Naïve Bayes

Prof. Jefferson T. Oliva

Aprendizado de Máquina e Reconhecimento de Padrões (AM28CP)
Engenharia de Computação
Departamento Acadêmico de Informática (Dainf)
Universidade Tecnológica Federal do Paraná (UTFPR)
Campus Pato Branco

DAINF
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CÂMPUS PATO BRANCO

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Sumário

- Eventos dependentes e Independentes
- Permutabilidade
- Teorema de Bayes
- Classificador Naïve Bayes

Introdução

- Muitos problemas de classificação não são determinísticos
  - Relação entre atributos preditivos e classe é probabilística
  - Caso os dados possuem muito ruído (ou incerteza), a mesma entrada pode ser classificada de maneiras diferentes dependendo do modelo ou do conjunto de dados utilizado
  - Algumas informações importantes não são capturadas pelos atributos preditivos usados
    - Conjunto de atributos com limitações
    - Problemas na representação dos dados
    - Fatores externos

Introdução

- Exemplo: predizer se uma pessoal terá alguma doença cardíaca
  - Atributos preditivos
    - Peso
    - Exercícios frequentes
  - Ignorar outros eventos
    - Bebida
    - Tabagismo
    - Estresse
    - Fatores genéticos
    - ...

Sumário

- Suponha uma caixa com duas bolas brancas e 3 pretas

- Em um hipotético experimento, considere a remoção aleatória de uma bola de uma determinada cor, com representação de Bernoulli (X_1)
  - $X_1 = 0$: caso a bola retirada seja branca
  - $X_1 = 1$: caso a bola retirada seja preta

- Pela abordagem clássica, a probabilidade de ser retirada uma bola de determinada cor é:
  - Branca

$$P(X_1 = 0) = \frac{2}{5}$$

  - Preta

$$P(X_1 = 1) = \frac{3}{5}$$

- Após a primeira retirada, e sem reposição, caso seja retirada mais uma bola, qual seria a probabilidade de da segunda bola ser branca?
  - A probabilidade da segunda retirada (X_2) é condicionada ao resultado da primeira
    - Se a primeira bola retirada for branca (restando 1 branca e 3 pretas), a probabilidade é $\frac{1}{4}$
    - Caso a primeira bola seja preta (restando duas brancas e duas pretas), então a probabilidade é $\frac{2}{4}$
- A probabilidade do evento B ser condicionada ao evento A é representada por $P(B|A)$
  - $P(X_2 = 0|X_1 = 0) = \frac{1}{4}$
  - $P(X_2 = 0|X_1 = 1) = \frac{2}{4}$
  - O evento X_2 é dependente do evento X_1

Eventos dependentes e Independentes

Probabilidade conjunta

- A probabilidade conjunta dos eventos A e B é expressa pela regra do produto:

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

- A probabilidade conjunta também pode ser representada por $P(B \cap A) = P(B, A)$

- Considere um experimento, o lançamento de uma moeda honesta e a observância da face virada para cima

- Para isso, considere a representação Bernoulli das seguintes ocorrências
  - $X_1 = 0$: caso o resultado do lançamento seja coroa
  - $X_1 = 1$: caso o resultado do lançamento seja cara

- Em um lançamento honesto, teríamos as seguintes probabilidades
  - $P(X_1 = 0) = \frac{1}{2}$
  - $P(X_1 = 1) = \frac{1}{2}$

- Em seguida, a moeda é lançada novamente e o resultado representado por X_2
- A probabilidade da ocorrência de cara no segundo lançamento também é igual a $P(X_2 = 1) = \frac{1}{2}$
- Neste caso, o evento X_2 é independente de X_1, e, temos a seguinte regra do produto

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

ou

$$P(X_2, X_1) = P(X_2)P(X_1)$$

Permutabilidade

Permutabilidade

- Propriedade da alteração no ordenamento de realizações em uma sequência de eventos sem que a probabilidade conjunta seja alterada

- Considere o lançamento de 5 moedas não viciadas, com representação semelhante à apresentada anteriormente, sendo observado o seguinte resultado $(1, 0, 1, 1, 0)$, isto é, (cara, coroa, cara, cara, coroa)

- Os lançamentos são independentes e que a probabilidade conjunta desse evento é:

$$P(1,0,1,1,0) = P(X_1 = 1) \times P(X_2 = 0) \times P(X_3 = 1) \times P(X_4 = 1) \times P(X_5 = 0) = \frac{1}{2^5}$$

- Nessa situação, caso fosse observado o evento $(1, 1, 1, 0, 0)$, a probabilidade não seria alterada

Permutabilidade

- A ordem da ocorrências dos resultados cara não altera a probabilidade conjunta, desde que seja a mesma quantidade de resultados de “sucesso”
- A independência de uma sequência de eventos garante a permutabilidade da mesma
- A independência não é condição necessária para a permutabilidade, apenas suficiente
  - Ainda que uma sequência não seja formada por eventos independente, é possível que sejam permutáveis
- Considere a mesma representação do exemplo da caixa com bolas apresentado anteriormente para o caso da retirada de 5 bolas sem reposição
  - Como visto, esses eventos não são independentes, pois a probabilidade de um determinado evento na sequência, depende do resultado observado nos eventos anteriores

Permutabilidade

- Dessa forma, vamos verificar se o evento (1,0,1,1,0) é permutável
  - Inicialmente, vamos calcular a probabilidade $P(1,0,1,1,0)$

    $$P(1,0,1,1,0) = P(X_1 = 1) \times P(X_2 = 0|X_1 = 1) \times P(X_3 = 1|X_2 = 0, X_1 = 1) \times P(X_4 = 1|X_3 = 1, X_2 = 0, X_1 = 1) \times P(X_5 = 0|X_4 = 1, X_3 = 1, X_2 = 0, X_1 = 1)$$

    $$P(1,0,1,1,0) = \frac{3}{5} \times \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{10}$$

  - Cálculo da propriedade $P(1,1,1,0,0)$

    $$P(1,1,1,0,0) = P(X_1 = 1) \times P(X_2 = 1|X_1 = 1) \times P(X_3 = 1|X_2 = 1, X_1 = 1) \times P(X_4 = 0|X_3 = 1, X_2 = 1, X_1 = 1) \times P(X_5 = 0|X_4 = 0, X_3 = 1, X_2 = 1, X_1 = 1)$$

    $$P(1,1,1,0,0) = \frac{3}{5} \times \frac{2}{4} \times \frac{1}{3} \times \frac{2}{2} \times \frac{1}{1} = \frac{1}{10}$$

Teorema de Bayes

Teorema de Bayes

- Considerando os eventos A e B permutáveis, o termo $P(A \cap B)$ é igual a $P(B \cap A)$ e, dessa forma, pode ser escrita como

$$P(A \cap B) = P(B \cap A)$$

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

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

- A probabilidade $P(A)$ é denominada probabilidade a priori
  - Informação sobre o evento A antes que se soubesse algo sobre o evento B

- Quando se tem conhecimento sobre B, a probabilidade relacionada ao evento A deve ser atualizada pela probabilidade do evento B
  - A probabilidade $P(A|B)$ é agora denominada probabilidade a posteriori

Teorema de Bayes

- Sejam dois eventos A e B
  - A: atributo alvo (presença de doença)
    - Possíveis valores: presença ou ausência
  - B: atributo preditivo (resultado do exame)
    - Possíveis valores: positivo ou negativo
  - $P(A)$: probabilidade do evento A (presença da doença) ocorrer
  - $P(B)$: probabilidade do evento B (exame positivo) ocorrer

Teorema de Bayes

- Probabilidade a priori pode ser estimada pela frequência

Paciente	Exame	Doença
001	positivo	presente
002	negativo	presente
003	negativo	ausente
004	positivo	presente
005	positivo	ausente
006	positivo	presente
007	negativo	ausente
008	negativo	presente
009	positivo	ausente
010	positivo	presente

- $P(negativo) = ?$
- $P(positivo) = ?$
- $P(presente) = ?$
- $P(ausente) = ?$

Teorema de Bayes

- Probabilidade a priori pode ser estimada pela frequência

Paciente	Exame	Doença
001	positivo	presente
002	negativo	presente
003	negativo	ausente
004	positivo	presente
005	positivo	ausente
006	positivo	presente
007	negativo	ausente
008	negativo	presente
009	positivo	ausente
010	positivo	presente

- $P(\mathit{negativo}) = 0,4$
- $P(\mathit{positivo}) = 0,6$
- $P(\mathit{presente}) = 0,6$
- $P(\mathit{ausente}) = 0,4$

Teorema de Bayes

- As frequências a priori são fáceis de serem estimadas
  - $P(B)$: probabilidade do resultado do exame ser positivo
  - $P(A)$: probabilidade do paciente estar doente (doença presente)
  - $P(B|A)$: probabilidade do resultado do exame ser positivo dado que o paciente está doente
- A probabilidade posteriori é considerada difícil de ser estimada
  - $P(A|B)$: probabilidade do paciente estar doente dado que seu exame deu positivo
  - O Teorema de Bayes tenta estimar esse tipo de probabilidade

Teorema de Bayes

- Teorema de Bayes permite o cálculo da probabilidade a posteriori de um evento
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- $posteriori = \frac{\text{verosimilhança} \times priori}{\text{evidência}}$
  - Verosimilhança - $P(D|H)$: probabilidade de observarmos D, supondo que a hipótese H seja verdadeira
  - Priori - $P(H)$: probabilidade de uma hipótese ser verdade antes da coleta dos dados
  - Evidência - $P(D)$: probabilidade de observarmos esses dados sob todas as hipóteses possíveis
  - Posteriori - $P(H|D)$: probabilidade da hipótese ser verdadeira dado o conjunto de dados coletados

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$$

Classificador Naïve Bayes

Classificador Naïve Bayes

- Todos os classificadores naïve Bayes (NB) assumem que o valor de uma característica particular é independente do valor de qualquer outra característica, dada a variável de classe
  - Uma fruta pode ser considerada uma laranja se a mesma for laranjada, redonda e ter diâmetro aproximadamente de 10 cm
  - O classificador NB considera que cada uma dessas características contribui independentemente para a probabilidade de que essa fruta seja uma laranja, independentemente de quaisquer correlações possíveis entre as características de cor, forma e diâmetro
- Em diversas aplicações, a estimativa de parâmetros para modelos NB utiliza o método da máxima verossimilhança

Classificador Naïve Bayes

- NB é um modelo de probabilidade condicional
  - Dada uma instância de problema a ser classificada, representada por um vetor x representando n características independentes
    - A seguinte instância de probabilidades é atribuída:

$$p(C_k|x_1, ..., x_n)$$

- Com o teorema de Bayes, a probabilidade condicional pode ser decomposta na seguinte equação

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

Classificador Naïve Bayes

- O numerador é equivalente ao modelo de probabilidade conjunta

$$p(C_k|x_1, ..., x_n)$$

- Uso da regra da cadeia para aplicações repetidas da definição de probabilidade condicional

$$p(C_k|x_1, ..., x_n) = p(x_1, ..., x_n, C_k) =$$
$$p(x_1|x_2..., x_n, C_k)p(x_2, ..., x_n, C_k) =$$
$$p(x_1|x_2..., x_n, C_k)p(x_2|x_3..., x_n, C_k)p(x_3..., x_n, C_k) = ... =$$
$$p(x_1, ..., x_n, C_k)p(x_2|x_3..., x_n, C_k)...p(x_{n-1}|x_n..., x_n, C_k)p(C_k)$$

Classificador Naïve Bayes

- As suposições de independência condicional "naïves" (ingênuas) são consideradas
  - Suponha que todas as características x não mutualmente independentes, condicionais à classe C_k

$$
\begin{aligned}
p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\
&\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\
&\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,,
\end{aligned}
$$

$$
P(B \mid A_1, \ldots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}
$$

$$
\hat{y} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).
$$

Classificador Naïve Bayes

- Naïve Bayes é um algoritmo que utiliza o Teorema de Bayes com a hipótese de independência de atributos

- Por que estimar independência entre atributos $A_1, ..., A_n$?
  - Estimar probabilidades conjuntas $P(A_1, A_2, ..., A_n)$ e $P(A_1, A_2, ..., A_n|B)$ demandaria uma quantidade mínima de exemplos de cada combinação possível de valores de $A_1, ..., A_n$
    - Impraticável, especialmente para quantidades elevadas de atributos!

- Apesar da hipótese ser quase sempre violada, o método (Naïve Bayes) se mostra bastante competitivo na prática

Classificador Naïve Bayes

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

$$P(B \mid A_1, \ldots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}$$

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Classificador Naïve Bayes

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	???

P(Yes|Sunny, Cool, High, True) = (2/9 × 3/9 × 3/9 × 3/9 × 9/14) / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = (3/5 × 1/5 × 4/5 × 3/5 × 5/14) / P(Sunny, Cool, High, True)

P(Yes|Sunny, Cool, High, True) = **0.0053** / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = **0.0206** / P(Sunny, Cool, High, True)

Play = No

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

- O que acontece se um determinado valor de atributo não aparece na base de treinamento, mas aparece no exemplo de teste?
  - Por exemplo: "Outlook = Overcast" para classe "No"
    - Probabilidade correspondente será zero: $P(Overcast|No) = 0$
    - Probabilidade a posteriori será também zero: $P(No|Overcast, ...)$
  - Não importa as probabilidades referentes aos demais atributos
  - Muito radical, especialmente considerando que a base de treinamento pode não ser totalmente representativa
    - Classes minoritárias com instâncias raras

- Uma solução:
  - Adicionar unidades fictícias para cada combinação de valor-classe (estimador de Laplace)
    - Probabilidades zero não existirão
    - Exemplo:
      $sunny = \frac{3+1}{5+3}$, $overcast = \frac{0+1}{5+3}$ e $rainy = \frac{2+1}{5+3}$
  - Obs.: a inclusão de uma unidade fictícia deve ser adicionada para todas as classes para evitar viés nas probabilidades de apenas uma classe

- Solução mais geral (Estimativa m):
  - Adição de múltiplas unidades fictícias para cada combinação de valor-classe
  - Exemplo
    $sunny = \frac{3+\frac{m}{3}}{5+m}$, $overcast = \frac{0+\frac{m}{3}}{5+m}$ e $rainy = \frac{2+\frac{m}{3}}{5+m}$
- Solução ainda mais geral: substituir o termo $\frac{1}{n}$ no numerador (onde n é quantidade de valores do atributo) por uma probabilidade p qualquer

- Exclusão de exemplos com valores ausentes do conjunto de treinamento

- Ou considerar apenas os atributos sem valores ausentes

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	???

Verossimilhança para "Yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

Verossimilhança para "No" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

Probabilidade Estimada ("Yes") = 0.0238 / (0.0238 + 0.0343) = 41%

Probabilidade Estimada ("No") = 0.0343 / (0.0238 + 0.0343) = 59%

- Alternativa 1: discretização

- Alternativa 2: assumir ou estimar alguma função de densidade de probabilidade para estimar as probabilidades
  - Distribuição Gaussiana (normal)

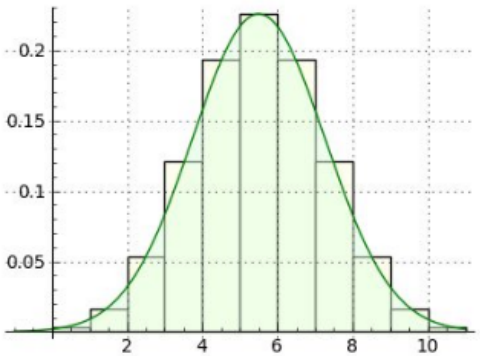
$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  - Exemplo

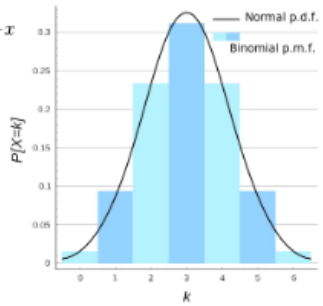
Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, ...	85, ...	80, ...	95, ...					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

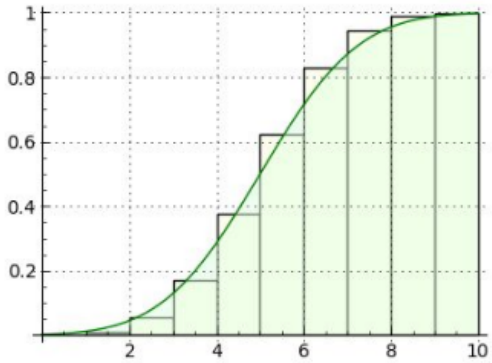
$$p(temperature|66, yes) = \frac{1}{\sqrt{2\pi 6,2^2}} \exp\left(-\frac{(66-73)^2}{26,2^2}\right) = 0,0340$$

- A distribuição binomial com parâmetros n e p, denotada $B(n, p)$ é a distribuição de probabilidade discreta do número de sucessos em uma sequência de n experimentos independentes
  - Se n for grande o suficiente, então a assimetria da distribuição não é muito grande
  - Uma aproximação razoável é dada pela distribuição normal



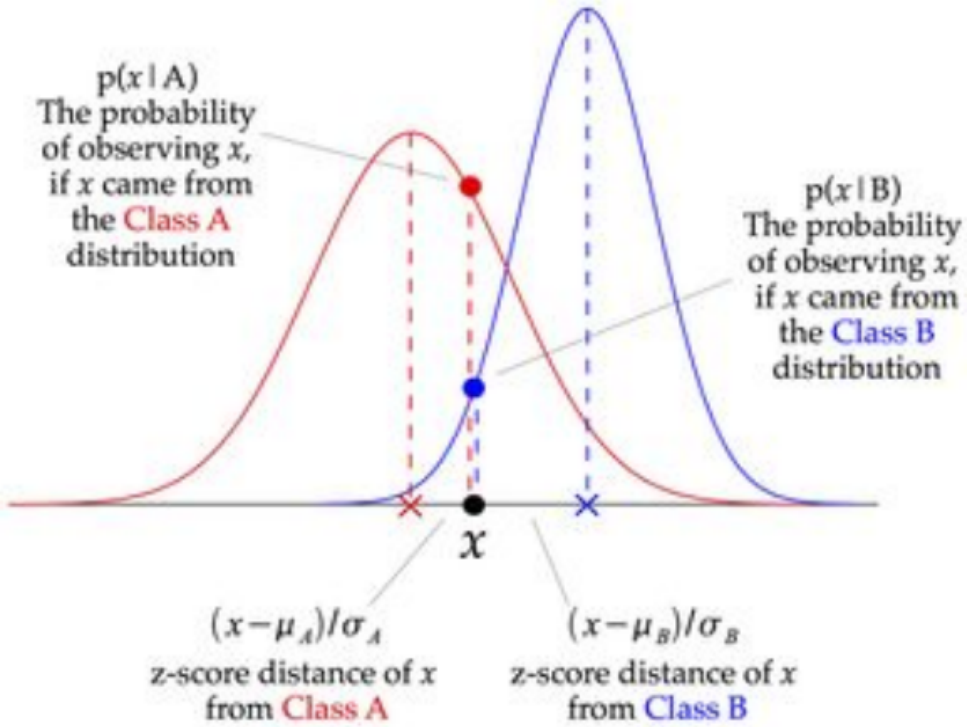
$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}$$
$$= (p + (1-p))^n$$
$$= (1)^n$$
$$= 1$$

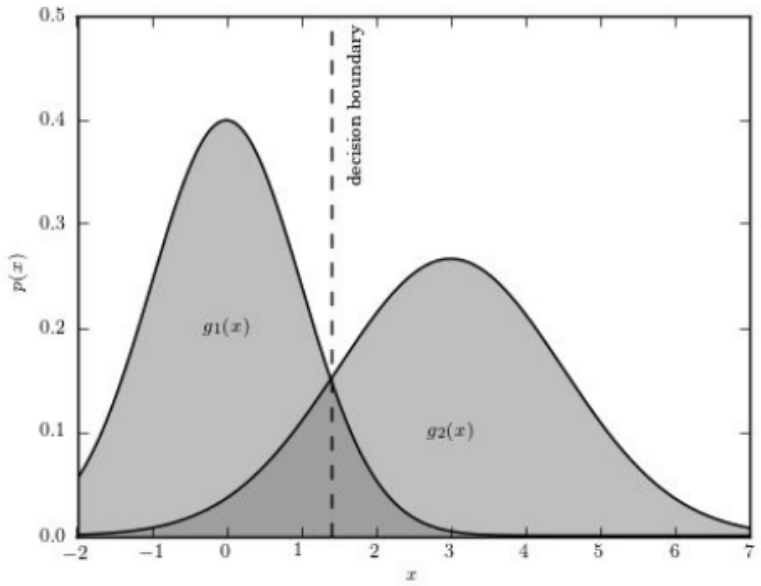




Classificador Naïve Bayes

Atributos numéricos





Classificador Naïve Bayes

- Características do classificador Naïve Bayes
  - Robusto a ruídos e atributos irrelevantes
  - Capaz de classificar instâncias com valores ausentes
  - Assume que atributos são igualmente importantes
  - Desempenho pode ser afetado pela presença de atributos correlacionados

Referências I


BISHOP, C. M.
Pattern Recognition and Machine Learning.
Springer, 2006.


DE CARVALHO, A. P. L. F.
Métodos probabilísticos. Aprendizado de Máquina.
Slides. Ciência de Computação e Matemática Computacional.
ICMC/USP, 2015.


CASANOVA, D.
Naïve Bayes. Aprendizado de Máquina.
Slides. Engenharia de Computação. Dainf/UTFPR, 2020.


DUDA, Richard O.; HART, Peter E.; STORK, David G.
Pattern classification.
2nd ed. New York, NY: J. Wiley & Sons, 2001.

Referências II


RASCHKA, S.; MIRJALILI, V.
Python Machine Learning.
Packt, 2017.