Inteligência Artificial

Guilherme Henrique de Souza Nakahata

Universidade Estadual do Paraná - Unespar

20 de Junho de 2024

Naive Bayes

- Algoritmo de classificação probabilístico;
- Baseado no Teorema de Bayes;
- Assume independência entre os preditores;
- Estima a hipótese mais provável.

Teorema de Bayes

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

Exemplo

Suponha que queremos classificar se um e-mail é spam ou não spam.

- C: Classe (spam ou não spam)
- $X_1, X_2, ..., X_n$: Características do e-mail (palavras, por exemplo)

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C) \cdot P(C)}{P(\mathbf{X})}$$

Onde:

- $P(C)$: Probabilidade da classe C
- $P(\mathbf{X}|C)$: Probabilidade dos preditores dado C
- $P(\mathbf{X})$: Probabilidade dos preditores

Exemplo: Continuação

Suposição de independência (Naive)

Naive Bayes assume que as características $X_1, X_2, ..., X_n$ são independentes, dado C:

$$P(\mathbf{X}|C) = P(X_1|C) \cdot P(X_2|C) \cdot ... \cdot P(X_n|C)$$

$$P(a_1, a_2, \ldots, a_n|v_j) = \prod_{a_i}^{n} P(a_i|v_j)$$

Implementação de Naive Bayes

- Treinamento:
  - Calculamos as probabilidades $P(X_i|C)$ para cada classe C.
- Classificação:
  - Usamos a probabilidades calculadas para determinar a classe mais provável para um novo conjunto de características.

Base de dados

Outlook	Temperature	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

Contagem e probabilidade

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								

Exemplo

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	?

- $P(\text{yes}|\text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$
- $P(\text{no}|\text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

Exemplo

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								

- $P(\text{yes} \mid \text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true}) = P(\text{Outlook} = \text{sunny} \mid \text{yes}) \times P(\text{Temperature} = \text{cool} \mid \text{yes}) \times P(\text{Humidity} = \text{high} \mid \text{yes}) \times P(\text{Windy} = \text{true} \mid \text{yes}) \times P(\text{yes})$;
- $P(\text{yes} \mid \text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true}) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0053$

Frame

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								

- $P(\text{no} \mid \text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true}) = P(\text{Outlook} = \text{sunny} \mid \text{no}) \times P(\text{Temperature} = \text{cool} \mid \text{no}) \times P(\text{Humidity} = \text{high} \mid \text{no}) \times P(\text{Windy} = \text{true} \mid \text{no}) \times P(\text{no})$;
- $P(\text{no} \mid \text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true}) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0206$

Exemplo

- $P(\text{yes} \mid \text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true}) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0053$
- $P(\text{no} \mid \text{Outlook} = \text{sunny}, \text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true}) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0206$

Exemplo

- Normalizando:
  - $P(\text{yes} \mid \ldots) = \frac{0{,}0053}{0{,}0053+0{,}0206} = 20,5\%$;
  - $P(\text{no} \mid \ldots) = \frac{0{,}0206}{0{,}0053+0{,}0206} = 79,5\%$;

Exemplo 2

Outlook	Temperature	Humidity	Windy	Play
Overcast	Cool	High	True	?

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								

Exemplo 2

Outlook	Temperature	Humidity	Windy	Play
Overcast	Cool	High	True	?

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								

- $P(\text{yes} \mid \ldots) = \frac{4}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0,0106$;
- $P(\text{no} \mid \ldots) = 0 \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = \mathbf{0}$;

Naive Bayes

- Valores ausentes nos dados de treino;
- Probabilidade será zero;
- Valor da posterior será zero;
- Independentemente das probabilidades e das outras evidências;
- Estimador laplaciano:
  - Adicionar uma pequena constante α a cada numerador;
  - Adicionar a soma dessas constaste aos denominadores;
  - Exemplo:
    - Probabilidade observadas: 2/9, 4/9 e 3/9;
    - Probabilidade suavizadas com *alpha* = 1: 3/12, 5/12, 4/12;

Exemplo 3

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	3/12	4/8	hot	3/12	3/8	high	4/11	5/7	false	7/11	3/7	10/16	6/16
overcast	5/12	1/8	mild	5/12	3/8	normal	7/11	2/7	true	4/11	4/7		
rainy	4/12	3/8	cool	4/12	2/8								

Exemplo 3

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	3/12	4/8	hot	3/12	3/8	high	4/11	5/7	false	7/11	3/7	10/16	6/16
overcast	5/12	1/8	mild	5/12	3/8	normal	7/11	2/7	true	4/11	4/7		
rainy	4/12	3/8	cool	4/12	2/8								

- $P(\text{yes} \mid \ldots) = \frac{5}{12} \times \frac{4}{12} \times \frac{4}{11} \times \frac{4}{11} \times \frac{10}{16} = 0,0115;$
- $P(\text{no} \mid \ldots) = \frac{1}{8} \times \frac{2}{8} \times \frac{5}{7} \times \frac{4}{7} \times \frac{6}{16} = 0,0048;$
- $P\text{yes} \mid \ldots) = 71\%$
- $P\text{no} \mid \ldots) = 29\%$

Atributos numéricos

- Probabilidades;
- Média;
- Desvio padrões;
- Naive Bayes Gaussiano (NB-G);

Naive Bayes Gaussiano

- Densidade de probabilidade da distribuição normal;

$$P(x|v_j) = \frac{1}{\sigma_{v_j}\sqrt{2\pi}} \times e^{-\frac{1}{2}\left(\frac{x-\mu_{v_j}}{\sigma_{v_j}}\right)^2}$$

- x = Valor numérico que queremos estimar a probabilidade condicional;
- μ = média dos valores observados;
- σ = desvio padrão dos valores observados;

Naive Bayes Gaussiano

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	Média	73	74,6	Média	79,1	86,2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	Desvio P.	6,2	7,9	Desvio P.	10,2	9,7	true	3/9	3/5		
rainy	3/9	2/5											

- Calculando a probabilidade para a temperatura 66;

$$P(Temp = 66|yes) = \frac{1}{6{,}2 \times \sqrt{2\pi}} \times e^{-\frac{(66-73)^2}{2\times 6{,}2^2}} = 0{,}034$$

Naive Bayes Gaussiano

Outlook	Temperature	Humidity	Windy	Play
Sunny	66	90	True	?

$$P(Temp = 66|yes) = \frac{1}{6,2\times\sqrt{2\pi}} \times e^{-\frac{(66-7\ )^2}{2\times 6,2^2}} = 0,034$$

$$P(Temp = 66|no) = \frac{1}{7,9\times\sqrt{2\pi}} \times e^{-\frac{(66-74,6)^2}{2\times 7,9^2}} = 0,028$$

$$P(Hum = 90|yes) = \frac{1}{10,2\times\sqrt{2\pi}} \times e^{-\frac{(90-79,1)^2}{2\times 10,2^2}} = 0,022$$

$$P(Hum = 90|no) = \frac{1}{9,7\times\sqrt{2\pi}} \times e^{-\frac{(90-8\ ,2)^2}{2\times 9,7^2}} = 0,038$$

Naive Bayes Gaussiano

Outlook	Temperature	Humidity	Windy	Play
Sunny	66	90	True	?

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	3/12	4/8	Média	73	74,6	Média	79,1	86,2	false	7/11	3/7	10/16	6/16
overcast	5/12	1/8	Desvio P.	6,2	7,9	Desvio P.	10,2	9,7	true	4/11	4/7		
rainy	4/12	3/8											

$$P(yes|\ldots) = \frac{3}{12} \times 0{,}034 \times 0{,}022 \times \frac{4}{11} \times \frac{10}{16} = 0{,}0000425$$

$$P(no|\ldots) = \frac{4}{8} \times 0{,}028 \times 0{,}038 \times \frac{4}{7} \times \frac{6}{16} = 0{,}000114$$

- P(yes | ...) = 27%; $P(no \mid \ldots) = 73\%$;

Naive Bayes

- Simples;
- Eficiente com grandes conjuntos de dados;
- Interpretar;
- Sensível a valores ausentes e outliers;
- Distribuição de dados desbalanceados;
- Simplicidade e eficiência computacional;
- Bons resultados em muitos problemas práticos;
- Assunção de independência pode ser irrealista.

Atividade

- Implemente o algoritmo Naive Bayes.

Obrigado!
Dúvidas?

Guilherme Henrique de Souza Nakahata

guilhermenakahata@gmail.com

https://github.com/GuilhermeNakahata/UNESPAR-2024