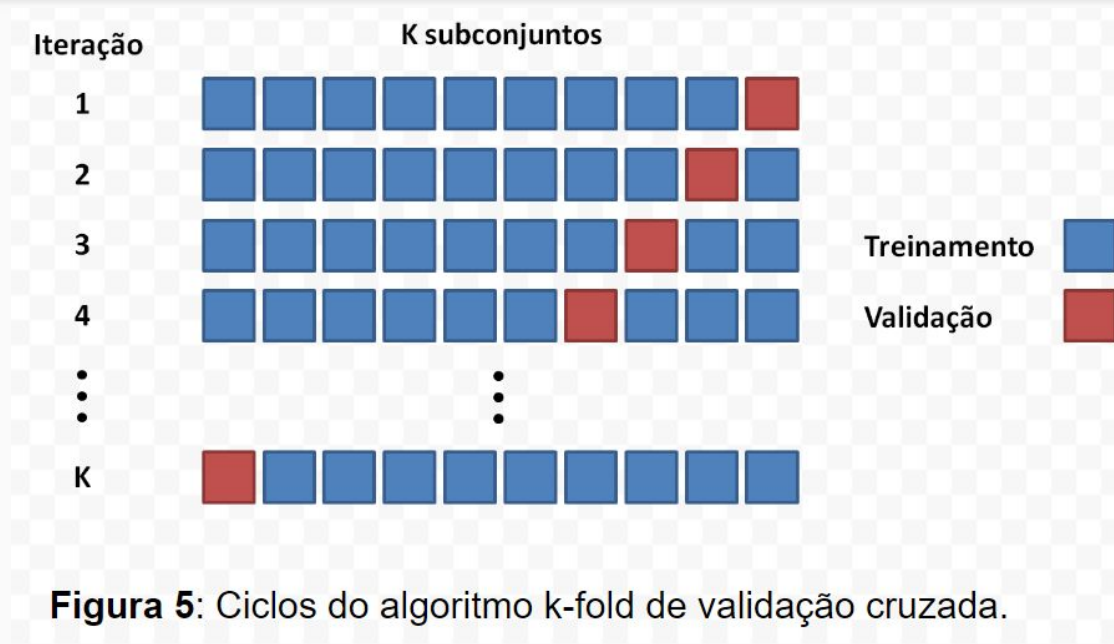


Figura 5: Ciclos do algoritmo k-fold de validação cruzada.

Notebook

Tarefa :)

- Comparar erro de treinamento e de teste para checar overfitting
- Brincar com pontos de corte na classificação: mudar os pontos de corte e checar como isso faz a matriz de confusão mudar;
- Estudar One-Hot Encoding:
 - <https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/>
- Interpretação dos coeficientes:
 - <https://christophm.github.io/interpretable-ml-book/logistic.html>
 - <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>