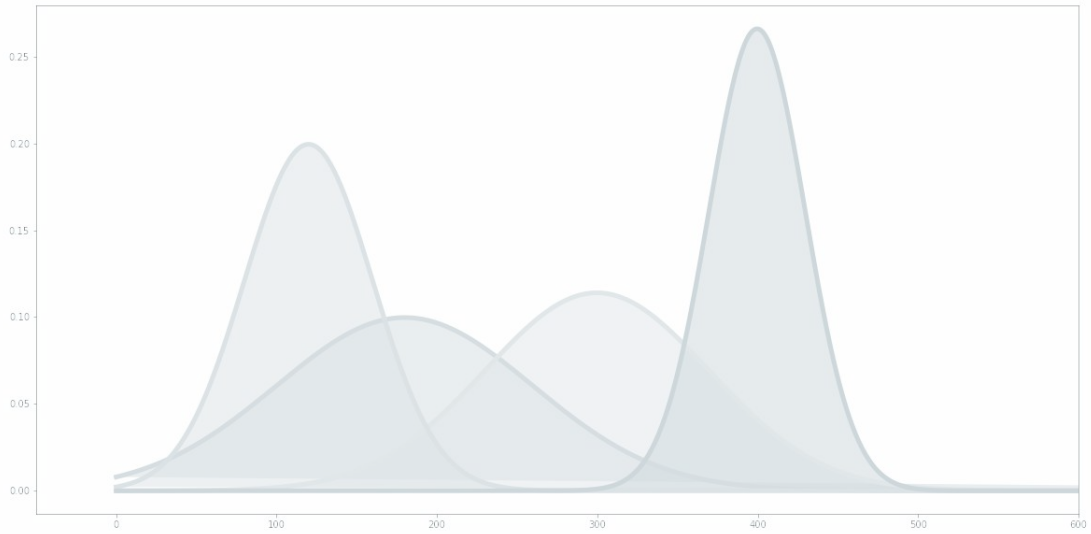


Aula 11

Naïve Bayes

Rafael Geraldeli Rossi

Conteúdo

Introdução

- É um dos classificadores probabilísticos mais simples
- Baseado no teorema de Bayes (cujo autor é Thomas Bayes), que descreve a probabilidade de ocorrer um evento baseado em condições relacionadas ao evento
- Chamado de "*naïve*" (do inglês ingênuo, ou simples) porque assume que a probabilidade de ocorrer cada valor de atributo em uma dada classe é independente dos valores dos outros atributos
- Isto é feito para simplificar as computações envolvidas nos cálculos das probabilidades

Teorema de Bayes

- O teorema de Bayes cujo autor é Thomas Bayes (por isso o nome) diz que:

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

na qual

- A e B são eventos
- $P(A)$ é a probabilidade de ocorrer um evento A e $P(B)$ é a probabilidade de ocorrer um evento B ($P(B) \neq 0$)
- $P(A|B)$ é a probabilidade de ocorrer um evento A dado que ocorreu um evento B (probabilidade condicional)
- $P(B|A)$ é a probabilidade de ocorrer um evento B dado que ocorreu o evento A

Exemplo do Teorema de Bayes

Probabilidade de ter câncer aos 65 anos

- Suponha que a probabilidade de ter câncer, independente da idade é de 1%
- Suponha que a probabilidade de se ter 65 anos é de 0.2%
- Suponha que, dentre as pessoas que têm câncer, a probabilidade de se ter 65 anos é de 0.5%
- Dada essas informações, qual a probabilidade de uma pessoa ter câncer dada que ela tem 65 anos?

$$P(cancer|65) = \frac{p(65|cancer) * p(65)}{p(cancer)} = \frac{0.5\% * 1\%}{0.2\%} = 2.5\%$$

Introdução

- Para o problema de aprendizado de máquina:
  - Seja X uma instância (exemplo ou objeto) descrita por n atributos $(X = (x_1, x_2, \ldots, x_n))$
  - Seja H uma hipótese, por exemplo, que a instância X pertence a uma classe C
  - Para problemas de classificação, o objetivo é determinar $P(H|X)$ ou $P(C|X)$, ou seja, a probabilidade da instância X pertencer a classe C dado que conhecemos a descrição dos atributos de X
  - $P(H|X)$ é a probabilidade *a posteriori* de H condicionada a X
  - $P(H)$ é a probabilidade *a priori* de H, ou seja a probabilidade de ocorrer a hipótese independente dos valores dos atributos (independente de X)

- Retomando o Teorema de Bayes instanciado para o problema de aprendizado supervisionado

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- $P(H)$, $P(X|H)$ e $P(X)$ podem ser estimados utilizando os dados o conjunto de dados de treinamento

Naïve Bayes

- O algoritmo de Classificação Naïve Bayes segue os seguintes passos
  1. Seja D um conjunto de exemplos de treinamento associados com os respectivos rótulos de classe, cada exemplo tendo n atributos mais a classe $(X = (x_1, x_2, \ldots, x_n, x_{n+1}))$
  2. Considerando m classes $C = (c_1, c_2, \ldots, c_m)$, dado uma tupla X, o classificador irá predizer que X pertence a classe que possuir a maior valor de probabilidade a posteriori condicionado a X, ou seja, um novo exemplo será rotulado como c_i se

$$P(c_i|X) > P(c_j|X) \text{ para } 1 \leq i, j \leq m,\ j \neq i$$

Naïve Bayes

- O algoritmo de classificação Naïve Bayes segue os seguintes passos

3. $P(c_i) = |c_{i,D}|/|D|$, na qual $|c_{i,D}|$ é o número de exemplos de treinamento das classes c_i em D

4. Dado um conjunto de dados com muitos atributos, o cálculo $P(X|c_i)$ é inviável, pois dificilmente haverá vários exemplos de treinamento com os mesmo valores de atributos (exemplos idênticos)

Naïve Bayes

- O algoritmo de Classificação Naïve Bayes segue os seguintes passos
  5. Neste caso, teria-se que calcular a probabilidade de cada exemplo distinto ocorrer para cada classe $\rightarrow$ muitas probabilidades seriam calculadas
  6. Além disso, para calcular a probabilidade de um novo exemplo pertencer a uma classe, teria-se que que ter exatamente um exemplo idêntico no conjunto de treinamento para que se tenha previamente calculado essa probabilidade

Naïve Bayes

- Para viabilizar o cálculo de $P(X|c_i)$, a suposição "*naïve*" de independência dos atributos é realizada:

$$P(X|c_i) = \prod_{k=1}^{n} P(x_k|c_i) = P(x_1|c_i) \times P(x_2|c_i) \times \cdots \times P(x_n|c_i) \quad (1)$$

- **ISTO SIGNIFICA QUE**: a probabilidade de ocorrer uma instância X dada a classe c_i é dada pelo produtório da probabilidade de cada valor de um atributo de X pertencer a classe c_i

Construção do Modelo

- As probabilidades $P(x_1|c_i), P(x_2|c_i), \ldots, P(x_n|c_i)$ podem ser facilmente estimadas a partir conjunto de treinamento
- Considere x_k o valor do atributo k na tupla X
- Para calcular $P(X|c_i)$ considera-se
  1. Se o atributo k é categórico, então $P(x_k|c_i)$ é o número de exemplos da classe c_i em D tendo o valor x_k, dividido por $|c_{i,D}|$
  2. Se o atributo k é contínuo, tipicamente assume-se uma distribuição gaussiana (distribuição normal) com média μ e desvio padrão σ definida por

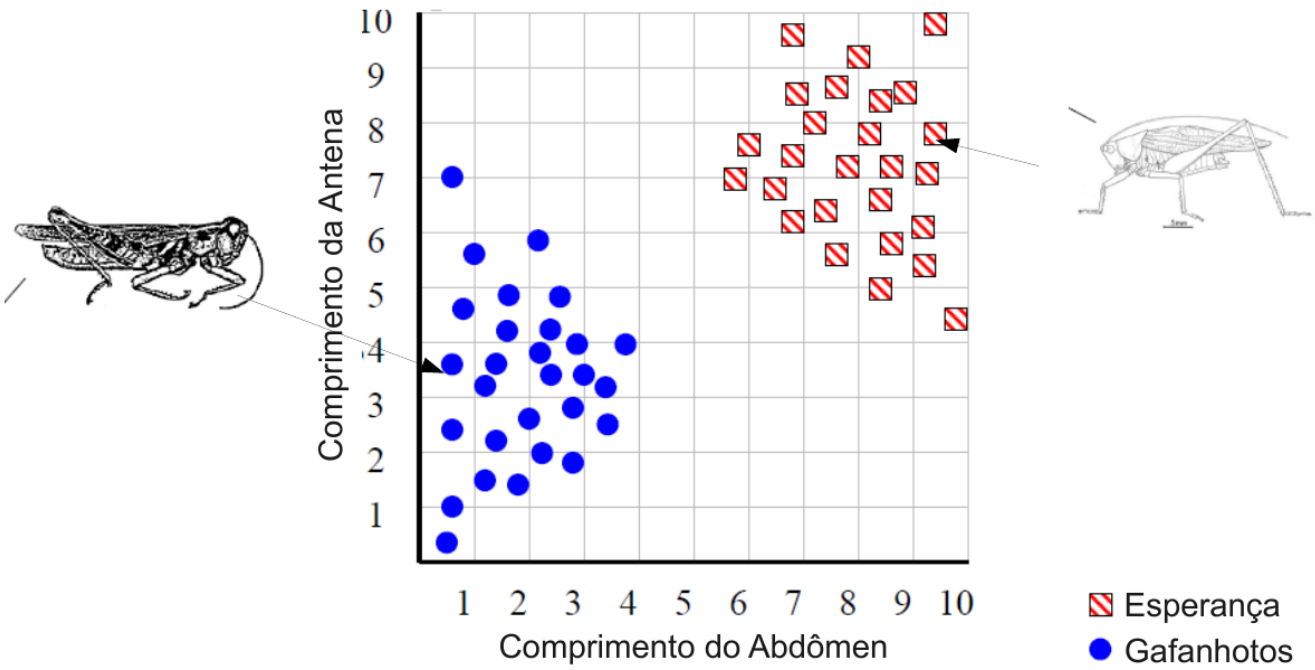
$$g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

tal que

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \tag{3}$$

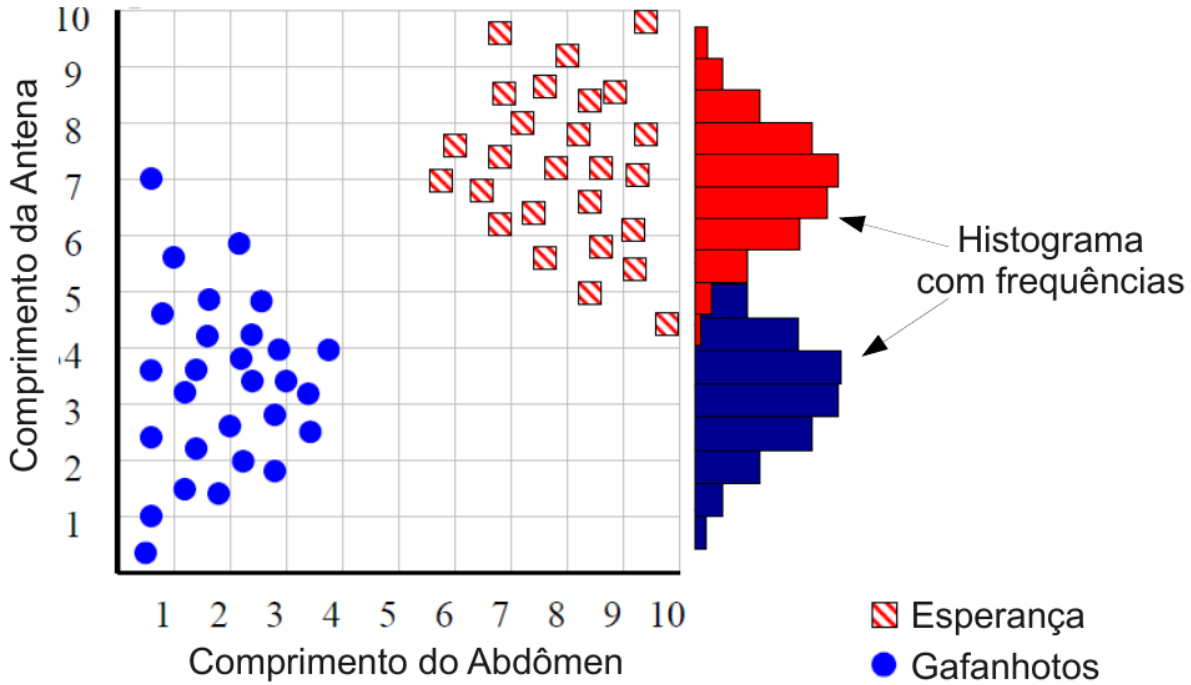
Naïve Bayes - Atributos Numéricos

- Tentando obter a probabilidade do tamanho da asa ocorrer para a classe `Esperança` ou `Gafanhoto`



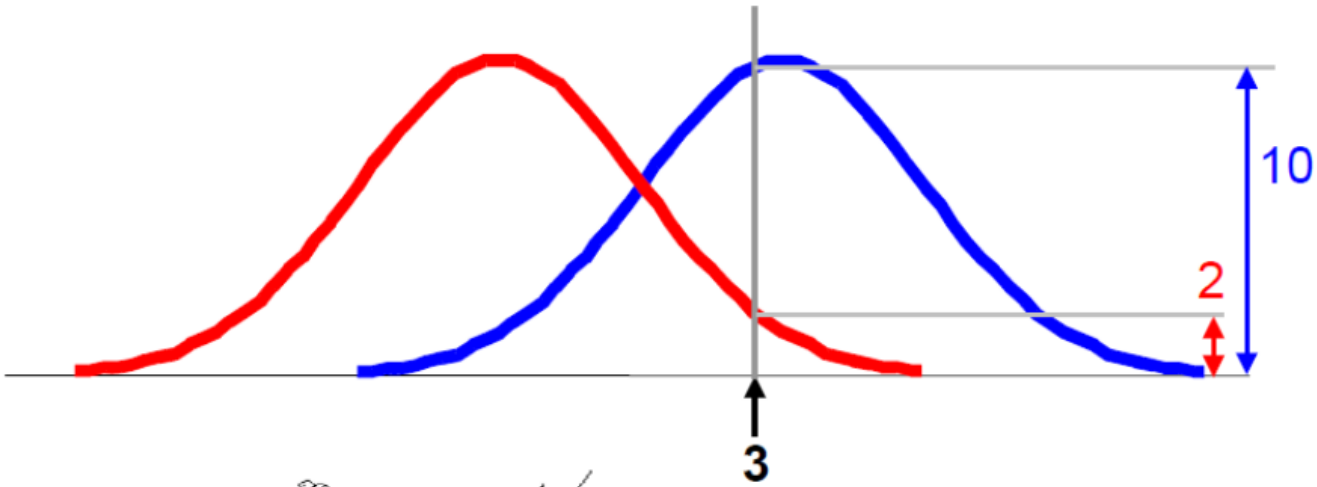
Naïve Bayes - Atributos Numéricos

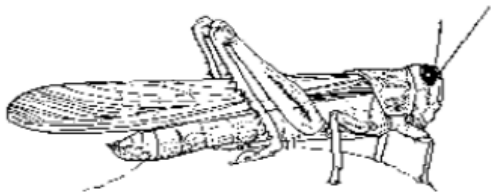
- Tentando obter a probabilidade do tamanho da asa ocorrer para a classe `Esperança` ou `Gafanhoto`



Naïve Bayes - Atributos Numéricos

- Tentando obter a probabilidade do tamanho da asa ocorrer para a classe `Esperança` ou `Gafanhoto`





Comp. da antena é **3**

Exemplo para atributos categóricos

Tabela: Conjunto de dados *Weather* [Witten and Frank, 2005].

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**

Exemplo para atributos categóricos

Tabela: Conjunto de dados *Weather* [Witten and Frank, 2005].

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**

Exemplos para atributos categóricos (continuação)

Tabela: Contagens e probabilidades do conjunto de dados *Weather* com atributos categóricos

Outlook			**Termperature**			**Humidity**			**Windy**			**Play**	
	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								

Exemplo para atributos continuos

Tabela: Conjunto de dados *Weather* com alguns atributos numéricos

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Exemplo para atributos continuos (continuação)

Tabela: Contagens e probabilidades do conjunto de dados *Weather* com alguns atributos numéricos

Outlook			**Temperature**			**Humidity**			**Windy**			**Play**	
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	μ	73	74,6	μ	79,1	86,2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	σ	6,2	7,9	σ	10,2	9,7	true	3/9	3/5		
rainy	3/9	2/5											

Classificação

- Uma vez obtidas as probabilidade de cada valor de atributo para cada uma das classes, pode-se calcular $P(X|c_i)$ e consequentemente $P(X|c_i)$

$$P(X|c_i) = \prod_{k=1}^{n} P(x_k|c_i) = P(x_1|c_i) \times P(x_2|c_i) \times \cdots \times P(x_n|c_i)$$

$$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)}$$

Exemplo com atributos categóricos

- Vamos utilizar os dados da Tabela 3
- Exemplo a ser classificado

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	???

$$X = (Outlook = sunny, Temperature = cool, Humidity = high, Wind = true)$$

$$P(Play = yes|X) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0053$$

$$P(Play = no|X) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0206$$

- Note que $P(Play = yes|X)$ e $P(Play = no|X)$ não são probabilidades reais, uma vez que a soma não é 1
- Portanto, para a obtenção das probabilidades há um passo adicional

$$P(Play = yes|X) = \frac{0,0053}{0,0053 + 0,0206} = 20,5\%$$

$$P(Play = no|X) = \frac{0,0206}{0,0053 + 0,0206} = 79,5\%$$

- Como $P(Play = no|X) > P(Play = yes|X)$, o exemplo será classificado como sendo da classe **no**

Exemplo com atributos categóricos e numéricos

- Vamos utilizar os dados da Tabela 5
- Exemplo a ser classificado

Outlook	Temperature	Humidity	Windy	Play
sunny	66	90	true	???

$$X = (Outlook = sunny, Temperature = 66, Humidity = 90, Wind = true)$$

$$P(Play = yes|X) = \frac{2}{9} \times 0,0340 \times 0,0221 \times \frac{3}{9} \times \frac{9}{14} = 0,000036$$

$$P(Play = no|X) = \frac{3}{5} \times 0,0279 \times 0,0381 \times \frac{3}{5} \times \frac{5}{14} = 0,000136$$

$$P(Play = yes|X) = \frac{0,000036}{0,000036 + 0,000198} = 25,0\%$$

$$P(Play = no|X) = \frac{0,000108}{0,000036 + 0,000108} = 75,0\%$$

- Como $P(Play = no|X) > P(Play = yes|X)$, o exemplo será classificado como sendo da classe **no**

Frequência Zero

- Se não foi encontrada nenhuma ocorrência do valor de um atributo para uma classe, a probabilidade será 0

- A probabilidade igual a 0 cancela o efeito de outras probabilidades envolvidas no cálculo da probabilidade de uma instância X pertencer a classe C_i

- Estimador de Laplace
  - Adicionar **uma** unidade fictícia para cada valor de um atributo para cada uma das classes
  - Considere o atributo `outlook` do conjunto de dados `Weather` na qual a frequência do valor `overcast` para a classe `no` é 0 De acordo com o Estimador de Laplace, as novas probabilidades $P(X|C)$ para os valores do atributo `outlook` são

$$Sunny : \frac{3+1}{5+3} \quad Overcast : \frac{0+1}{5+3} \quad Rainy : \frac{2+1}{5+3}$$

  - Este procedimento deve ser realizado para todas as classes

- Estimativa m
  - Solução mais geral
  - Adicionar múltiplas unidade fictícias para cada valor de um atributo para cada uma das classes

$$P(x_i|C_j) = \frac{n_c + mp}{n + m}$$

na qual n é o número total de instâncias da classe C_j, n_c é o número de exemplos de treinamento da classe C_j que possuem o valor x_i, m é conhecido como tamanho da amostra (número de unidades fictícias), e p é um parâmetro definido pelo usuário

- Estimativa m
  - Considere o atributo `outlook` do conjunto de dados *Weather* na qual a frequência do valor `overcast` para a classe `no` é 0 De acordo com a Estimativa m, considerando $p = 1/3$, as novas probabilidades $P(X|C)$ para os valores do atributo `outlook` são

$$Sunny : \frac{3 + \frac{m}{3}}{5 + m} \quad Overcast : \frac{0 + \frac{m}{3}}{5 + m} \quad Rainy : \frac{2 + \frac{m}{3}}{5 + m}$$

Valores Ausentes

- **Para construir o modelo**: se uma instância de treinamento possui um valor ausente, este simplesmente não é incluído no cálculo das frequências

- **Para predizer a classe de um novo exemplo:** se uma instância de teste possui um valor ausente, este deve ser omitido do cálculo da probabilidade

Discussão

- **Vantagens**
  - Abordagem simples
  - Pode produzir resultados bons, podendo obter melhores resultados que algoritmos mais complexos
  - Naturalmente suporta aprendizado incremental
  - Rápido
- **Desvantagens**
  - Atributos redundantes podem prejudicar a classificação
  - O fato de assumir uma distribuição normal para atributos numéricos é uma restrição
    - Utilizar outra forma de distribuição se conhecida
    - Discretizar os dados
    - ...

Material Complementar

- Distribuição Normal

  `https://www.ime.usp.br/~hbolfar/aula_2013/Aula6-A12012.pdf`

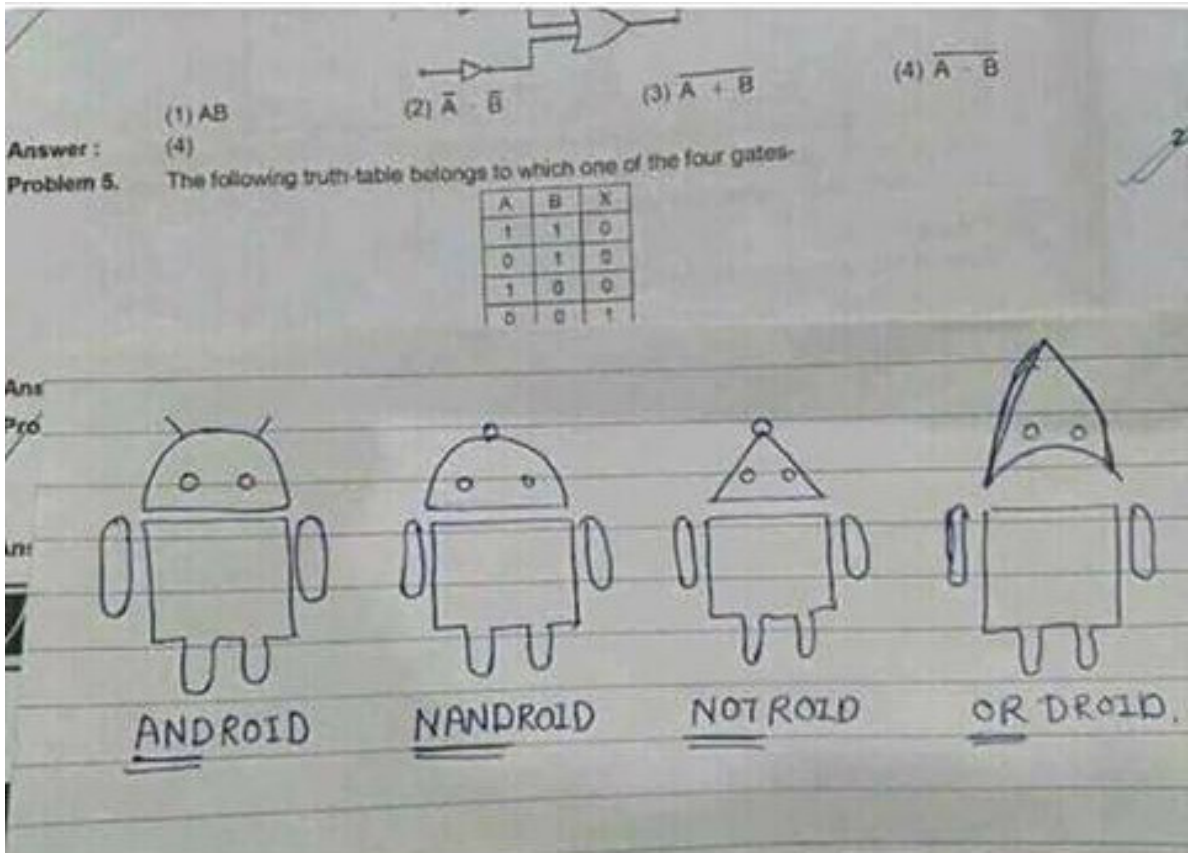
- Naive Bayes classifier

  `https://en.wikipedia.org/wiki/Naive_Bayes_classifier`

- All about Naive Bayes

  `https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf`

Imagem do Dia



Inteligência Artificial
http://lives.ufms.br/moodle/

Rafael Geraldeli Rossi
rafael.g.rossi@ufms.br

Slides baseados em [Han et al., 2011], [Tan et al., 2005] e [Witten and Frank, 2005]

Referências Bibliográficas I


Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques.
The Morgan Kaufmann Series in Data Management Systems.
Elsevier.


Tan, P.-N., Steinbach, M., and Kumar, V. (2005).
Introduction to Data Mining.
Addison-Wesley.


Witten, I. H. and Frank, E. (2005).
Data Mining: Practical machine learning tools and techniques.
Morgan Kaufmann, 2 edition.