Aprendizagem de Máquina

Alceu Britto Jr.

Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná (PUCPR)

Aprendizagem Bayesiana

Referências

- Duda R., Hart P., Stork D. Pattern Classification 2ed. Willey Interscience, 2002. Capítulos 2 & 3
- Mitchell T. Machine Learning. WCB McGraw–Hill, 1997. Capítulo 6.
- Theodoridis S., Koutroumbas K. Pattern Recognition. Academic Press, 1999. Capítulo 2

Introdução

- O pensamento Bayesiano fornece uma abordagem probabilística para aprendizagem
- Está baseado na suposição de que as quantidades de interesse são reguladas por **distribuições de probabilidade**.
- Distribuição de probabilidade: é uma função que descreve a probabilidade de uma variável aleatória assumir certos valores.

Teorema de *Bayes*

- $P(c|X)$ é chamada de probabilidade *a posteriori* de c porque ela reflete nossa confiança que c se mantenha após termos observado o vetor de treinamento X.
- $P(c|X)$ reflete a influência do vetor de treinamento X.
- Em contraste, a probabilidade *a priori* $P(c)$ é independente de X.

Teorema de *Bayes*

- Geralmente queremos encontrar a classe mais provável $c \in C$, sendo fornecidos os exemplos de treinamento X.

- Ou seja, a classe com o máximo *a posteriori (MAP)*

$$c_{MAP} \equiv \arg\max_{c \in C} P(c \mid X)$$

$$= \arg\max_{c \in C} \frac{P(X \mid c)P(c)}{P(X)}$$

$$= \arg\max_{c \in C} P(X \mid c)P(c)$$

Exemplo

- Atributo alvo: *PlayTennis* (*yes, no*)

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Exemplo

- Logo, temos que estimar:
  - duas probabilidades *a priori* das classes:

    $P(yes) = ?$ $P(no) = ?$
  - probabilidade *a priori* do vetor x_t:

    $P(x_t) = ?$
  - duas probabilidades condicionais:

    $P(x_t \mid yes) = ?$ $P(x_t \mid no) = ?$

- Como fazer isso dadas as 14 instâncias de treinamento da tabela?

Exemplo

$P(yes) = 9/14 = 0{,}643$
$P(no) = 5/14 = 0{,}357$

$P(< \text{outlook} = \text{sunny}, \text{temperature} = \text{hot}, \text{humidity} = \text{high}, \text{wind} = \text{weak} > \mid yes) = ?$
$P(< \text{outlook} = \text{overcast}, \text{temperature} = \text{hot}, \text{humidity} = \text{high}, \text{wind} = \text{weak} > \mid yes) = ?$
$P(< \text{outlook} = \text{rain}, \text{temperature} = \text{hot}, \text{humidity} = \text{high}, \text{wind} = \text{weak} > \mid yes) = ?$

...

$P(< \text{outlook} = \text{rain}, \text{temperature} = \text{cool}, \text{humidity} = \text{normal}, \text{wind} = \text{strong} > \mid yes) =$

....ou seja, temos que estimar todas as probabilidades condicionais, considerando todas as classes possíveis e todos os vetores de características possíveis:

2 x [3 x 3 x 2 x 2] = 72 probabilidades condicionais

Exemplo

$$2 \times [3 \times 3 \times 2 \times 2] = 72$$

pois:

- temos 2 classes
- temos 4 atributos e seus possíveis valores:
  - Outlook (sunny/overcast/rain) [3 valores possíveis]
  - Temperature (hot/mild/cool) [3 valores possíveis]
  - Humidity (high/normal) [2 valores possíveis]
  - Wind (weak/strong) [2 valores possíveis]

- Logo, temos 72 probabilidades condicionais possíveis.

- e $P(x_t)$?

Classificador Ótimo de Bayes

- Limitações práticas

  - Como estimar com confiança todas estas probabilidades condicionais?
    - Conjunto de treinamento com muitas instâncias!
    - Conhecer a distribuição de probabilidade!

  - A probabilidade *a priori* calculada geralmente não reflete a população.

Classificador Naïve Bayes

- *Naïve Bayes* é um dos métodos de aprendizagem mais práticos.

- Quando usar ?
  - disponibilidade de um conjunto de treinamento grande ou moderado.
  - os atributos que descrevem as instâncias forem condicionalmente independentes dada a classe.

- Aplicações bem sucedidas:
  - diagnóstico médico
  - classificação de documentos de textuais

Classificador Naïve Bayes

- O classificador Naïve Bayes é baseado na suposição simplificadora de que os valores dos atributos são condicionalmente independentes dado o valor alvo.

- Ou seja, a probabilidade de observar a conjunção de atributos a_1, a_2,..., a_n é somente o produto das probabilidades para os atributos individuais:

$$P(a_1, a_2, ..., a_n \mid c_j) = \prod_i P(a_i \mid c_j)$$

Classificador Naïve Bayes

- Temos assim o classificador Naïve Bayes:

$$\hat{c}_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(a_i \mid c_j)$$

onde c_{NB} indica o valor alvo fornecido pelo algoritmo Naïve Bayes.

Classificador Naïve Bayes

- Em resumo, o algoritmo Naïve Bayes envolve
  - Aprendizagem: os termos $P(c_j)$ e $P(a_i|c_j)$ são estimados baseado nas suas frequências no conjunto de treinamento.
  - Estas probabilidades “aprendidas” são então utilizadas para classificar uma nova instância aplicando a equação vista anteriormente (c_{NB})

Classificador Naïve Bayes

Algoritmo Naïve Bayes

Treinamento_Naïve_Bayes(*conjunto de exemplos*)
  Para cada valor alvo (classe) c_j
    $P'(c_j)$ ← estimar $P(c_j)$
    Para cada valor de atributo a_i de cada atributo a
      $P'(a_i|c_j)$ ← estimar $P(a_i|c_j)$

Classica_Naïve_Bayes(x_t)

$$\hat{c}_{NB} = \arg\max_{c_j \in C} P'(c_j) \prod_{a_i \in x} P'(a_i \mid c_j)$$

Classificador Naïve Bayes

- Exemplo: Considere novamente os 14 exemplos de treinamento de *PlayTennis* e uma nova instância que o Naïve Bayes deve classificar:

  x_t = <outlook=sunny, temperature=cool, humidity=high, wind=strong>

- A tarefa é predizer o valor alvo (*yes* ou *no*) do conceito *PlayTennis* para esta nova instância.

Classificador Naïve Bayes

- Atributo alvo: *PlayTennis* (*yes, no*)

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Classificador Naïve Bayes

- O valor alvo c_{NB} será dado por:

$$c_{NB} = \arg\max_{c_j \in \{yes, no\}} P(c_j) \prod_i P(a_i \mid c_j)$$

$$= \arg\max_{c_j \in \{yes, no\}} P(c_j) P(Outlook = sunny \mid c_j) P(Temperature = cool \mid c_j)$$

$$P(Humidity = high \mid c_j) P(Wind = strong \mid c_j)$$

- Note que a_i foi instanciado utilizando os valores particulares do atributo da instância x_t.

- Para calcular c_{NB} são necessárias 10 probabilidades que podem ser estimadas a partir dos exemplos de treinamento.

Classificador Naïve Bayes

- Probabilidades *a priori*:

  $P(PlayTennis = yes) = 9/14 = 0.64$

  $P(PlayTennis = no) = 5/14 = 0.36$

- Probabilidades condicionais:

  $P(Wind=strong \mid PlayTennis = yes) = 3/9 = 0.33$

  $P(Wind=strong \mid PlayTennis = no) = 3/5 = 0.60$

  . . .

Classificador Naïve Bayes

- Usando estas estimativas de probabilidade e estimativas similares para os valores restantes dos atributos, calculamos c_{NB} de acordo com a equação anterior (omitindo nome dos atributos) :

  $P(yes)\ P(sunny|\ yes)\ P(cool|\ yes)\ P(high|\ yes)\ P(strong|\ yes) = 0{,}0053$

  $P(no)\ P(sunny|\ no)\ P(cool|\ no)\ P(high|\ no)\ P(strong|\ no) = 0{,}026$

- Então o classificador atribui o valor alvo *PlayTennis = no* para esta nova instância.

Resumo

- Naïve Bayes:
  - é chamado de naïve (simples, não sofisticado), porque assume que os valores dos atributos são condicionalmente independentes.
  - se a condição é encontrada, ele fornece a classificação MAP, caso contrário, pode ainda fornecer bons resultados.