Aprendizagem de Máquina

Alessandro L. Koerich / Alceu S Britto

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

Aprendizagem Bayesiana

Plano de Aula

- Introdução
- Teorema de Bayes
- Classificador Ótimo de Bayes
- Classificador Naïve Bayes
- Resumo

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.

Introdução

- Decisões ótimas podem ser tomadas com base nestas probabilidades conjuntamente com os dados observados.
- Fornece a base para algoritmos de aprendizagem que manipulam probabilidades, bem como para outros algoritmos que não manipulam probabilidades explicitamente.

Introdução

- Os métodos Bayesianos são importantes por dois motivos:
  1. Fornecem algoritmos práticos de aprendizagem:
     - *Naïve Bayes*
     - Redes Bayesianas
     - Combinam conhecimento a priori com os dados observados
     - Requerem probabilidades a priori
  2. Fornecem uma estrutura conceitual útil:
     - "Norma de Ouro" para avaliar outros algoritmos de aprendizagem. Norma de Ouro $\rightarrow$ menor erro possível

Características da Aprendizagem Bayesiana

- Conhecimento *a priori* pode ser combinado com os dados observados para determinar a probabilidade de uma hipótese.
- Métodos Bayesianos podem acomodar hipóteses que fazem predições probabilísticas. Ex.: o paciente tem uma chance de 93% de possuir a doença.
- Novas instâncias podem ser classificadas combinando a probabilidade de múltiplas hipóteses ponderadas pelas suas probabilidades.

Dificuldades Práticas

- Métodos Bayesianos requerem o conhecimento inicial de várias probabilidades.
  - Quando não conhecidas, podem ser estimadas:
    - a partir de conhecimento prévio
    - dados previamente disponíveis
    - suposições a respeito da forma da distribuição.
- Custo computacional linear com o número de hipóteses para determinar a hipótese ótima de Bayes.

Teorema de Bayes

P(*c*|*X*): probabilidade da classe *c* dado o vetor *X*

P(*X*|*c*): probabilidade do vetor *X* dada a classe *c.*

P(*c*): probabilidade *a priori* da classe *c*

$$P(c \mid X) = \frac{P(X \mid c)P(c)}{P(X)}$$

P(*X*): probabilidade *a priori* do vetor de treinamento *X*

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)*

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

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

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

Teorema de *Bayes*

- Desprezamos o termo $P(X)$ porque ele é uma constante independente de c.
- Se assumirmos que cada classe em C é igualmente provável *a priori*, i.e.

$$P(c_i) = P(c_j) \quad \forall\ c_i \text{ e } c_j \text{ em } C$$

- Então, podemos simplificar e escolher a classe de máxima probabilidade condicional (*maximum likelihood = ML*).

Teorema de *Bayes*

- O termo $P(X|c)$ é chamado de probabilidade condicional (ou *likelihood*) de X
- Sendo fornecido c, qualquer classe que maximiza $P(X|c)$ é chamada de uma hipótese ML.

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

Teorema de Bayes: Exemplo

- Considere um problema de diagnóstico médico onde existem duas classes possíveis:
  - O paciente tem H1N1
  - O paciente não tem H1N1
- As características disponíveis são um exame de laboratório com dois resultados possíveis:
  - $\oplus$: positivo
  - $\ominus$: negativo

Teorema de Bayes: Exemplo

- Temos o conhecimento prévio que na população inteira somente 0,008 tem esta doença.
- O exame retorna um resultado positivo correto somente em 98% dos casos nos quais a doença está presente.
- O exame retorna um resultado negativo correto somente em 97% dos casos nos quais a doença não esteja presente.
- Nos outros casos, o teste retorna o resultado oposto.

Teorema de Bayes: Exemplo

- P(H1N1) = ?
- P(⊕|H1N1) = ?
- P(⊕|¬H1N1) = ?

P(¬H1N1) = ?

P(⊖|H1N1) = ?

P(⊖| ¬H1N1) = ?

Teorema de Bayes: Exemplo

- Supondo que um paciente fez um exame de laboratório e o resultado deu positivo.
- O paciente tem H1N1 ou não ?

Aplicando o Teorema de Bayes

- Calculando a classe com maior probabilidade *a posteriori*:
  - $P(\oplus|H1N1)\ P(H1N1) = 0,98 \times 0,008 = 0,0078$
  - $P(\oplus|\neg H1N1)\ P(\neg H1N1) = 0,03 \times 0,992 = 0,0298$
- Assim: $c_{MAP} = \neg H1N1$

Classificador Ótimo de Bayes

- Consideramos até agora a questão:

  "Qual a classe mais provável (c_{MAP}) *dado os exemplos de treinamento X?*"

- Entretanto, a questão mais significativa é na verdade:

  "*Qual é a classificação mais provável de uma nova instância dado os dados de treinamento?*"

- *A classe MAP* (c_{MAP}) é ou não a classificação mais provável?

Classificador Ótimo de Bayes

- A classificação mais provável de uma nova instância *x* é obtida através da maior probabilidade *a posteriori*.

- Assim, a $P(c_j|x)$ que a correta classificação para a instância *x* seja c_j é:

$$\hat{P}(c_j \mid x) = \max_{c_j \in C} P(c_j \mid x)$$

$$\hat{c} = \arg\max_{c_j \in C} P(c_j \mid x)$$

- Qualquer sistema que classifique novas instâncias de acordo com a equação acima é chamada de um *classificador ótimo de Bayes*.

Exemplo

- Exemplo: Considere as 14 instâncias de treinamento de *PlayTennis* e uma nova instância de teste (x_t) que devemos classificar:

x_t = <*Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong*>

- Nossa tarefa é predizer o valor alvo (*yes* ou *no*) do conceito *PlayTennis* para esta nova instância, ou seja:

$$\hat{P}(c_j \mid x_t) = \max_{c_j \in [yes, no]} P(c_j \mid x_t)$$

$$\hat{c} = \arg\max_{c_j \in [yes, no]} P(c_j \mid x_t)$$

Exemplo

- Então, dado x_t, devemos estimar duas probabilidades *a posteriori*:

$$P(c_j = yes \mid x_t) \qquad P(c_j = no \mid x_t)$$

- Aplicando o teorema de Bayes…

$$P(c_j \mid x_t) = \frac{P(x_t \mid c_j) P(c_j)}{P(x_t)}$$

- Ou seja, para estimar a probabilidade *a posteriori*, devemos conhecer:

$$P(x_t) = ? \qquad P(x_t \mid c_j) = ? \qquad P(c_j) = ?$$

.

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

$P(yes) = 9/14 = 0{,}643$

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

$P(<\text{outlook} = \text{sunny}, \text{temperatur e} = \text{hot}, \text{humidity} = \text{high}, \text{wind} = \text{weak} >| yes) = ?$

$P(<\text{outlook} = \text{overcast}, \text{temperatur e} = \text{hot}, \text{humidity} = \text{high}, \text{wind} = \text{weak} >| yes) =$

$P(<\text{outlook} = \text{rain}, \text{temperatur e} = \text{hot}, \text{humidity} = \text{high}, \text{wind} = \text{weak} >| yes) = ?$

…

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

….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 x [3 x 3 x 2 x 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 textuais

Classificador Naïve Bayes

- Se aplica a tarefas de aprendizagem onde:
  - cada instância *x* é descrita por uma conjunção de valores de atributos
  - a função alvo *f*(*x*) pode assumir qualquer valor de um conjunto *V*.
  - um conjunto de exemplos de treinamento da função alvo é fornecido
  - uma nova instância é descrita pela *tupla* de valores de atributos $<a_1, a_2, ..., a_n>$.
- A tarefa é predizer o valor alvo (ou classe) para esta nova instância.

Classificador Naïve Bayes

- A solução *Bayesiana* para classificar uma nova instância consiste em:
  - atribuir o valor alvo mais provável (c_{MAP}) dados os valores dos atributos $<a_1, a_2, ..., a_n>$ que descrevem a instância.

$$c_{MAP} = \arg\max_{c_j \in C} P(c_j \mid a_1, a_2, ..., a_n)$$

- Mas podemos usar o teorema de Bayes para reescrever a expressão . . .

Classificador Naïve Bayes

$$c_{MAP} = \arg\max_{c_j \in C} P(c_j \mid a_1, a_2, ..., a_n)$$

$$c_{MAP} = \arg\max_{c_j \in C} \frac{P(a_1, a_2, ..., a_n \mid c_j) P(c_j)}{P(a_1, a_2, ..., a_n)}$$

$$= \arg\max_{c_j \in C} P(a_1, a_2, ..., a_n \mid c_j) P(c_j)$$

- Devemos agora estimar os dois termos da equação acima baseando-se nos dados de treinamento.
  - $P(c_j)$ é fácil de estimar . . .
  - Porém, $P(a_1,a_2,...,a_n \mid c_j)$ . . .

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:

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

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

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

- 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

- Métodos Bayesianos:
  - acomodam conhecimento prévio e os dados observáveis;
  - atribuem probabilidade a posteriori para cada classe candidata, baseando–se na probabilidade a priori e nos dados.
  - podem determinar a hipótese mais provável (MAP), tendo os dados.
- Bayes Ótimo:
  - combina predições de todas classes, ponderadas pela probabilidade a posteriori, para calcular a classificação mais provável de uma 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** fornecer também bons resultados.

Rede Bayesiana - WEKA