

APRENDIZADO DE MÁQUINAS

Aprendizado probabilístico

TÓPICOS

1. Paradigmas de Aprendizado de Máquina
2. Métodos Bayesianos
3. Teorema de Bayes
4. Aprendizado Bayesiano
5. Classificador Naive-Bayes
6. Exemplo: Problema da Balança
7. Análise do Algoritmo

PARADIGMAS DE APRENDIZADO DE MÁQUINAS

1. Simbólico (árvores de decisão, regras ou rede semântica)
2. Estatístico (aprendizado Bayesiano)
3. Baseado em Exemplos (Nearest Neighbours e raciocínio baseado em casos)
4. Conexionista (redes neurais)
5. Evolutivo (algoritmos genéticos)



Fonte: Freepik.com

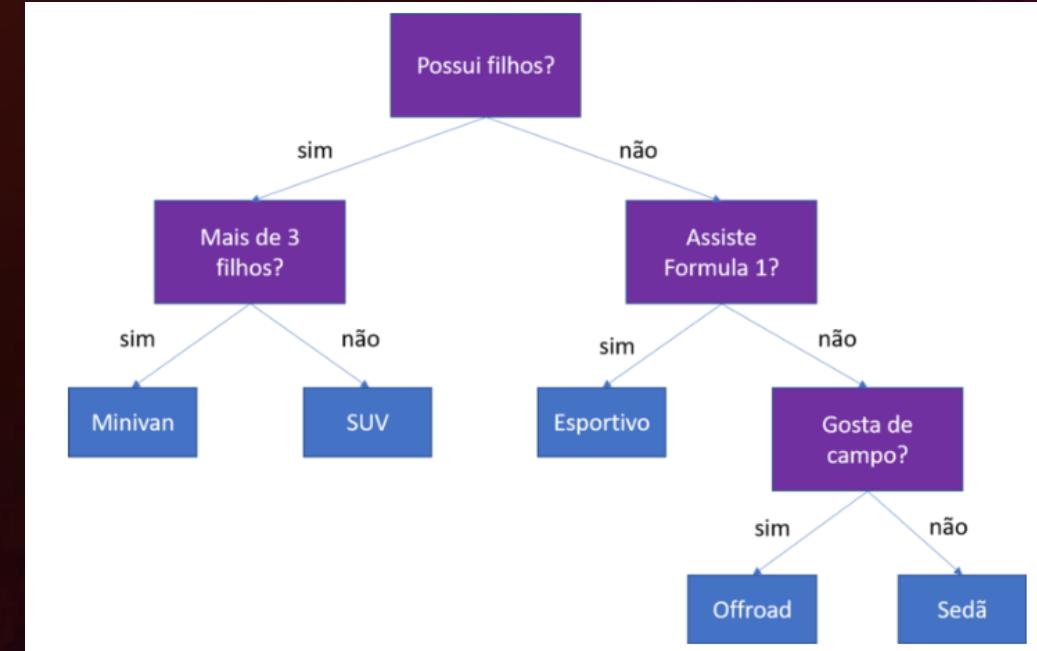
PARADIGMAS DE APRENDIZADO DE MÁQUINAS

Simbólico

Neste paradigma, um conceito é representado em uma estrutura simbólica e o aprendizado é realizado através da apresentação de exemplos e contraexemplos deste conceito.

As estruturas simbólicas estão tipicamente representadas em alguma expressão lógica, como por exemplo, regras de produção.

Exemplos de técnicas que utilizam este paradigma são:
Agentes Inteligentes e Árvores de Decisão



Fonte: Medium.com

PARADIGMAS DE APRENDIZADO DE MÁQUINAS

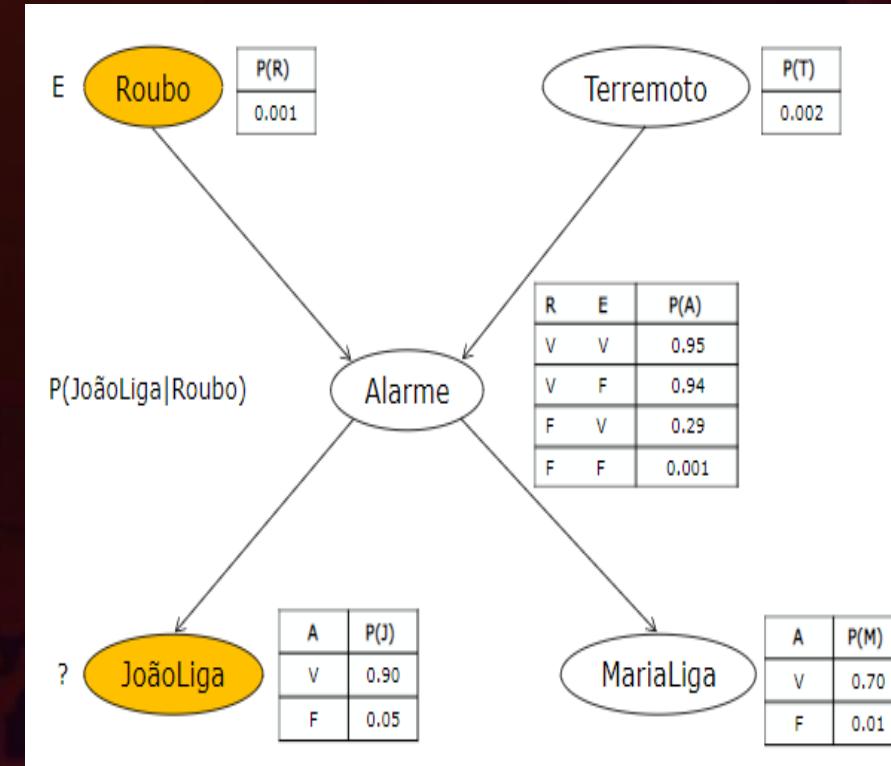
Estatístico

No paradigma estatístico é utilizado um modelo estatístico que encontre uma hipótese que possua uma boa aproximação do conceito a ser induzido.

O aprendizado consiste em encontrar os melhores parâmetros para o modelo.

Estes modelos podem ser paramétricos (quando fazem alguma suposição sobre a distribuição dos dados, ou podem ser não paramétricos, quando não fazem suposição sobre a distribuição dos dados).

Dentre os modelos estatísticos utilizados em aprendizagem de máquina, podemos destacar os modelos Bayesianos.



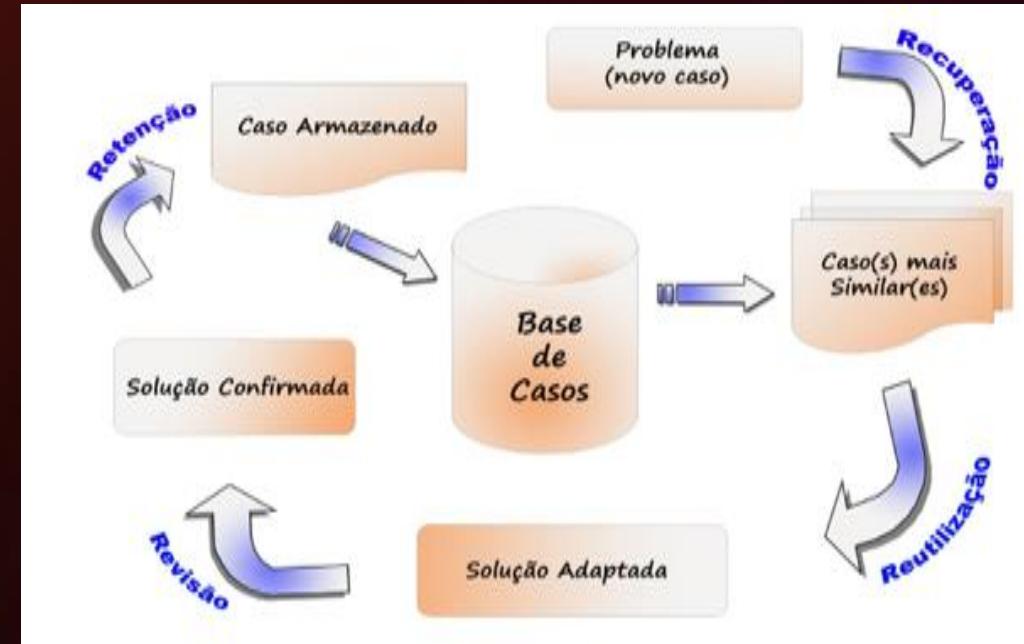
PARADIGMAS DE APRENDIZADO DE MÁQUINAS

Baseado em exemplos

Uma forma de classificar um novo padrão é lembrar-se de exemplos parecidos classificados anteriormente, e assim atribuir ao novo exemplo uma classe de um padrão parecido.

Esta é a ideia central deste paradigma.

A técnica de raciocínio baseada em casos é um exemplo de técnica que utiliza este paradigma.



Fonte: Wangenheim e Wangenheim (2003)

PARADIGMAS DE APRENDIZADO DE MÁQUINAS

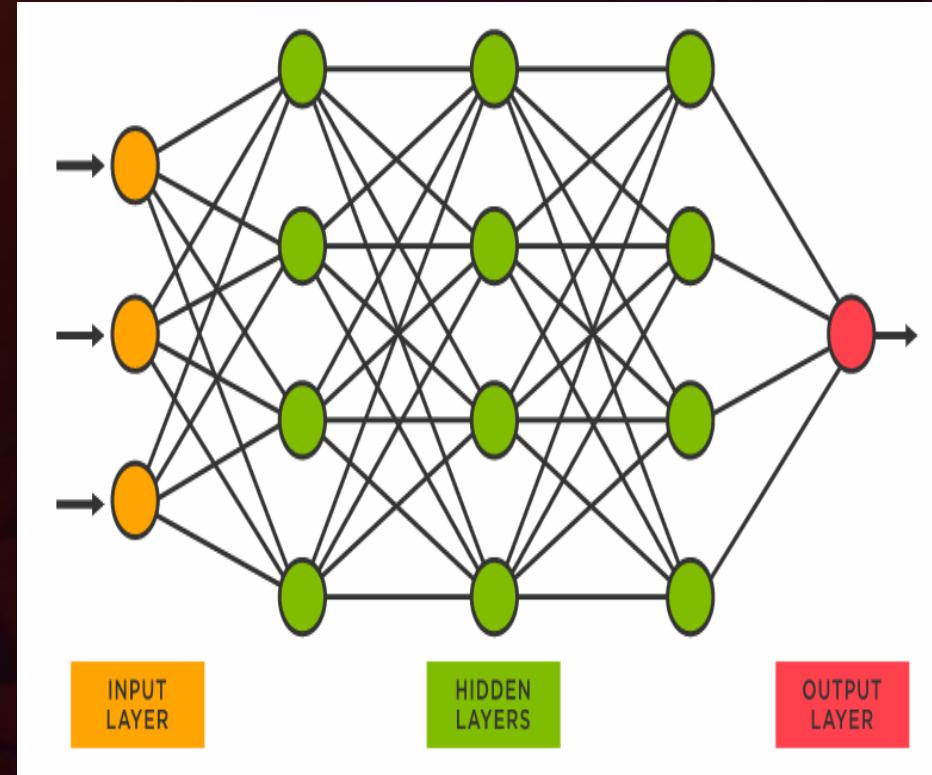
Conexionista

O nome **conexionista** vem da área de pesquisa de **Redes Neurais Artificiais (RNA)**. Uma rede neural artificial é um **modelo computacional** inspirado no funcionamento do cérebro humano.

Uma RNA possui três componentes principais: unidade de processamento “os neurônios”, conexões “sinapses” e uma topologia.

As redes neurais possuem como principal característica aprender através de exemplos e poder de generalização.

As redes Multi Layer Perceptron (MLP) e Self Organizing Map (SOM) são exemplos de técnicas que utilizam este paradigma de aprendizado.



PARADIGMAS DE APRENDIZADO DE MÁQUINAS

Evolutivo

Este paradigma foi inspirado na teoria da evolução das espécies de Charles Darwin.

O algoritmo inicia com uma população de indivíduos, onde cada indivíduo representa uma possível solução.

Os indivíduos competem entre si, os indivíduos com menor desempenho são descartados, e os indivíduos com melhores desempenhos são selecionados para reprodução (Crossover); os novos indivíduos gerados podem ou não sofrer mutação.

A população evolui através de várias gerações, até que uma solução ótima seja encontrada.

Algoritmos genéticos e Programação genética são exemplos de técnicas que utilizam este paradigma.



MÉTODOS BAYESIANOS

- Fornece algoritmos práticos de aprendizagem
- Aprendizagem de Redes Bayesianas
- Combina conhecimento *a priori* (probabilidade *a priori* ou incondicional) com dados de observação.
- Assumem que a probabilidade de um evento A (que pode ser uma classe), dado um evento B (que pode ser o conjunto dos atributos de entrada) não depende apenas da relação entre A e B mas também da probabilidade de observar A, independentemente de observar B. (Mitchell, 1997)

TEOREMA DE BAYES

Axiomas de Kolmogorov:

- $P(E) \geq 0$;
- Se Ω é o espaço de eventos, então $P(\Omega) = 1$;
- Se A e B são eventos disjuntos, então $P(A \cup B) = P(A) + P(B)$.

Probabilidade Condisional:

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

Teorema de Bayes:

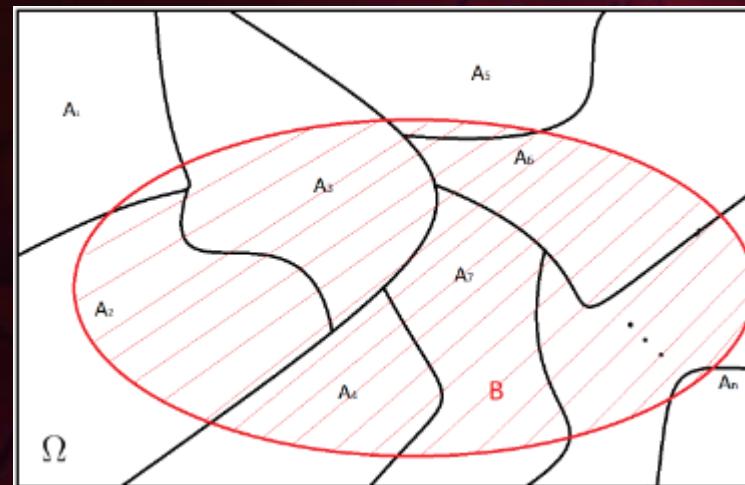
$$P(A \cap B) = P(B \cap A)$$

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

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

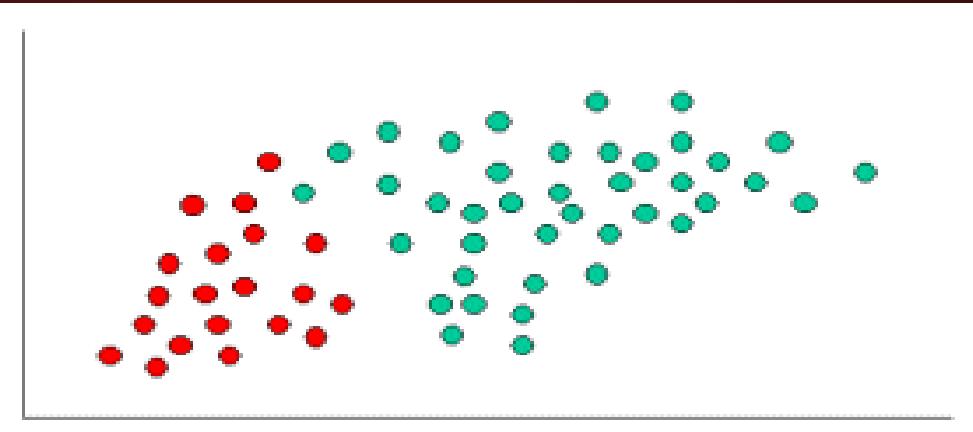
Lei da Probabilidade Total:

$$P(A) = \sum_{i=1}^n P(A | B_i) \times P(B_i)$$



TEOREMA DE BAYES

Considere o exemplo de dados:



Os objetos podem ser classificados em vermelho ou verde

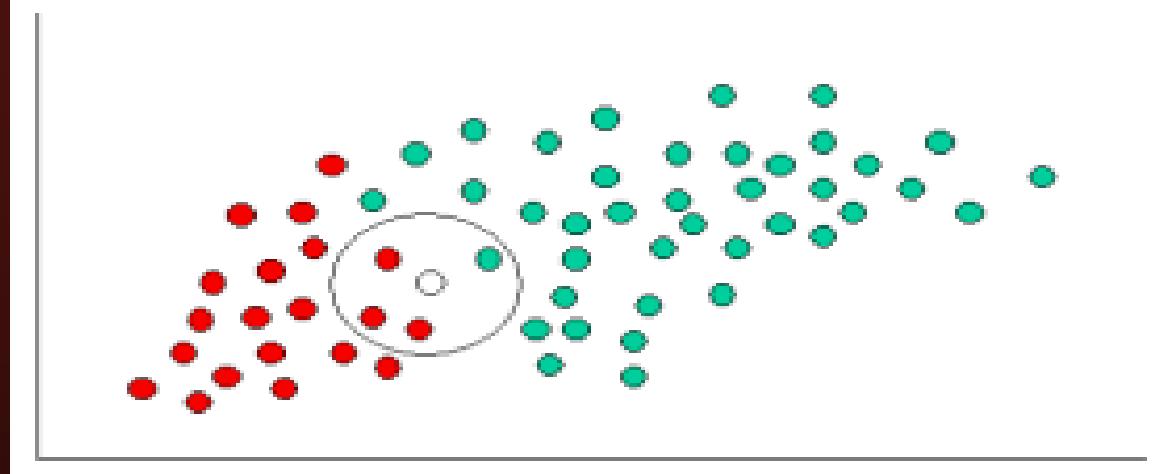
Como há mais objetos verdes que vermelhos, a probabilidade, *a priori*, é que um novo objeto seja verde

Probabilidade *a priori* de verde = número de objetos verdes/ número total de objetos = $40/60 = 4/6$

Probabilidade *a priori* de vermelho = número de objetos vermelhos / número total de objetos = $20/60 = 2/6$

TEOREMA DE BAYES

Queremos classificar um novo objeto X (ponto branco)



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

$$P(\text{novo objeto ser verde}) = 4/6 \cdot 1/40 = 1/60$$

$$P(\text{novo objeto ser vermelho}) = 2/6 \cdot 3/20 = 1/20$$

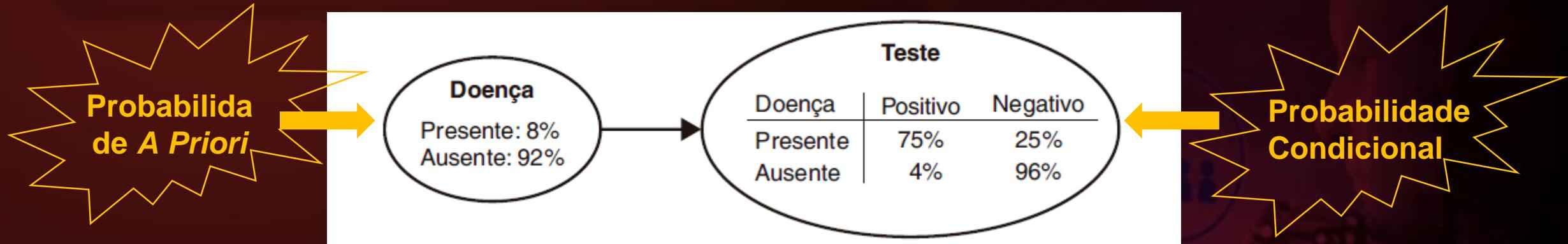
Como os objetos estão agrupados, é razoável considerar que quanto mais objetos de uma classe “parecidos” com X, maior a chance de X ser daquela classe.

Vamos considerar o “parecido” pelo círculo na figura (estar dentro do círculo) e calcular a probabilidade:

Probabilidade de “parecido” dado que é verde = número de objetos verdes no círculo/ número total de verdes= 1/40

Probabilidade de “parecido” dado que é vermelho = número de objetos vermelhos no círculo/ número total de vermelhos= 3/20

APRENDIZADO BAYESIANO



Fonte: Inteligência Artificial – Uma abordagem de aprendizado de máquina (Faceli et al.)

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

$$P(\text{Doença} = \text{presente}) = 0,08 \text{ e}$$
$$P(\text{Doença} = \text{ausente}) = 0,92$$

$$P(\text{Teste} = \text{positivo} | \text{Doença} = \text{presente}) = 0,75$$

$$P(\text{Teste} = \text{negativo} | \text{Doença} = \text{ausente}) = 0,96$$

$$\begin{aligned} P(\text{Teste} = \text{positivo}) &= P(\text{Teste} = \text{positivo} | \\ &\quad \text{Doença} = \text{presente}) \times P(\text{Doença} = \text{presente}) \\ &+ P(\text{Teste} = \text{positivo} | \text{Doença} = \text{ausente}) \times \\ &\quad P(\text{Doença} = \text{ausente}) = 0,75 \times 0,08 + 0,04 \times \\ &\quad 0,92 = 0,0968 \end{aligned}$$

$$\begin{aligned} P(\text{Teste} = \text{negativo}) &= P(\text{Teste} = \text{negativo} | \\ &\quad \text{Doença} = \text{presente}) \times P(\text{Doença} = \text{presente}) \\ &+ P(\text{Teste} = \text{negativo} | \text{Doença} = \text{ausente}) \times \\ &\quad P(\text{Doença} = \text{ausente}) = 0,25 \times 0,08 + 0,96 \times \\ &\quad 0,92 = 0,9032 \end{aligned}$$

APRENDIZADO BAYESIANO - APLICAÇÃO

Novo caso: O Teste do paciente foi positivo. Podemos concluir que o paciente está doente ?



$$P(\text{Doença} = \text{presente} | \text{Teste} = \text{positivo}) ?$$



$P(y_i | x)$ = probabilidade de um exemplo x pertencer a classe y_i



$$\text{Método MAP : } y_{MAP} = \arg \max_i P(y_i | x)$$



Função Discriminante:

$$P(y_i | x) = \frac{P(y_i)P(x|y_i)}{P(x)}$$

Hipótese mais provável:

$$h_{MV} = \arg \max_i P(x | y_i)$$



$$P(\text{Doença} = \text{presente} | \text{Teste} = \text{positivo}) = P(\text{Doença} = \text{presente}) \\ \times P(\text{Teste} = \text{positivo} | \text{Doença} = \text{presente}) = 0,08 \times 0,75 = 0,06$$

$$P(\text{Doença} = \text{ausente} | \text{Teste} = \text{positivo}) = P(\text{Doença} = \text{ausente}) \\ \times P(\text{Teste} = \text{positivo} | \text{Doença} = \text{ausente}) = 0,92 \times 0,04 = 0,0368$$



Paciente Doente

CLASSIFICADOR NAIVE-BAYES

Para um conjunto de atributos de um exemplo, independentes entre si:

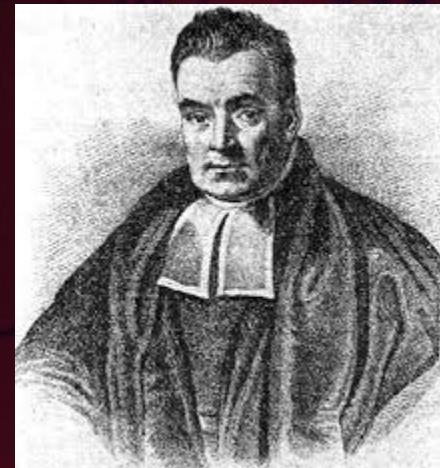
$$P(x | y_i) = P(x^1 | y_i) \times \dots \times P(x^d | y_i) \text{ onde } x^j \text{ é o } j\text{-ésimo atributo do exemplo } x$$



$$P(y_i | x) \propto P(y_i) \prod_{j=1}^d P(x^j | y_i)$$



$$y_{MAP} = \arg \max_i P(y_i | x)$$



**Fórmula do
Naive-Bayes:**

$$\log(P(y_i | x)) \propto \log(P(y_i)) + \sum_j \log(P(x^j | y_i))$$

Classificador
Naive-Bayes

**Caso particular
(2 classes):**

$$\log \frac{P(y_1 | x)}{P(y_2 | x)} \propto \log \frac{P(y_1)}{P(y_2)} + \sum_j \log \frac{P(x^j | y_1)}{P(x^j | y_2)}$$

EXEMPLO: PROBLEMA DA BALANÇA

Cada exemplo é classificado em uma de três posições da balança:

- Balança está inclinada à esquerda
- Balança está inclinada à direita
- Balança equilibrada (sem inclinação)

Atributos:

- Peso do lado esquerdo
- Dimensão do braço esquerdo
- Peso do lado direito
- Dimensão do braço direito

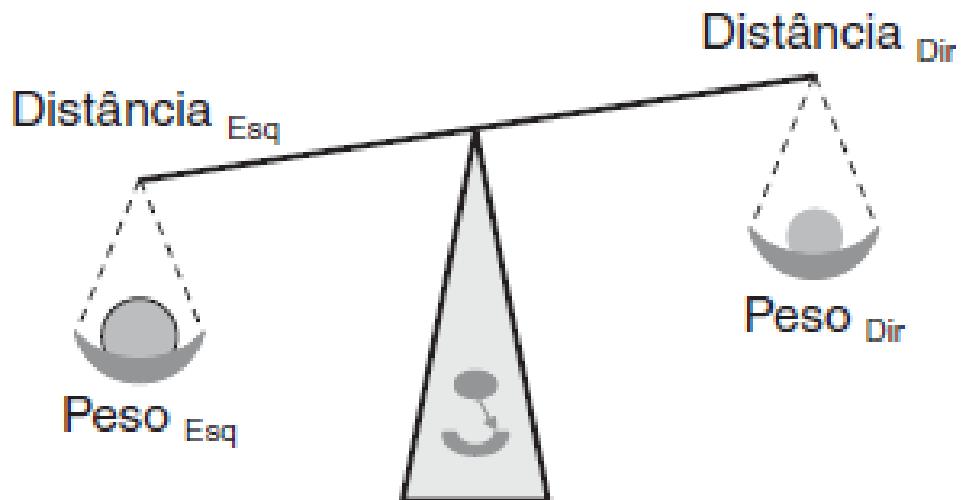
Para encontrar uma Classe:

Procura-se o maior valor entre:

$$\text{Distância}_{\text{esq}} \times \text{Peso}_{\text{esq}} \text{ e } \text{Distância}_{\text{dir}} \times \text{Peso}_{\text{dir}}$$

Domínio dos atributos: {1,2,3,4,5}

Conjunto de dados: 625 exemplos



EXEMPLO: PROBLEMA DA BALANÇA

Distribuição dos valores dos atributos por classe							
Distribuição normal			Discretização				
Peso esq	Média	Desvio Padrão	V1	V2	V3	V4	V5
Balanceada	2,938	1,42	10	11	9	10	9
Esquerda	3,611	1,23	17	43	63	77	88
Direita	2,399	1,33	98	71	53	38	28
Distância esq	Média	Desvio Padrão	V1	V2	V3	V4	V5
Balanceada	2,938	1,42	10	11	9	10	9
Esquerda	3,611	1,22	17	43	63	77	88
Direita	2,399	1,33	98	71	53	38	28
Peso dir	Média	Desvio Padrão	V1	V2	V3	V4	V5
Balanceada	2,938	1,42	10	11	9	10	9
Esquerda	2,399	1,33	98	71	53	38	28
Direita	3,611	1,22	17	43	63	77	88
Distância dir	Média	Desvio Padrão	V1	V2	V3	V4	V5
Balanceada	2,938	1,42	10	11	9	10	9
Esquerda	2,399	1,33	98	71	53	38	28
Direita	3,611	1,22	17	43	63	77	88

Nesse exemplo todos os atributos são numéricos.

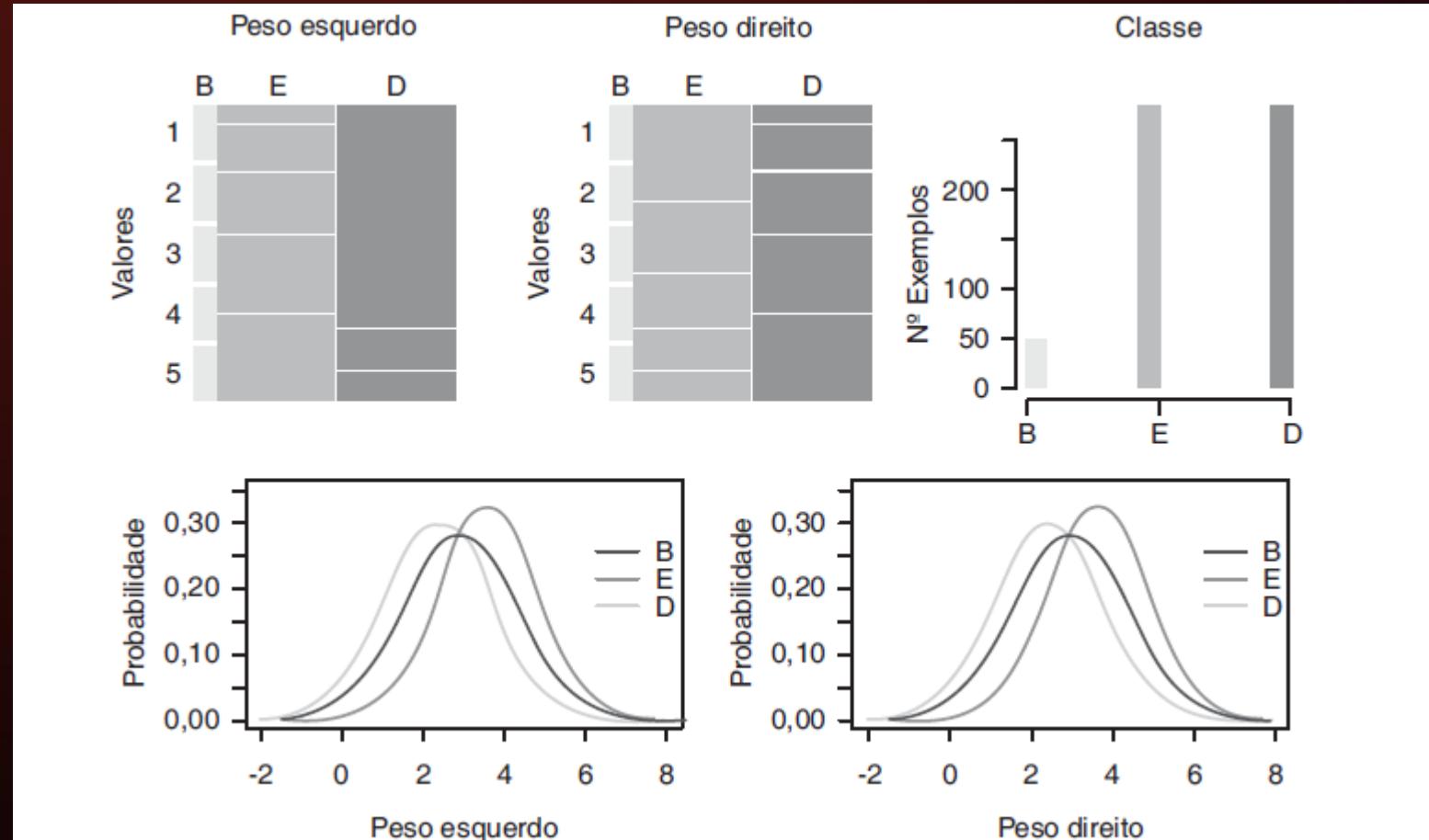
Vamos assumir a hipótese de que eles são normalmente distribuídos.

Calculando a média e o desvio padrão e assumindo $k = \min(10; \text{número de valores diferentes})$ obtemos 5 intervalos.

	Balanceada	Esquerda	Direita
Contagem	49	288	288
P(Classe)	0,078	0,461	0,461

EXEMPLO: PROBLEMA DA BALANÇA

Abaixo as distribuições discretizadas para os atributos: Peso_{esq} , Peso_{dir} e Classe e função densidade de probabilidade por classe assumindo uma distribuição normal:



Fonte: Inteligência Artificial – Uma abordagem de aprendizado de máquina (Faceli et al.)

EXEMPLO: PROBLEMA DA BALANÇA

Supondo um novo exemplo x com os valores $Peso_{esq} = 3$, $Distância_{esq} = 2$, $Peso_{dir} = 1$ e $Distância_{dir} = 3$

Neste caso, a balança está para a esquerda, pois: $3 \times 2 > 1 \times 3$

Utilizando o NB:

$$P(\text{Classe} = \text{esquerda}|x) = 0,71$$

$$P(\text{Classe} = \text{direita}|x) = 0,20$$

$$P(\text{Classe} = \text{balanceada}|x) = 0,09$$

$$\begin{aligned} \log P(\text{Classe} = \text{Esquerda} | x) = \\ \log P(\text{Classe} = \text{Esquerda}) + \log P(Peso_{Esq} = 3 | \text{Classe} = \text{Esquerda}) \\ + \log P(Distância_{Esq} = 2 | \text{Classe} = \text{Esquerda}) + \log P(Peso_{Dir} = 1 | \text{Classe} = \text{Esquerda}) \\ + \log P(Distância_{Dir} = 3 | \text{Classe} = \text{Esquerda}) = \\ \log(0,461) + \log\left(\frac{63}{288}\right) + \log\left(\frac{43}{288}\right) + \log\left(\frac{98}{288}\right) + \log\left(\frac{53}{288}\right) \\ = -3,03 \end{aligned}$$

NAIVE-BAYES: ANÁLISE DO ALGORITMO

Aspectos positivos:

- Todas as probabilidades da equação podem ser calculadas a partir de um conjunto de treinamento em uma única passagem;
- De fácil implementação
- Algoritmo Naive-Bayes tem bom desempenho em uma grande variedade de domínios;
- É robusto à presença de ruídos e atributos irrelevantes;
- Teorias aprendidas são de fácil compreensão pelos especialistas do domínio;

Aspectos negativos:

- O impacto das variáveis irrelevantes deve ser levado em conta em consideração ao desempenho do NB;
- O tratamento de atributos com valores contínuos não é direto (sendo necessário discretizar-los);

APRENDIZADO DE MÁQUINAS

Aprendizado probabilístico