Naïve Bayesian Learning

Marcílo Souto
DIMAp/UFRN

Teorema de Bayes

- Probabilidade de um evento *H* dada evidência *E*:

$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]}$$

- Probabilidade a *priori* de *H*: $\Pr[H]$
  - Probabilidade do evento *antes* da evidência ser vista
- Probabilidade a *posteriori* *H*: $\Pr[H \mid E]$
  - Probabilidade do evento *ápos* a evidência é vista

Aplicação do Teorema de Bayes: Diagnóstico Médico

•Seja

M=doença meningite

S= dor no pescoço

•Um Médico sabe:

$Pr(S/M)=0.5$

$Pr(M)=1/50000$

$Pr(S)=1/20$

$$Pr(M/S)=\frac{P(S/M)P(M)}{Pr(S)}$$

$$=\frac{0{,}5*(1/50000)}{1/20}=0{,}002$$

•A probabilidade de uma pessoa ter meningite dado que ela está com dor no pescoço é 0,02% ou ainda 1 em 5000.

Naïve Bayes para classificação

- Aprendizado: qual a probabilidade da classe dada uma instância?
  - Evidência E = uma instância (padrão)
  - Evento H = valor da classe para a instância (Play=yes, Play=no)
- Suposição Naïve Bayes: evidência pode ser dividida em partes independentes (i.e., os atributos das instâncias são independentes)

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]\ldots\Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Conjunto de Dados “Tempo”

	Outlook	Temperature	Humidity	Windy	Play
D1	overcast	cool	normal	true	yes
D2	overcast	hot	high	false	yes
D3	overcast	hot	normal	false	yes
D4	overcast	mild	high	true	yes
D5	rainy	cool	normal	false	yes
D6	rainy	mild	high	false	yes
D7	rainy	mild	normal	false	yes
D8	sunny	cool	normal	false	yes
D9	sunny	mild	normal	true	yes
D10	rainy	cool	normal	true	no
D11	rainy	mild	high	true	no
D12	sunny	hot	high	false	no
D13	sunny	hot	high	true	no
D14	sunny	mild	high	false	no

Exemplo para o conjunto de dados de Tempo

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Evidência E

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes] \times$$
$$\Pr[Temperature = Cool \mid yes] \times$$
$$\Pr[Humdity = High \mid yes] \times$$
$$\Pr[Windy = True \mid yes] \times \frac{\Pr[yes]}{\Pr[E]}$$

Probabilidade para classe "yes"

Probabilidades para o Conjunto

Outlook			**Temperature**			**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								

- Um dia novo:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

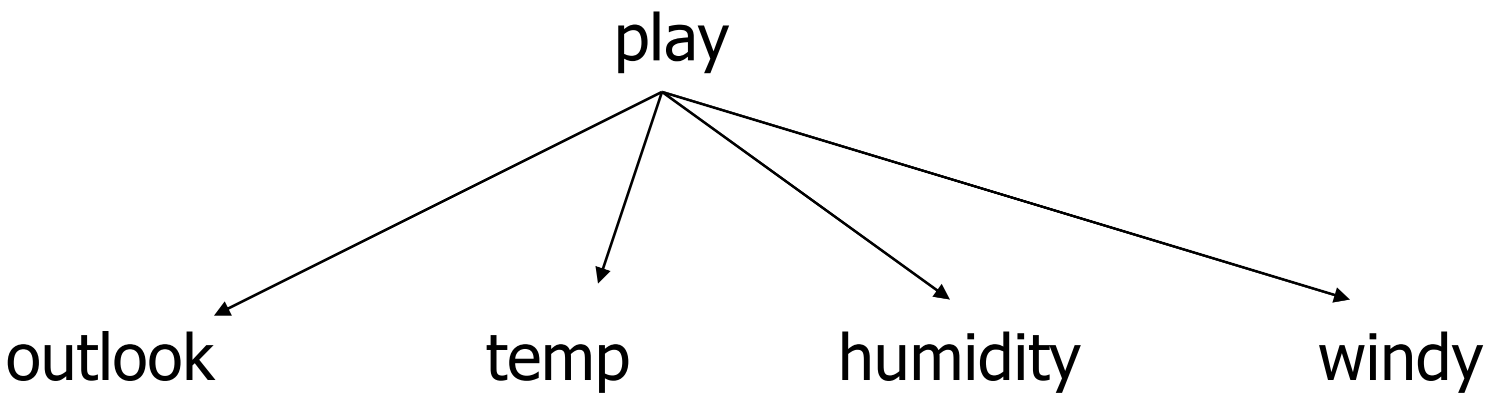
Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Por que Naïve?

- Suponha que todos os atributos são independentes, dada a classe
- O que isso significa?



Pr(outlook=sunny | windy=true, play=yes)=
Pr(outlook=sunny|play=yes)

Conjunto de Dados “Tempo”

Outlook	Windy	Play
overcast	FALSE	yes
rainy	FALSE	yes
rainy	FALSE	yes
overcast	TRUE	yes
sunny	FALSE	yes
rainy	FALSE	yes
sunny	TRUE	yes
overcast	TRUE	yes
overcast	FALSE	yes

A suposição é satisfeita?

- #yes=9
- #sunny=2
- #windy, yes=3
- #sunny|windy, yes=1

Pr(outlook=sunny|windy=true, play=yes)=1/3

Pr(outlook=sunny|play=yes)**=2/9**

Pr(windy|outlook=sunny,play=yes)=1/2

Pr(windy|play=yes)=3/9

Assim, a suposição NÃO é satisfeita.

Outlook	Windy	Play
overcast	FALSE	yes
rainy	FALSE	yes
rainy	FALSE	yes
overcast	TRUE	yes
sunny	FALSE	yes
rainy	FALSE	yes
sunny	TRUE	yes
overcast	TRUE	yes
overcast	FALSE	yes

O problema de “freqüência-zero”

- E se um valor de atributo não ocorre para cada valor da classe (e.g., “Humidity = high” para a classe “yes”)?
  - Probabilidade será zero!
  - Probabilidade a *posteriori* probability também será zero!

    (Não importa os valores dos outros atributos!)
- Solução: adicione 1 ao contador de cada valor de atributo para cada combinação de classe (*Laplace estimator)*
- Resultado: probabilidades nunca serão zero

$$\Pr[Humdity = High \mid yes] = 0$$

$$\Pr[yes \mid E] = 0$$

Estimativas modificadas de probabilidade

- Em alguns casos a adição de uma constante diferente de zero pode ser apropriada
- Exemplo: atributo *outlook* para a classe *yes*

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

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

Valores faltosos

- Treinamento: instância não é incluida na freqüência do contador
- Classificação: atributo será omitido do cálculo
- Exemplo:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of "yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

Likelihood of "no" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

P("yes") = $0.0238 / (0.0238 + 0.0343) = 41\%$

P("no") = $0.0343 / (0.0238 + 0.0343) = 59\%$

Atributos numéricos?

- Suposição usual: os valores dos atributos têm uma *distribuição normal ou Gaussiana* (dada a classe)
- A *função de densidade probabilidade* para a distribuição normal é definida por dois parâmetros:
  - Média da amostra μ:

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

  - Desvio padrão σ:

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

  - Função de densidade *f(x)*:

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

Estatísticas para a base Tempo

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					
				…	…		…	…					
Sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	*std dev*	6.2	7.9	*std dev*	10.2	9.7	True	3/9	3/5		
Rainy	3/9	2/5											

- Exemplo:

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Classificando um novo dia

- Novo dia:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of "yes" = $2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$

Likelihood of "no" = $3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$

P("yes") = 0.000036 / (0.000036 + 0. 000136) = 20.9%

P("no") = 0. 000136 / (0.000036 + 0. 000136) = 79.1%

- Valores faltosos durante treinamento: não são incluidos no cálculo da média e do desvio padrão

Observações

- Junto com árvores de decisão, redes neurais, vizinhos mais-próximos, é um dos métodos de aprendizagem mais práticos
- Quando usa-lo
  - Quando se tem disponível um conjunto de treinamento médio ou grande
  - Os atributos que descrevem as instâncias forem condicionalmente independentes dada uma classificação
  - Note também que nem todos os atributos numéricos vão obedecer uma distribuição normal
- Aplicações:
  - Diagnóstico
  - Classificação de documentos (texto)