Naive Bayes

Prof. André Gustavo Hochuli

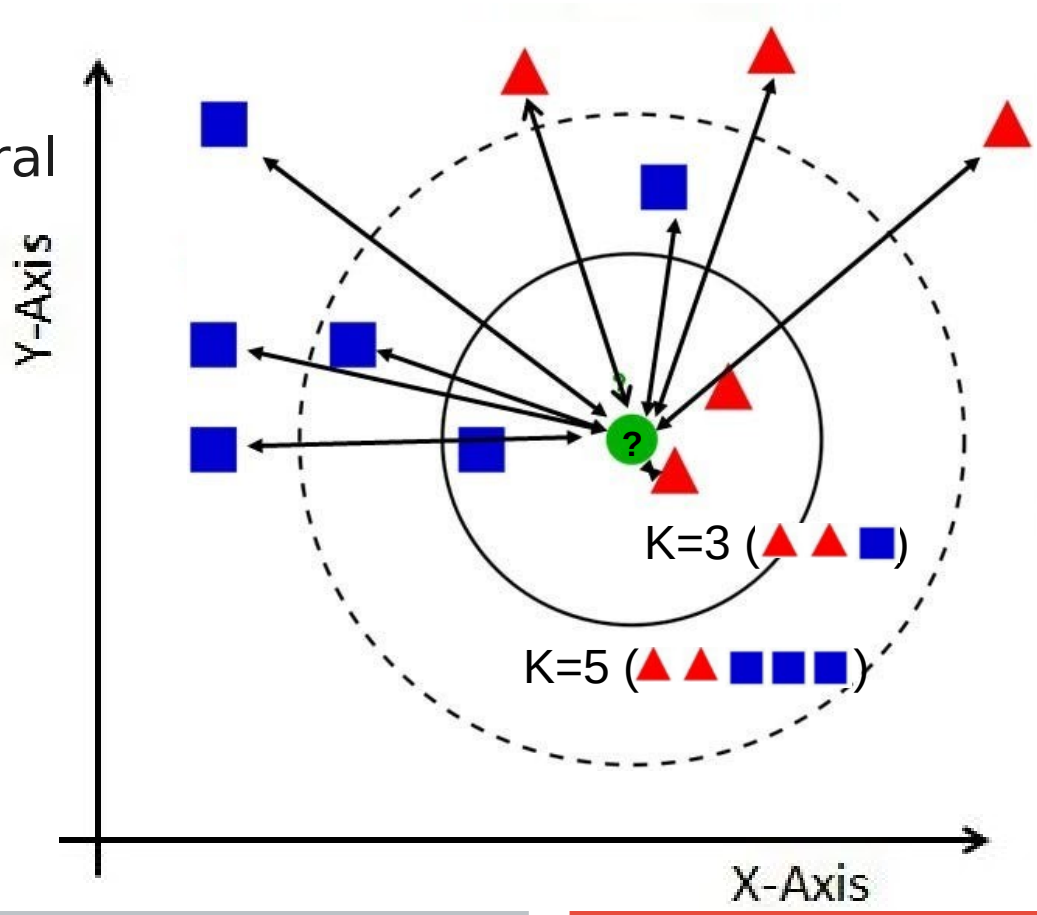
gustavo.hochuli@pucpr.br
aghochuli@ppgia.pucpr.br
github.com/andrehochuli/teaching

Plano de Aula

- Discussões Iniciais
- Aprendizado por Instâncias
- Algoritmo Naive Bayes
- Métricas de Avaliação
- Exercícios

Discussões Iniciais

- KNN
  - Simples
  - Desempenho vs Espaço Amostral



Teorema de Bayes

- Um piloto tem 50% de chances de vencer se chover, e 25% caso não ocorra chuva. Sabe-se que a probabilidade de chuva na corrida é de 30%.

- Dado que o piloto venceu, qual a probabilidade de ter chovido?

- Define-se que:

  - P(C) = 0.3
  - P(NC) = 1 – P(C) = 0.7

  - P(V) = ??
  - P(NV) = ??

Teorema de Bayes (Probabilidades)

- **Probabilidades Condicionais**
  - **P(A|B)= Probabilidade de acontecer A, dado que ocorreu B.**

  - Infere-se do texto que:
    - Vitória se ocorreu chuva P(V|C) = 50%
    - Vitória se não ocorreu chuva P(V|NC) = 0,25%

  - Se o piloto venceu, qual a probabilidade de ter chovido? Então:

    - P(C|V) = ???

Teorema de Bayes (Probabilidades)

- **Probabilidades Condicionais**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B) \cdot P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B|A) \cdot P(A)$$

$$\Downarrow$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Teorema de Bayes (Probabilidades)

- **Probabilidades Condicionais**

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

- **Teorema de Bayes**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Teorema de Bayes (Probabilidades)

- **Aplicando ao caso do piloto**
  - P(C) = 30%
  - P(NC) = 70%
  - P(V|C) = 50%
  - P(V|NC) = 25%

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

  - P(C|V) = ???

$$P(C|V) = \frac{P(V|C) \cdot P(C)}{P(V)} \rightarrow P(C|V) = \frac{0,5 \times 0,3}{P(V)}$$

Teorema de Bayes (Probabilidades)

- **Obtém-se P(V) pelo teorema da probabilidade total**

$$P(A) = \sum_{j=1}^{m} P(A \mid B_j) P(B_j)$$

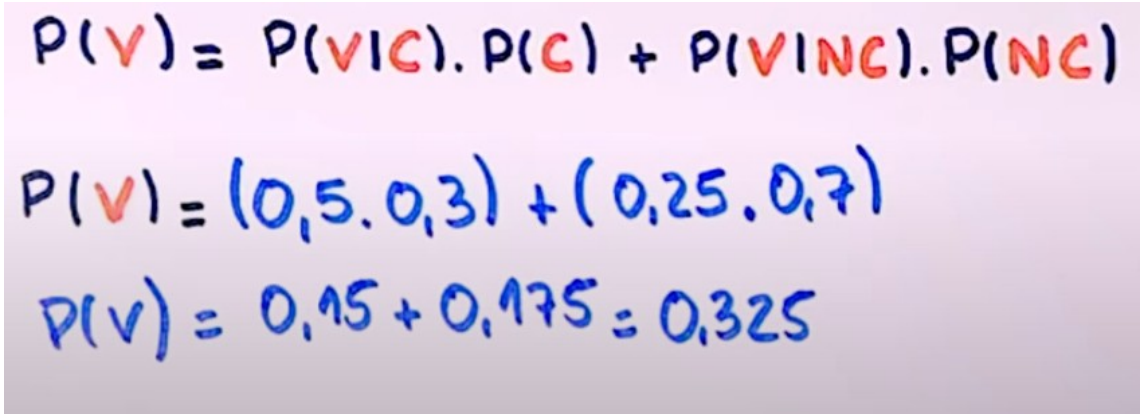
- **Vencer com ou sem chuva**

P(C) = 30%

P(NC) = 70%

P(V|C) = 50%

P(V|NC) = 25%



Teorema de Bayes (Probabilidades)

- Então, P(V) = 0.325, logo P(C|V):

$$P(C|V) = \frac{0,5 \times 0,3}{P(V)}$$

$$P(C|V) = \frac{0,5 \times 0,3}{0,325} = \frac{0,15}{0,325} = 0,46$$

Classificador Naive Bayes

- Naive (“Ingênuo”) : Variáveis/Características Independentes
- Estende o Teorema de Bayes para Múltiplas Variáveis

$$P(Y|X_1, X_2, X_3, \dots, X_n) = \frac{P(X_1|Y)P(X_2|Y)P(X_3|Y) \dots P(X_n|Y)P(Y)}{P(X_1)P(X_2)P(X_3) \dots P(X_n)}$$

$$P(Y|X_1, X_2, X_3, \dots, X_n) = \frac{P(X_1|Y)P(X_2|Y)P(X_3|Y) \dots P(X_n|Y)P(Y)}{\cancel{P(X_1)P(X_2)P(X_3) \dots P(X_n)}}$$

$$P(Y|X_1, X_2, X_3, \dots, X_n) = P(X_1|Y)P(X_2|Y)P(X_3|Y) \dots P(X_n|Y)P(Y)$$

Classificador Naive Bayes

- O dataset abaixo descreve um potencial comprador (computadores)
- Assume-se que as características são independentes
  - I.E: Renda alta não implica em crédito excelente

Income	Student	Credit Rating	Buys computer
high	no	fair	YES
medium	no	fair	NO
medium	no	excellent	NO
high	yes	fair	YES
high	no	fair	NO
medium	yes	fair	YES
low	yes	excellent	YES
low	yes	fair	YES
high	no	excellent	NO
medium	no	fair	YES
low	yes	fair	NO
low	yes	fair	NO
low	yes	fair	NO
high	yes	fair	YES
high	no	excellent	YES
high	no	excellent	YES
high	no	excellent	YES
medium	no	excellent	YES

Classificador Naive Bayes

- Sendo assim:
  - X_1: Income
  - X_2: Student
  - X_3: Credit
- Classe
  - Y: Buys Computer
- Naive Bayes para 3 features:

$$P(Y|X_1, X_2, X_3) = P(X_1|Y)P(X_2|Y)P(X_3|Y)P(Y)$$

Income	Student	Credit Rating	Buys computer
high	no	fair	YES
medium	no	fair	NO
medium	no	excellent	NO
high	yes	fair	YES
high	no	fair	NO
medium	yes	fair	YES
low	yes	excellent	YES
low	yes	fair	YES
high	no	excellent	NO
medium	no	fair	YES
low	yes	fair	NO
low	yes	fair	NO
low	yes	fair	NO
high	yes	fair	YES
high	no	excellent	YES
high	no	excellent	YES
high	no	excellent	YES
medium	no	excellent	YES

Classificador Naive Bayes

- Probabilidades *a priori* (Treino)

Student			P(X2)	
	YES	NO	P(YES)	P(NO)
yes	5	3	5/11	3/7
no	6	4	6/11	4/7
Total	11	7	100%	100%

Credit Rating			P(X3)	
	YES	NO	P(YES)	P(NO)
fair	6	5	6/11	5/7
excellent	5	2	5/11	2/7
Total	11	7	100%	100%

Income			P(X1)	
	YES	NO	P(YES)	P(NO)
high	6	2	6/11	2/7
medium	3	2	3/11	2/7
low	2	3	2/11	3/7
Total	11	7	100%	100%

Buys Computer		P(Y)
	Count	P(Y)
YES	11	P(YES) = 11/18
NO	7	P(NO) = 7/18
Total	18	100%

Classificador Naive Bayes

- Teste

Income	Student	Credit Rating
low	yes	excellent

Income P(X1)

	YES	NO	P(YES)	P(NO)
high	6	2	6/11	2/7
medium	3	2	3/11	2/7
low	2	3	2/11	3/7
Total	11	7	100%	100%

Student P(X2)

	YES	NO	P(YES)	P(NO)
yes	5	3	5/11	3/7
no	6	4	6/11	4/7
Total	11	7	100%	100%

Credit Rating P(X3)

	YES	NO	P(YES)	P(NO)
fair	6	5	6/11	5/7
excellent	5	2	5/11	2/7
Total	11	7	100%	100%

Buy Computer P(Y)

	Count	P(Y)
YES	11	P(YES) = 11/18
NO	7	P(NO) = 7/18
Total	18	100%

$$P(YES_{test}) = P(Income = low|YES) * P(Student = yes|YES) * P(Credit\ Rating = excellent|YES) * P(YES_{train}) = \frac{2}{11} * \frac{5}{11} * \frac{5}{11} * \frac{11}{18} = 0{,}023$$

$$P(NO_{test}) = P(Income = low|NO) * P(Student = yes|NO) * P(Credit\ Rating = excellent|NO) * P(NO_{train}) = \frac{3}{7} * \frac{3}{7} * \frac{2}{7} * \frac{7}{18} = 0{,}0204$$

Classificador Naive Bayes

- Input (Test)

Income	Student	Credit Rating
low	yes	excellent

$P(YES_{test}) = 0{,}023$

$P(NO_{test}) = 0{,}0204$

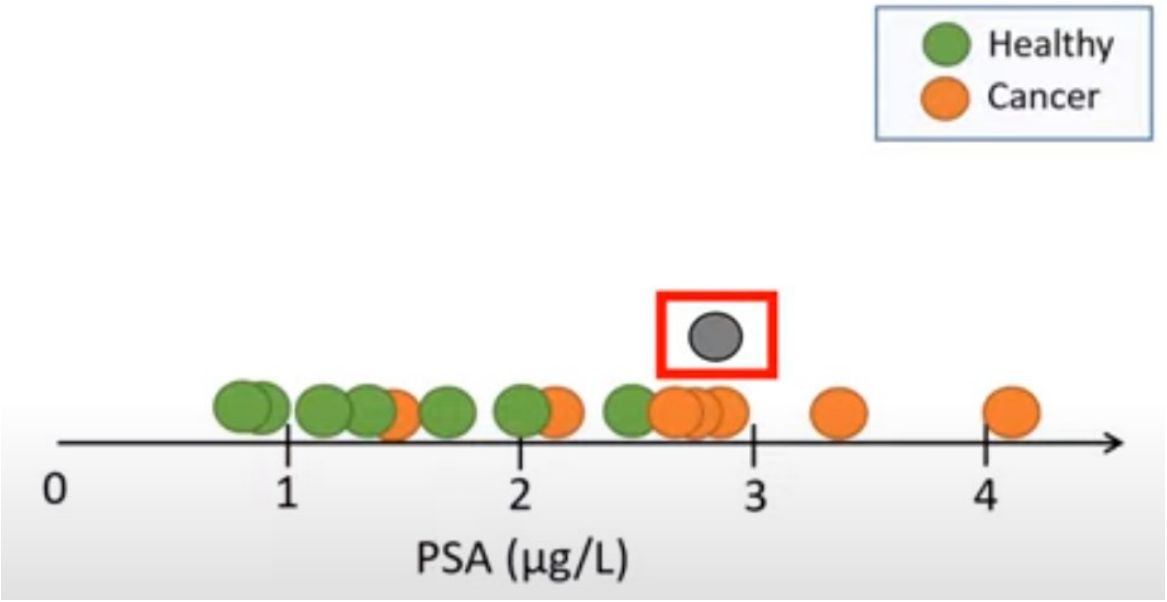
0,52939 YES

0,47061

Classificador Naive Bayes

- E quando as variáveis não são categóricas ?

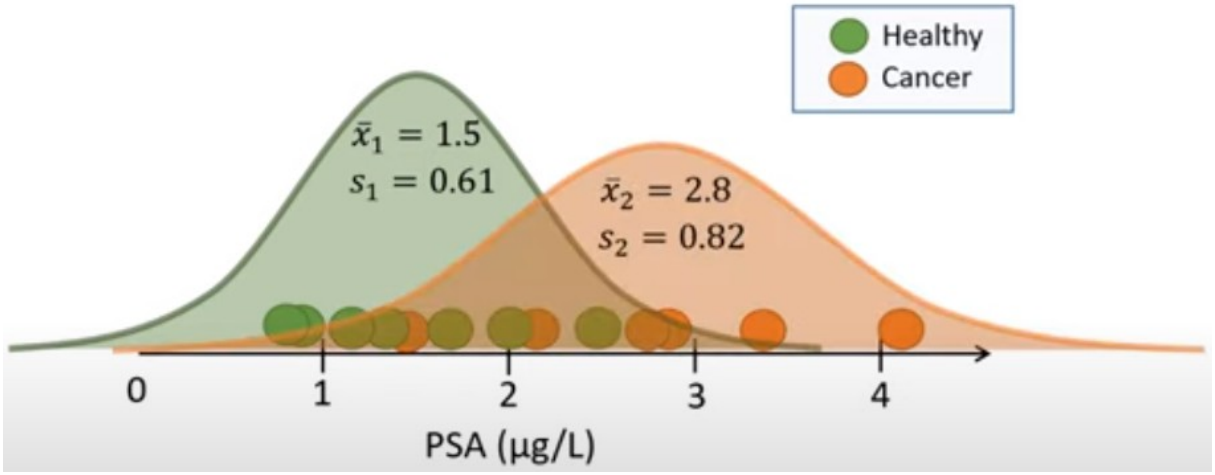
Status	PSA
Cancer	4.1
Cancer	3.4
Cancer	2.9
Cancer	2.8
Cancer	2.7
Cancer	2.1
Cancer	1.6
Healthy	2.5
Healthy	2.0
Healthy	1.7
Healthy	1.4
Healthy	1.2
Healthy	0.9
Healthy	0.8



Classificador Naive Bayes

- Distribuição Normal (Gaussiana)

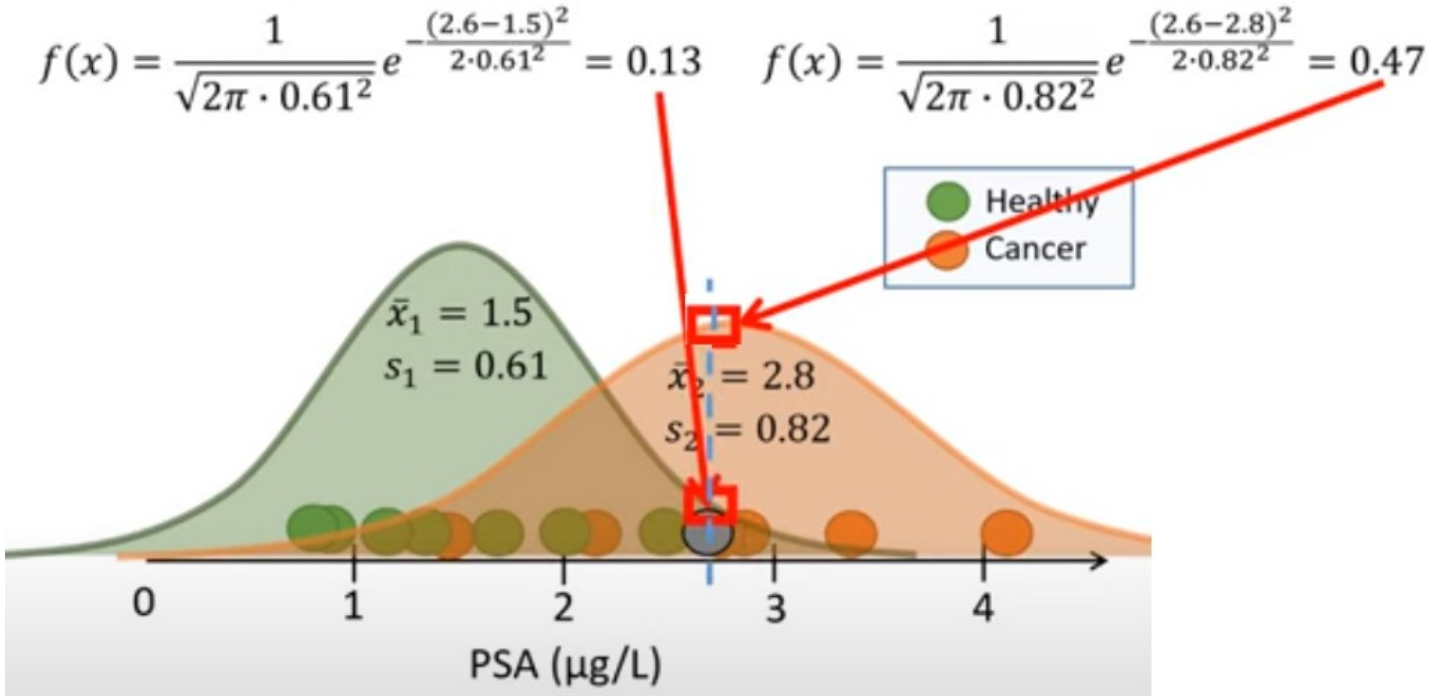
Status	PSA
Cancer	4.1
Cancer	3.4
Cancer	2.9
Cancer	2.8
Cancer	2.7
Cancer	2.1
Cancer	1.6
Healthy	2.5
Healthy	2.0
Healthy	1.7
Healthy	1.4
Healthy	1.2
Healthy	0.9
Healthy	0.8



Classificador Naive Bayes

- Distribuição Normal (Gaussiana)

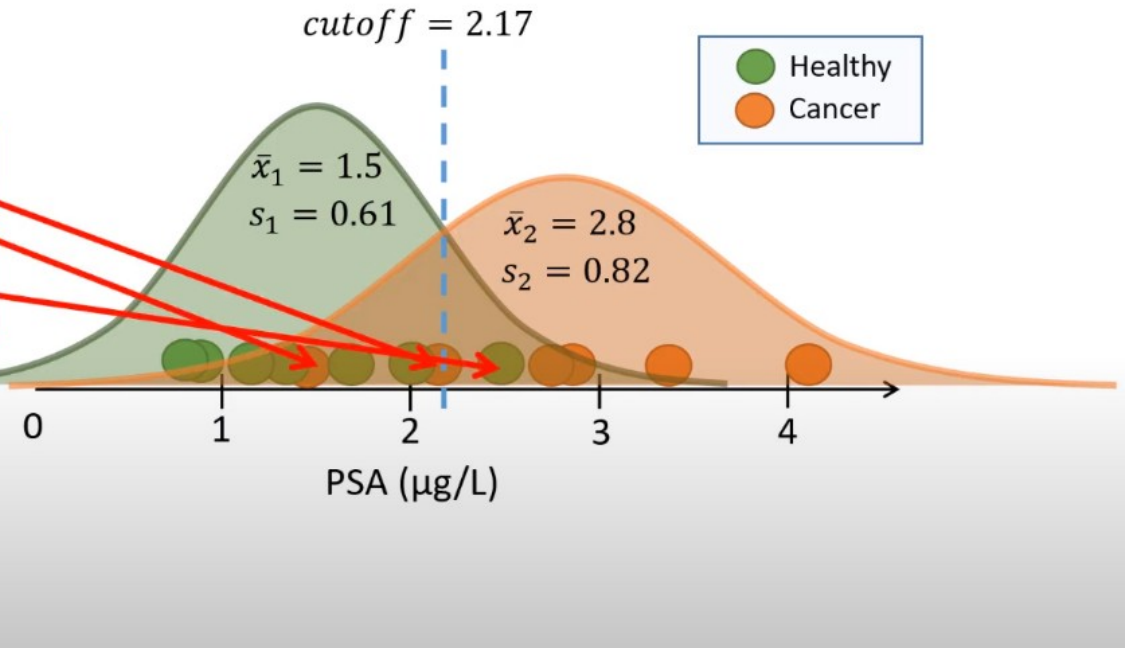
Status	PSA
Cancer	4.1
Cancer	3.4
Cancer	2.9
Cancer	2.8
Cancer	2.7
Cancer	2.1
Cancer	1.6
Healthy	2.5
Healthy	2.0
Healthy	1.7
Healthy	1.4
Healthy	1.2
Healthy	0.9
Healthy	0.8



Classificador Naive Bayes

- Distribuição Normal (Gaussiana)

Status	PSA	
Cancer	4.1	Cancer
Cancer	3.4	Cancer
Cancer	2.9	Cancer
Cancer	2.8	Cancer
Cancer	2.7	Cancer
Cancer	2.1	Healthy
Cancer	1.6	Healthy
Healthy	2.5	Cancer
Healthy	2.0	Healthy
Healthy	1.7	Healthy
Healthy	1.4	Healthy
Healthy	1.2	Healthy
Healthy	0.9	Healthy
Healthy	0.8	Healthy



Let’s code!

- Vamos implementar o Naive Bayes com o Scikit learn

Link: Tópico_02_Aprendizado_Supervisionado_Naive_Bayes.ipynb

Considerações Finais

- Vantagens
  - Implementação simples
  - Se ajusta bem com datasets pequenos

- Desvantagens:
  - Características devem ser independentes
  - Bases complexas normalmente apresentam dados dependentes
  - Se um atributo novo ocorrer no test, a probabilidade será zerada visto que não estava presente no treino