Modelos Preditivos

Departamento de Engenharia Informática (DEI/ISEP)
Fátima Rodrigues
mfc@isep.ipp.pt

Raciocínio

Capacidade humana em trabalhar com conhecimento, factos e estratégias de resolução de problemas por forma a obter conclusões

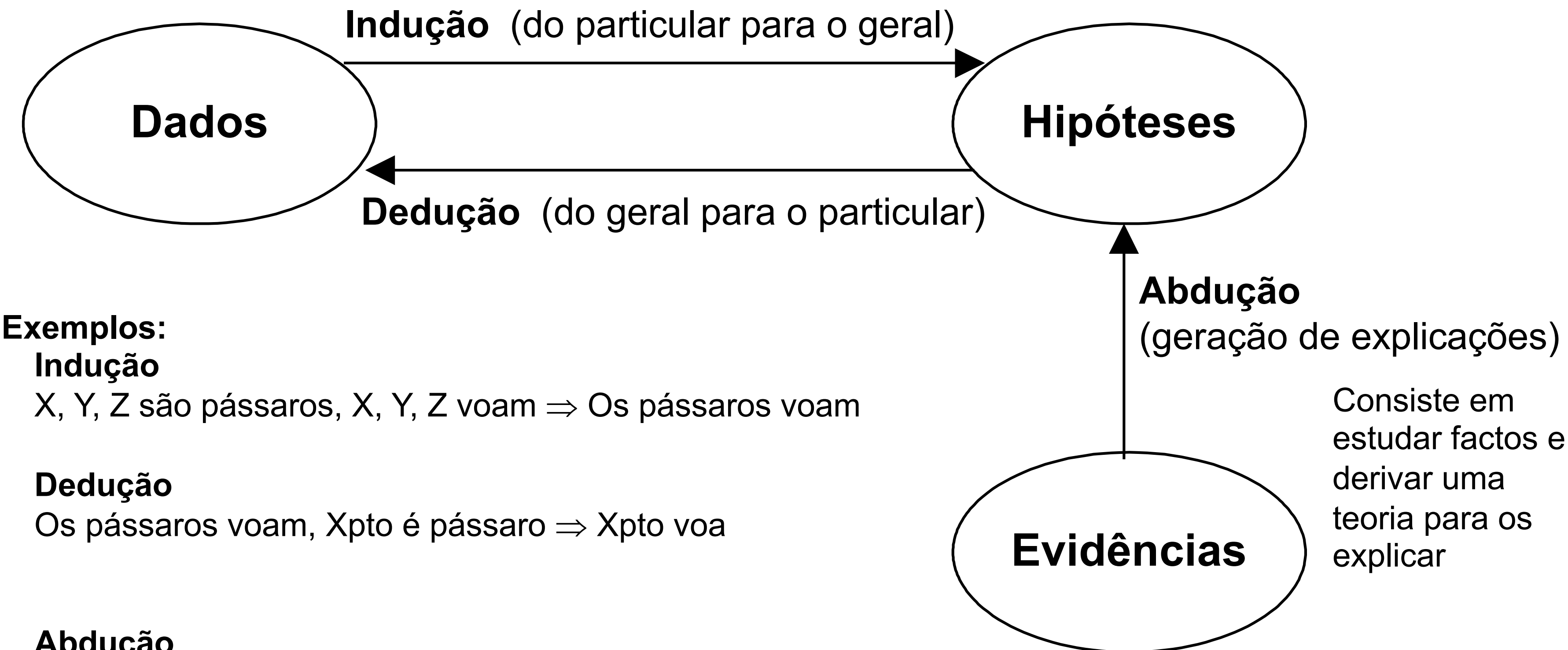
Entender

- como os humanos raciocinam
- como trabalham com a informação relativa a um dado problema

Permite

- delinear o processo de inferência num algoritmo de Data Mining

Mecanismos de Raciocínio



Exemplos:

Indução
X, Y, Z são pássaros, X, Y, Z voam ⇒ Os pássaros voam

Dedução
Os pássaros voam, Xpto é pássaro ⇒ Xpto voa

Abdução
Todos os pássaros voam. Xpto voa.
Hipótese: Xpto é um pássaro

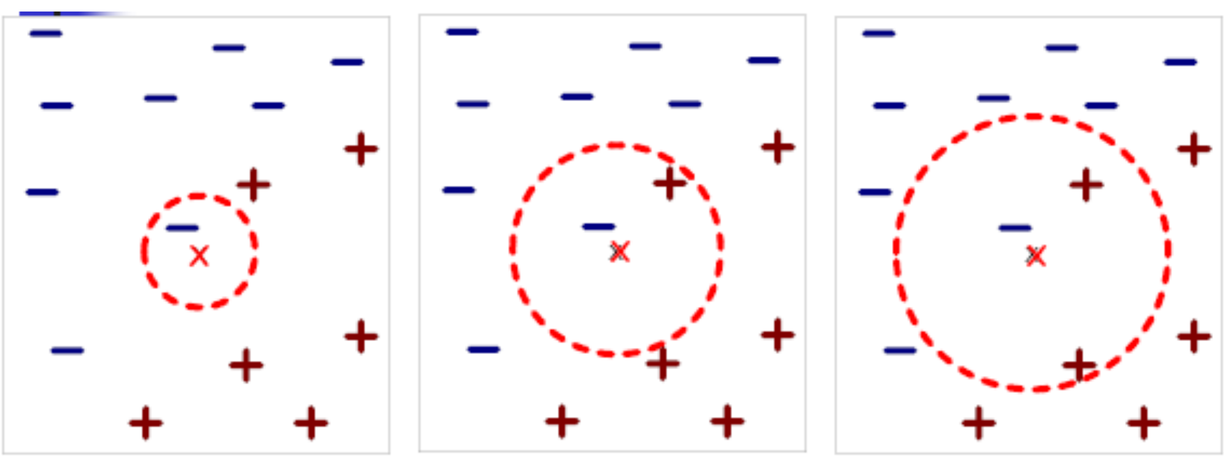
Aprendizagem Baseada em Instâncias

Aprendizagem Baseada em Instâncias

- Método dos k-vizinhos mais próximos
- Raciocínio Baseado em Memória

Método mais antigo (1967) e difundido

Prevê valores desconhecidos de registos num conjunto de dados baseando-se na combinação dos valores dos **k-registos mais próximos**, **mais similares** do conjunto de dados histórico



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Aprendizagem baseada em instâncias

A previsão de um valor (discreto/contínuo) é feita com base nos valores mais próximos

É baseada em dois conceitos fundamentais

- **Função de distância**

  mede a distância entre dois registos

- **Função de combinação**

  combina os resultados a partir dos vizinhos

Distância entre exemplos

Se o conjunto de treino tem N atributos o novo exemplo a classificar x_q tem também N atributos $x_q \in \Re$, *t=1,...N*

A distância entre dois pontos x_i e y_i é normalmente calculada usando a **distância euclidiana**

$$dist(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Os atributos devem ser **normalizados**

Propriedades da Função de Distância

Para que uma função *possa* ser usada como função *distância* é necessário e suficiente que as seguintes condições para quaisquer objetos *d, p, q* sejam satisfeitas:

1. $d(p, q) \geq 0$
2. $d(p, p) = 0$
3. $d(p, q) = d(q, p)$ *(Simetria)*
4. $d(p, r) \leq d(p, q) + d(q, r)$ (desigualdade triangular)

As mais importantes funções nesta categoria são:

- Distância de Minkowski

$$dist(i, j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{in} - x_{jn}|^q}$$

- Distância Euclidiana (q=2)
- Distância de Manhattan (q=1)

Algoritmo K-Vizinhos mais Próximos

Tendo N exemplos de treino $(x_n, f(x_n))$ $n = 1,2,..,N$

Para um novo registo x_q

encontra os K-vizinhos mais próximos em $(x_n, f(x_n))$

Se $f(x_n)$ é discreto

$f(x_q)$ é a moda $f(x_k)$ (k-vizinhos mais próximos)

Se $f(x_n)$ é contínuo

$f(x_q)$ é a média (k-vizinhos mais próximos)

$$f(x_q) := \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

Problema da Dimensionalidade

Para calcular a distância entre os pontos, o método utiliza ***todos*** os atributos da instância

Problema

Quando alguns destes atributos não são igualmente importantes?

Ao atribuir-se a mesma importância aos atributos, ou seja, se os atributos não forem distinguidos está-se a distorcer a noção de semelhança entre dois casos

$$dist(i,j) = \sqrt{w_i \left|x_{i1} - x_{j1}\right|^2 + w_2 \left|x_{i2} - x_{j2}\right|^2 + \ldots + w_n \left|x_{in} - x_{jn}\right|^2}$$

$$\sum_{n} w_i = 1$$

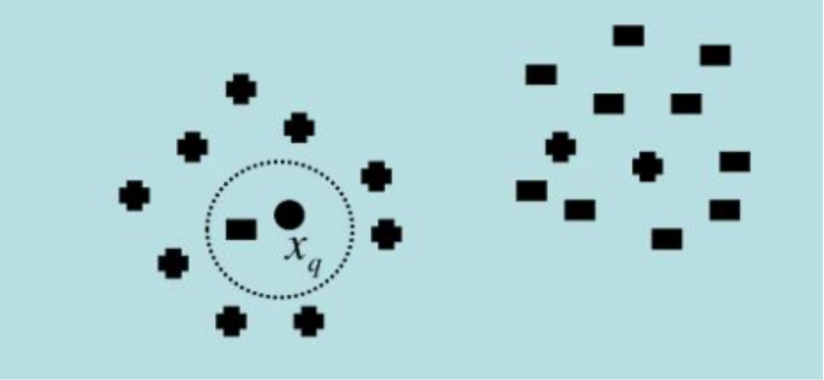
A determinação dos **pesos dos atributos** depende do objetivo da aprendizagem

Escolha K

Valores frequentes de K: 1, 3, 5, ..., valores ímpares

Menores valores de K:

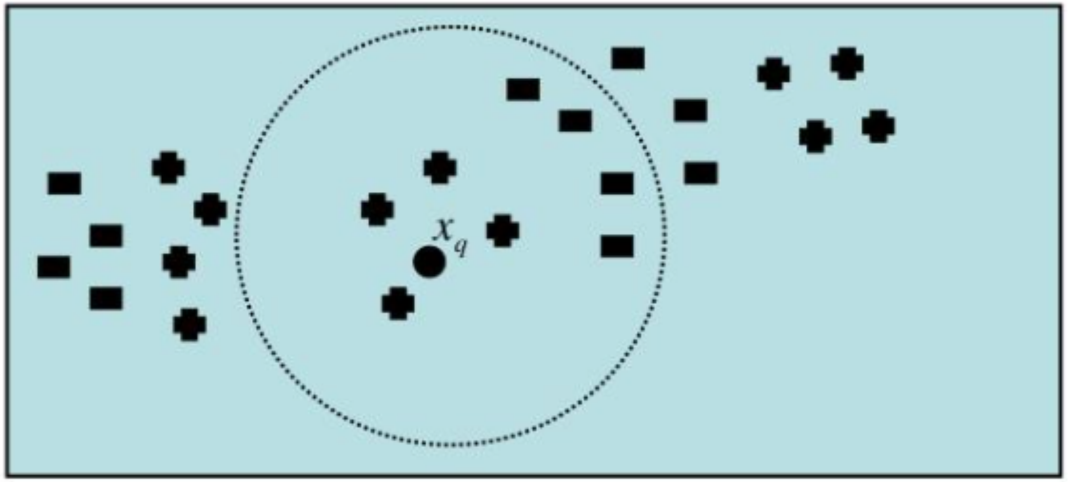
- pode aumentar a contribuição de exemplos *ruidosos*



Escolha K

Valores elevados de K:

- pode aumentar a contribuição de exemplos pouco similares e logo, menos relevantes



- K deve ser estimado experimentalmente

Vantagens / Desvantagens

- ✓ Processa qualquer número de variáveis dependentes
- ✓ Pode ser usada para prever quer valores contínuos quer valores discretos
- ✓ Qualquer função de distância pode ser usada
- ✓ Fácil de implementar
- ✓ Facilmente atualizado por expansão ou substituição do conjunto de treino
- ✓ É fácil entender os resultados

- ✗ Computacionalmente intensivo o calculo das distâncias entre casos
- ✗ Os resultados dependem da função distância usada, da função combinação e do número de vizinhos usados
- ✗ Necessita de grandes conjuntos de treino
- ✗ Sensibilidade a valores isolados e variáveis irrelevantes
- ✗ Ausência de qualquer modelo para “mostrar” ao utilizador

Aprendizagem Baseada em Probabilidades

Teoria Bayesiana - Motivação

Aprendizagem Bayesiana

Usa dados acerca de eventos passados para estimar a probabilidade de eventos futuros

↪ requer **Probabilidades à priori**

A Teoria Bayesiana faz duas assunções:

- *Os atributos são igualmente importantes*
- *Os atributos são estatisticamente independentes* (sabendo a sua classe)
  - i.e., saber o valor de um atributo não diz nada acerca do valor de outro atributo (se a classe for conhecida)

- Assunções de independência normalmente não são corretas!
- No entanto … no esquema de aprendizagem Bayesiana funciona bem

Teoria de Probabilidades

Teoria da Probabilidade

- É uma aproximação matemática para processar informação incerta
- Propõe a existência de um valor **P(E) – Probabilidade -** que consiste na possibilidade de ocorrência de um **evento E** a partir de uma sequência de **experiências de eventos aleatórios**
- Se uma determinada experiência for realizada um número considerável de vezes, então podemos ter quase a certeza que a frequência relativa do evento **E** é aproximadamente igual a **P(E)**
- O conjunto de todos os possíveis resultados de uma experiência (eventos) é denominado **espaço da amostra S**
- A probabilidade total de todos os possíveis resultados de um ensaio é sempre **100%**

Probabilidade Discreta

Experiências com resultados discretos

$$P(E) = \frac{W(E)}{n}$$

A probabilidade de um evento discreto é dada pelo nº de vezes que esse evento correu dividido pelo nº total de experiências

Exemplo

Se 10 mensagens em 50 forem spam ⇒ P(spam) = 10/50 = 20%

Para um espaço amostral com dois possíveis eventos: spam/não-spam
P(não-spam) = 1 – P(spam) = 1-0.20 = 80%

Isso funciona porque spam e não-spam são **mutuamente exclusivos** e **exaustivos**: os eventos não podem ocorrer ao mesmo tempo e são os únicos resultados possíveis no espaço amostral

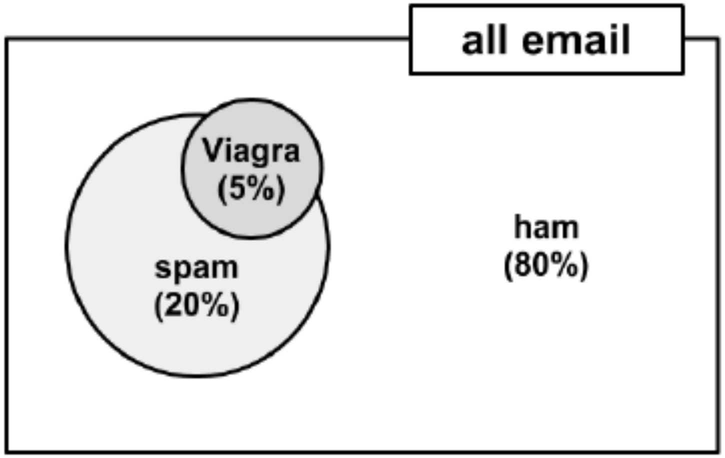
Probabilidade Conjunta

- Muitas vezes, estamos interessados em monitorar vários eventos não mutuamente exclusivos no mesmo espaço amostral

- Se alguns eventos ocorrem simultaneamente com o evento de interesse, é útil usá-los para fazer previsões

Exemplo

Considere, por exemplo, um segundo evento com base no resultado que a mensagem de e-mail contém a palavra Viagra



Probabilidades Conjunta

- Pretende-se estimar a probabilidade de Spam e Viagra ocorrerem conjuntamente, ou seja P(Spam $\cap$ Viagra)

- Se Spam e Viagra fossem dois **eventos independentes**

$$P(Spam \cap Viagra) = P(Spam) \times P(Viagra)$$

P(Spam) = 20%
P(Viagra) = 5% $\Rightarrow$ P(Spam $\cap$ Viagra) = 0.20 × 0.05 = 0.01 (1%)

Mas como a P(Spam) e P(Viagra) são altamente dependentes, significa que **o cálculo não é correto**

Probab. Condicional – Teorema de Bayes

As relações entre os eventos dependentes podem ser descritas por meio do **Teorema de Bayes**, como mostrado na seguinte fórmula:

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

A notação P(A|B) pode ser lido como a probabilidade do evento A, uma vez que o evento B ocorreu

Isto é conhecido como **probabilidade condicional**, uma vez que a probabilidade de A é dependente (isto é, é condicionada) pelo evento B

Probabilidade Condicional

- Se uma mensagem for selecionada aleatoriamente da caixa de correio e essa mensagem contiver a palavra *Viagra* (evento B), qual é a probabilidade da mensagem ser Spam (evento A)?

Aplicando o teorema de Bayes calcula-se a **probabilidade posterior** que mede o quão provável a mensagem poderá ser Spam

Probab. marginal

Probab. à priori

$$P(Spam \mid Viagra) = \frac{P(Viagra \mid Spam)P(Spam)}{P(Viagra)}$$

Probab. à posteriori

Teorema de Bayes

Supondo que temos 100 mensagens (20 spam/80 ham) na nossa caixa de correio:

- 5 mensagens contêm a palavra *Viagra*, e se destas 4 forem Spam

  P(Viagra|Spam) = 4/20 = 0.20 -> probabilidade marginal

$$P(Spam \mid Viagra) = \frac{P(Viagra \mid Spam)P(Spam)}{P(Viagra)}$$

$$P(Spam \mid Viagra) = \frac{0.2 \times 0.2}{0.05} = 0.8$$

Portanto, **80%** é a probabilidade de uma mensagem ser spam, uma vez que contém a palavra *Viagra*

Teorema de Bayes

Se adicionarmos mais termos ao nosso filtro de spam:

	Viagra		Dinheiro		Fitness		Unsubscribe		
Frequência	Sim	Não	Sim	Não	Sim	Não	Sim	Não	
Spam	4	16	10	10	0	20	12	8	20
Não Spam	1	79	14	66	8	72	23	57	80
Total									100

Qual a probabilidade de uma mensagem ser spam se Viagra = Sim, Dinheiro = Não, Fitness = Não, e Unsubscribe = Sim:

$$P(Spam \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 \mid Spam)P(Spam)}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

Fórmula computacionalmente muito difícil de resolver !

Classificador Bayesiano Naive (Ingénuo)

Suposição Naive Bayes

$$P(a_i, a_2, .., a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

Ou seja, as variáveis a_1, a_2, ..., a_n, são independentes

Seja o valor v_j do atributo a prever de uma instância da amostra com n atributos previsores com valores a_1, a_2, ..., a_n **a probabilidade do valor v_j** do atributo a prever é igual **ao produto das probabilidades individuais de cada atributo**

Fazendo a substituição no Classificador Bayesiano Naive

$$V_{MPP} = \arg\max_{v_j \in V} P(v_j) \times \prod_i P(a_i \mid v_j)$$

v_{MPP} é o valor previsto pelo classificador Bayesiano

Suposição Naive Bayes

$$P(Spam \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 \mid Spam)P(Spam)}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

Naive Bayes assume independência condicional de classe, o que significa que os **eventos são independentes** desde que **condicionados ao mesmo valor de classe**

Assumindo independência condicional a fórmula pode ser simplificada:

$$P(Spam \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \mid spam)P(\neg W_2 \mid spam)P(\neg W_3 \mid spam)P(W_4 \mid Spam)P(Spam)}{P(W_1)P(\neg W_2)P(\neg W_3)P(W_4)}$$

O resultado desta fórmula deve ser comparado com a probabilidade da mensagem ser não-spam

Suposição Naive Bayes

	Viagra		Dinheiro		Fitness		Unsubscribe		
Frequência	Sim	Não	Sim	Não	Sim	Não	Sim	Não	
Spam	4	16	10	10	0	20	12	8	20
Não Spam	1	79	14	66	8	72	23	57	80
Total									100

$$P(Spam \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{4}{20} \times \frac{10}{20} \times \frac{20}{20} \times \frac{12}{20} \times \frac{20}{100} = 0.012$$

$$P(\text{não Spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{1}{80} \times \frac{66}{80} \times \frac{72}{80} \times \frac{23}{80} \times \frac{80}{100} = 0.002$$

Suposição Naive Bayes

$$P(Spam \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{4}{20} \times \frac{10}{20} \times \frac{20}{20} \times \frac{12}{20} \times \frac{20}{100} = 0.012$$

$$P(\text{não Spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{1}{80} \times \frac{66}{80} \times \frac{72}{80} \times \frac{23}{80} \times \frac{80}{100} = 0.002$$

Convertendo estes números em probabilidades:

$$P(spam) = \frac{0.012}{(0.012 + 0.002)} = 0.857 \qquad P(não\ spam) = 1 - 0.857 = 0.143$$

Segundo o padrão de palavras da mensagem:

esta mensagem é spam com **85,7%** de probabilidade, e não-spam com **14,3%** de probabilidade

Classificador Bayesiano: Problemas

Se a probabilidade condicional de um dos atributos for nula, a probabilidade da classe será também nula

Por outro lado se os exemplos de treino não cobrirem todos os valores possíveis dos atributos, não será possível classificar determinados registos

Exemplo:

Considere-se outra mensagem com os termos: Viagra, Dinheiro, Fitness e Unsubscribe:

$$P(Spam \mid W_1 \cap W_2 \cap W_3 \cap W_4) = \frac{4}{20} \times \frac{10}{20} \times \frac{0}{20} \times \frac{12}{20} \times \frac{20}{100} = 0$$

$$P(\text{não spam} \mid W_1 \cap W_2 \cap W_3 \cap W_4) = \frac{1}{80} \times \frac{14}{80} \times \frac{8}{80} \times \frac{23}{80} \times \frac{80}{100} = 0.00005$$

$$P(spam) = \frac{0}{(0 + 0.0099)} = 0$$

Contudo a mensagem contém muitos termos relacionados com spam!

Estimativa Laplace

Este problema pode ser tratado usando a **estimativa Laplace** para cálculo das probabilidades condicionais

O estimador de Laplace essencialmente adiciona uma pequena quantidade a cada uma das contagens na tabela de frequência, o que assegura que cada característica tem uma probabilidade não nula

Tipicamente, o estimador de Laplace é ajustado para 1, o que garante que cada combinação de características da classe é encontrado nos dados pelo menos uma vez

Estimativa Laplace

Tabela de frequência inicial

	Viagra		Dinheiro		Fitness		Unsubscribe		
Frequência	Sim	Não	Sim	Não	Sim	Não	Sim	Não	
Spam	4	16	10	10	0	20	12	8	20
Não Spam	1	79	14	66	8	72	23	57	80
Total									100

Tabela de frequência com a estimativa Laplace

	Viagra		Dinheiro		Fitness		Unsubscribe		
Frequência	Sim	Não	Sim	Não	Sim	Não	Sim	Não	
Spam	5	17	11	11	1	21	13	9	22
Não Spam	2	80	15	67	9	73	24	58	82
Total									104

Estimativa Laplace

	Viagra		Dinheiro		Fitness		Unsubscribe		
Frequência	Sim	Não	Sim	Não	Sim	Não	Sim	Não	
Spam	5	17	11	11	1	21	13	9	22
Não Spam	2	80	15	67	9	73	24	58	82
Total									104

$$P(Spam \mid W_1 \cap W_2 \cap W_3 \cap W_4) = \frac{5}{22} \times \frac{11}{22} \times \frac{1}{22} \times \frac{13}{22} \times \frac{22}{104} = 0.00065$$

$$P(não\ Spam \mid W_1 \cap W_2 \cap W_3 \cap W_4) = \frac{2}{82} \times \frac{15}{82} \times \frac{9}{82} \times \frac{24}{82} \times \frac{82}{104} = 0.00011$$

O estimador de Laplace pode ser ajustado para qualquer valor, o valor 1 é o mais usado

Teorema de Bayes

A aplicação do teorema de Bayes como classificador requer que se conheçam:

- duas probabilidades a priori - P(decisão$_i$)
- uma probabilidade condicional - P(x|decisão$_i$)

Este classificador é ótimo no sentido em que, em média, nenhum outro classificador pode obter melhores resultados usando a mesma informação

Na prática estas probabilidades são desconhecidas

- **Estimativas fiáveis** destas probabilidades a partir de um conjunto de exemplos **requerem um número infinito de exemplos**

Assumem-se simplificações no calculo de P(x|decisão)

Assume-se que os atributos são independentes da decisão

↪**Classificador Bayesiano Naive**

Classificador Bayesiano Naive: Exemplo

Tempo	Temperatura	Humidade	Vento	Joga ?
Sol	elev	elev	N	N
Sol	elev	elev	S	N
Nublado	elev	elev	N	S
Chuva	med	elev	N	S
Chuva	baixa	normal	N	S
Chuva	baixa	normal	S	N
Nublado	baixa	normal	S	S
Sol	med	elev	N	N
Sol	baixa	normal	N	S
Chuva	med	normal	N	S
Sol	med	normal	S	S
Nublado	med	elev	S	S
Nublado	elev	normal	N	S
Chuva	med	elev	S	N

→ **Sol** **baixa** **elev** **S** **?**

Segundo o Classificador Bayesiano Naive é necessário calcular

$$v_{NB} = \underset{v_j \in V}{argmax} \; P'(v_j) \prod_{a_i \in x} P'(a_i \mid v_j)$$

Classificador Bayesiano Naive: Exemplo

Tabelas de Frequência

Tempo	Não	Sim
Sol	3	2
Nublado	0	4
Chuva	2	3

Temper.	Não	Sim
elev	2	2
med	2	4
baixa	1	3

Humidade	Não	Sim
elev	4	3
normal	1	6

Vento	Não	Sim
N	2	6
S	3	3

Tabelas de Probabilidades

Tempo	Não	Sim
Sol	3/5	2/9
Nublado	0/5	4/9
Chuva	2/5	3/9

Temp.	Não	Sim
elev	2/5	2/9
med	2/5	4/9
baixa	1/5	3/9

Humidade	Não	Sim
elev	4/5	3/9
normal	1/5	6/9

Vento	Não	Sim
N	2/5	6/9
S	3/5	3/9

$$P(\text{jogar=Sim}|x) = P(\text{jogar=Sim}) \times P(\text{Sol}|\text{Sim}) \times P(\text{baixa}|\text{Sim}) \times P(\text{elev}|\text{Sim}) \times P(\text{S}|\text{Sim})$$

$$= 9/14 \times 2/9 \times 3/9 \times 3/9 \times 3/9 = 0.0053$$

$$P(\text{jogar=Não}|x) = P(\text{jogar=Não}) \times P(\text{Sol}|\text{Não}) \times P(\text{baixa}|\text{Não}) \times P(\text{elev}|\text{Não}) \times P(\text{S}|\text{Não})$$

$$= 5/14 \times 3/5 \times 1/5 \times 4/5 \times 3/5 = \mathbf{0.021}$$

$P(\text{jogar=Não}|x) > P(\text{jogar=Sim}|x)$ $\Rightarrow$ **Sol baixa elev S N**

Probabilidades a partir de Atributos Contínuos

Atributos contínuos

- Discretizar

- Partição baseada em dois valores: (A < v) ou (A > v)

- Estimativa baseada na densidade de Probabilidade :
  - Assume-se que o atributo segue uma distribuição normal ou binomial
  - Usa-se a amostra para estimar os parâmetros da distribuição

$$média: \ \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad desvio\ padrão: \ \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - e para estimar a Probabilidade condicional $P(A_i|c)$

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Atributos Numéricos

Tempo	Temperatura	Humidade	Vento	Joga ?
Sol	85	85	N	N
Sol	80	90	S	N
Nublado	83	86	N	S
Chuva	70	96	N	S
Chuva	68	80	N	S
Chuva	65	70	S	N
Nublado	64	65	S	S
Sol	72	95	N	N
Sol	69	70	N	S
Chuva	75	80	N	S
Sol	75	70	S	S
Nublado	72	90	S	S
Nublado	81	75	N	S
Chuva	71	91	S	N

→ Sol 66 90 S ?

É necessário calcular

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Classificador Bayesiano Naive: Exemplo

Tabelas de Probabilidades

Tempo	Não	Sim
Sol	3/5	2/9
Nublado	0/5	4/9
Chuva	2/5	3/9

Temper	Não	Sim
	μ =74,6	μ =73,0
	σ = 7,893	σ = 6,164

Humidade	Não	Sim
	μ =86,2	μ =79,11
	σ =9,731	σ =10,126

Vento	Não	SIm
N	2/5	6/9
S	3/5	3/9

$$P(Temper = 66 \mid Joga = S) = \frac{1}{\sqrt{2\pi(6.164)^2}} e^{-\frac{(66-73)^2}{2\times 6{,}164^2}} = 0.0590$$

$$P(Temper = 66 \mid Joga = N) = \frac{1}{\sqrt{2\pi(7.893)^2}} e^{-\frac{(66-74{,}6)^2}{2\times 7{,}893^2}} = 0.0472$$

$$P(Humidade = 90 \mid Joga = S) = 0.041$$

$$P(Humidade = 90 \mid Joga = N) = 0.042$$

Classificador Bayesiano Naive: Exemplo

Tabelas de Frequência

Tempo	Não	Sim
Sol	3	2
Nublado	0	4
Chuva	2	3

Temper.	Não	Sim
	65	68
	71	75
	85	....
	80	69
	72	75

Humidade	Não	Sim
	70	70
	91	70
	85	....
	90	90
	95	90

Vento	Não	Sim
N	2	6
S	3	3

Tabelas de Probabilidades

Tempo	Não	Sim
Sol	3/5	2/9
Nublado	0/5	4/9
Chuva	2/5	3/9

Temper	Não	Sim
	μ =74,6,0	μ =73,0
	σ =7,89	σ =6,16

Humidade	Não	Sim
	μ =86,2	μ =79,11
	σ =9,73	σ =10,22

Vento	Não	SIm
N	2/5	6/9
S	3/5	3/9

P(jogar=Sim|x) = P(jogar=Sim)×P(Sol|Sim)×P(66|Sim)×P(90|Sim)×P(S|Sim)

= 9/14 × 2/9 × 0,0590 × 0,041 × 3/9 = 0.00012

P(jogar=Não|x) = P(jogar=Não)×P(Sol|Não)×P(66|Não)×P(90|Não)×P(S|Não)

= 5/14 × 3/5 × 0.0472 × 0.042 × 3/5 = **0.00025**

P(jogar=Não|x) > P(jogar=Sim|x) $\Rightarrow$ **Sol 66 90 S N**

Classificador Naive Bayes - Vantagens

- Robusto no **tratamento de valores isolados**
  - Valores isolados afetam pouco o cálculo das probabilidades
- Robusto no **tratamento de atributos irrelevantes**
  - Atributos afetam pouco as probabilidades relativas entre classes
- Fácil de implementar
- Capaz de classificar amostras com valores ausentes
- Considera todos os atributos como igualmente importantes
- Complexidade computacional linear em todas as variáveis do problema

Classificador Naive Bayes - Desvantagens

- Calculo de um número muito elevado de probabilidades
- Amostras suficientemente representativas (elevada dimensão) para se obterem probabilidades precisas
  - Deve ser usado se disponível um conjunto de treino médio ou grande
- Suposição de independência condicional
- Não capta dependências entre variáveis
- Desempenho pode ser afetado pela presença de atributos correlacionados