SCC0173 – Mineração de Dados Biológicos

Classificação II: Algoritmo Probabilístico Naïve Bayes

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

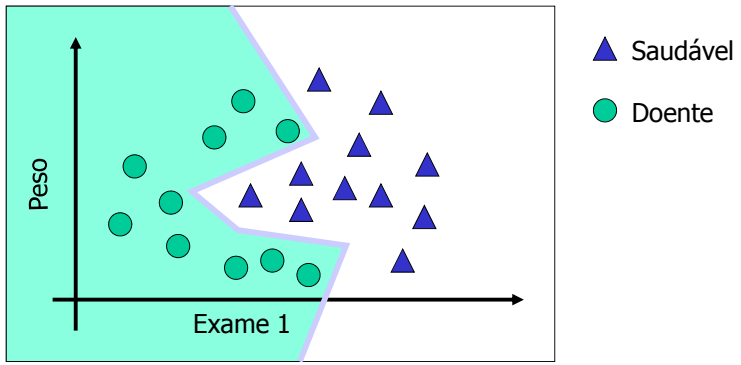
Créditos

- O material a seguir consiste de adaptações e extensões dos originais:
  - gentilmente cedidos pelo Prof. Eduardo R. Hruschka
  - de (Tan et al., 2006)
  - de (Witten & Frank, 2005)

Aula de Hoje

- Revisão de Elementos da Teoria de Probabilidades
  - Probabilidade Condicional
  - Teorema de Bayes
- Classificação via Aprendizado de Máquina Probabilístico
  - Algoritmo Naïve Bayes

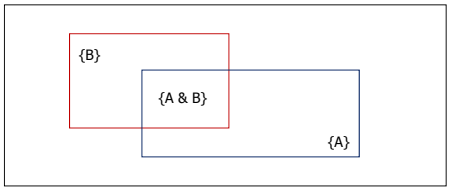
Relembrando Classificação...



Probabilidade Condicional

- Se A e B não forem eventos independentes, tem-se:
  - P(A & B) = P(A)*P(B|A)

  onde P(B|A) = P(A & B) / P(A) é a probabilidade que B ocorra dado que A ocorreu (**probabilidade condicional** de B dado A)
- Exemplo: várias bolas de 2 cores diversas em uma caixa
  - A = bola com 1 cor azul
  - B = bola com 1 cor vermelha
  - A & B = bola azul e vermelha



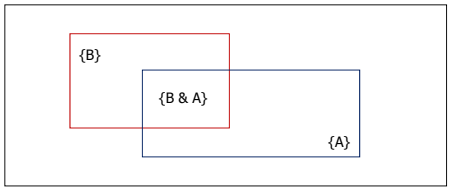
Teorema de Bayes

(Witten & Frank, 2005)

Thomas Bayes (1702-1761)

- Note que P(A & B) = P(B & A) e portanto
  - P(B|A)*P(A) = P(A|B)*P(B)
- **Teorema de Bayes**:
  - P(B|A) = P(A|B)*P(B) / P(A)

- {A} = conj. bolas azuis
- {B} = conj. bolas vermelhas
- {A & B} = conj. bolas azuis e vermelhas



Exemplo do Teorema de Bayes

- Dado:
  - Médico sabe que meningite causa pescoço rígido 50% das vezes
  - Probabilidade **a priori** de qualquer paciente ter meningite é de 1/50000
  - Probabilidade **a priori** de qualquer paciente ter pescoço rígido é 1/20
- Se um paciente tem pescoço rígido (**evidência**), qual é a **probabilidade a posteriori** que ele tenha meningite ?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Teorema de Bayes

- Para várias vars aleatórias A_1, A_2, ..., A_n, e B: raciocínio análogo
- **Teorema de Bayes**:
  - $P(B|A_1,A_2,...,A_n) = P(A_1, A_2,...,A_n|B)*P(B) / P(A_1,A_2,...,A_n)$
- Se as vars aleatórias A_1 ... A_n forem independentes entre si tem-se:
  - $P(A_1,A_2,...,A_n) = P(A_1)*P(A_2)* ... *P(A_n)$ [**independência**]
  - $P(A_1, A_2,...,A_n|B) = P(A_1|B)*P(A_2|B)* ... * P(A_n|B)$ [**independência condicional**]
  - e finalmente...

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

Naïve Bayes

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

- Algoritmo que utiliza o Teorema de Bayes com a hipótese de independência entre atributos
- Apesar da hipótese ser quase sempre violada...
  - o método se mostra bastante competitivo na prática !
    - é amplamente utilizado em aplicações reais
    - está implementado no software **Weka**

Exemplo:

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	2/9	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								

- Idéia é estimar a probabilidade de cada valor B do atributo meta (valor da classe) dados os valores A_1 ... A_n dos demais atributos

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

Base de Dados "Weather" (Witten & 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

Continuando...

(Witten & Frank, 2005)

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
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								

- Para um novo dia:

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

Problema da Freqüência Zero

- 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, ...) = 0
    - Não importa as probabilidades referentes aos demais atributos !
    - Muito radical, especialmente considerando que a base de treinamento pode não ser totalmente representativa
      - Por exemplo, classes minoritárias com instâncias raras

Problema da Freqüência Zero

- Possível solução (**Estimador de Laplace**):
  - Adicionar 1 unidade fictícia para cada combinação de valor-classe
    - Como resultado, probabilidades nunca serão zero !
  - Exemplo (atributo Outlook – classe No):

$$\frac{3+1}{5+3} \qquad \frac{0+1}{5+3} \qquad \frac{2+1}{5+3}$$

Sunny ***Overcast*** ***Rainy***

  - Nota: Deve ser feito para todas as classes, para não inserir viés nas probabilidades de apenas uma classe

Exercício

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

1. Associe valores categóricos ao atributo "Taxable Income" mapeando valores menores que 80K em "Small" e valores maiores que 80K em "Not Small"
2. Calcule a tabela de probabilidades Naive Bayes referente aos dados
3. Use a tabela e classifique cada uma das suas instâncias
4. Calcule o erro de classificação
5. Repita os exercícios acima usando o estimador de Laplace quando aplicável

Valores Ausentes

- Treinamento:
  - excluir exemplo do conjunto de treinamento
- Classificação:
  - considerar apenas os demais atributos
- Exemplo:

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%

Outros Problemas que Também Podem Ser Solucionados...

- Atributos Redundantes
  - seleção de atributos
  - veremos posteriormente no curso...
- Atributos com Valores Numéricos
  - a seguir...

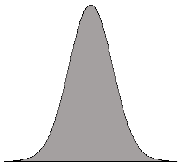
Atributos Numéricos

- **Alternativa 1:** Discretização
- **Alternativa 2:** Assumir ou estimar alguma função de densidade de probabilidade para estimar as probabilidades
  - Usualmente distribuição Gaussiana (Normal)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu)^2$$

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



Karl Gauss
1777-1855

(Witten & Frank, 2005)

Estatísticas para a BD "Weather"

(Witten & Frank, 2005)

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									

- Densidade como estimativa proporcional de probabilidade:
  - Exemplo: temperature = 66 | yes

$$\frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2\times 6.2^2}} = 0.0340$$

Exercício

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(Taxable\ Income = 120 \mid No) = ????$$

$$P(Taxable\ Income = 120 \mid Yes) = ????$$

Naive Bayes: Características

- **Complexidade computacional linear**
  - em todas as variáveis do problema !
- Robusto a ruídos isolados
  - Afetam pouco o cálculo das probabilidades
- Robusto a atributos irrelevantes
  - Afetam pouco as probabilidades relativas entre classes
- Assume que atributos são igualmente importantes
- Desempenho pode ser (mas muitas vezes não é) afetado pela presença de atributos correlacionados

Exercício

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

1. Calcule a tabela de probabilidades Naive Bayes referente aos dados ao lado
2. Use a tabela e classifique cada uma das suas instâncias
3. Calcule o erro de classificação
4. Repita o item 1 deixando de fora 5 instâncias escolhidas aleatoriamente (para teste)
5. Use a tabela para classificar as 5 instâncias de teste e calcule o erro de classificação

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	???