UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Disciplina: Aprendizado de Máquina
Métodos Probabilísticos

Prof. Jefferson Morais
Email: jmorais@ufpa.br

Métodos probabilísticos

- Informações disponíveis podem ser incompletas ou imprecisas
  - Ex. presença de ruído, atributos que não incluem todas as características que permitem predição acurada
  - → Uso de algoritmos baseados no **teorema de Bayes**
    - **Métodos probabilísticos Bayesianos**: modelam relacionamento probabilístico entre atributos de entrada e alvo

P(A|B) em que A pode ser uma classe (doença) e B o conjunto de valores de atributos de entrada (sintomas):
não depende apenas da relação entre A e B, mas também da probabilidade de observar A independentemente de observar B

Métodos probabilísticos

- **Estimativa de P(B)**: frequência com que esse evento ocorre
  - Também é possível estimar probabilidade que B ocorra para cada classe P(B|A)
  - **Interesse é calcular P(A|B)**
    - Probabilidade de objeto pertencer à classe A

Teorema de Bayes calcula P(A|B) usando:
(i) probabilidade *a priori* da classe P(A)
(ii) a probabilidade de observar objetos que pertencem à classe P(B|A)
(iii) probabilidade de ocorrência desses objetos P(B)

Probabilidades

- **Espaço amostral (Ω)**: todos os possíveis resultados de um experimento
  - Também chamado espaço de resultados
  - Valores do conjunto de atributos de entrada
- **Evento (E)**: subconjunto de resultados em Ω
  - Resultado de experimento/observaçã
  - Ex.: Jogar um dado de 6 faces
    - Ω = {1, 3, 3, 4, 2, 5, 1, 6}
    - E = valor do dado < 4 = {1, 3, 3, 2, 1}

Conceitos básicos

- P(E) satisfaz axiomas de *Kolmogorov*
  - P(E) ≥ 0
  - P(Ω) = 1
  - Se A e B são eventos disjuntos → P(AUB) = P(A) + P(B)
  - Eventos disjuntos ou mutuamente exclusivos:

    $P(A \cap B) = \emptyset$
  - Caso Contrário, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probabilidades

- Ex.: características de estudantes

Tipo/número	Mulheres	Homens	Total
Calouros	482	596	1078
Graduação	1916	2316	4232
Pós-graduação	1916	4236	6152
Total estudantes	3832	6552	10384

Probabilidades

- Ex.: características de estudantes
  - Supor que um calouro é selecionado
    - Espaço amostral = 1078 calouros
    - Qual é a probabilidade de uma mulher ter sido selecionada?
      - Evento = ser mulher
      - 45% dos calouros são mulheres (482 / 1078)
      - Se todos têm mesma probabilidade de serem escolhidos, probabilidade de selecionar mulher é 45%

Eventos

- Ex.: um calouro é selecionado
  - Resultado é a pessoa em específico selecionada
    - Evento fundamental
  - Outros eventos
    - Seleção de uma mulher
    - Seleção de alguém de São Paulo
    - Seleção de mulher de Minas Gerais
    - Seleção de qualquer pessoa
      - Evento universal
    - Seleção de nenhum aluno
      - Evento nulo

Eventos

- Diferentes eventos podem ou não se sobrepor
  - Ocorrer para o mesmo resultado

Eventos que não se sobrepõem: mutuamente exclusivos
Ex.: aluno selecionado ser homem ou mulher

- Conjunto de eventos **exaustivo**: ao menos um deles ocorre
  - aluno escolhido tem:
    - Evento 1: menos que 25 anos
    - Evento 2: mais que 17 anos
      - São exaustivos, mas não são mutuamente exclusivos

Eventos

- **Partição**: conjunto de eventos mutuamente exclusivos e exaustivos
  - **Partição fundamental**: contém todos os eventos fundamentais
    - Ex.: Eventos selecionar mulher e selecionar homem formam uma partição
    - Ex.: Eventos fundamentais associados a cada uma das pessoas formam partição fundamental

Partição

- Lei da probabilidade total
- Se B_1, B_2, ..., B_n formam uma partição em **Ω**, então para qualquer evento A:

$$P(A) = \sum_{i=1:n} P(A|B_i) \times P(B_i)$$

Além disso, para qualquer partição: $\sum_{i=1:n} P(B_i) = 1$

Eventos Conjuntos

- Probabilidade de ter duas propriedades diferentes
  - Ex: escolha de caloura (mulher) de Minas Gerais
    - P(M) = probabilidade de ser mulher
    - P(MG) = probabilidade de ser de Minas Gerais
    - P(M,MG) = probabilidade de ser mulher de Minas Gerais Probabilidade
- Probabilidade Conjunta

Probabilidade Conjunta

- Probabilidade de dois eventos A e B ocorrerem simultaneamente

$$P(A \cap B) \text{ ou } P(A,B)$$

- Se eventos são independentes

$$P(A \cap B) = P(A) * P(B)$$

$$P(A|B) = P(A)$$

  - A ocorrência de um não afeta a probabilidade de ocorrência do outro

Eventos conjuntos

- Independência não é usual
  - Fórmula mais geral para a probabilidade do evento conjunto (ambos ocorrerem)
  - **Probabilidades condicionais**: probabilidade de um evento dado que outro ocorreu
    - Ex.: **P(M|MG)** = probabilidade condicional de selecionar mulher, dado que o calouro escolhido é de Minas Gerais

$$P(A,B) = P(B)\,P(A|B) = P(A)\,P(B|A)$$

—— **Teorema de *Bayes***

Eventos conjuntos

- P(M, MG) = P(MG) P(M|MG)
  - Probabilidade de calouro ser mulher de Minas Gerais é probabilidade de estudante ser de Minas Gerais vezes a probabilidade de que, sendo mineira, a pessoa é mulher
- P(M, MG) = P(M) P(MG|M)
  - Probabilidade de calouro ser mulher de Minas Gerais é probabilidade de estudante ser mulher vezes a probabilidade de que a pessoa escolhida, sendo mulher, é mineira

Probabilidade condicional

- Lei da probabilidade condicional

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

Probabilidade em AM

- Sejam dois eventos A e B
  - A: atributo alvo (presença de uma doença)
    - Variável aleatória com dois valores: presença e ausência
  - B: atributo de entrada (resultado de um exame)
    - Variável aleatória com dois valores: positivo e negativo
  - P(A): probabilidade do evento A ocorrer (doença)
    - P(A) = 1 – P(!A)
  - P(B): probabilidade do evento B ocorrer (exame +)
    - P(B) = 1 – P(**!**B)

Probabilidade em AM

- Exemplo

Paciente	Teste	Doença
001	positivo	presente
002	negativo	presente
003	negativo	ausente
004	positivo	presente
005	positivo	ausente
006	positivo	presente
007	negativo	ausente
008	negativo	presente
009	positivo	ausente
010	positivo	presente

Probabilidade *a priori* pode ser estimada pela frequência

$P(-) = 4/10 = 0,4$
$P(+) = 6/10 = 0,6$
$P(\text{presente}) = 6/10 = 0,6$
$P(\text{ausente}) = 4/10 = 0,4$

O que se deseja em AM é a probabilidade *a posteriori*

Probabilidade condicional

- Probabilidade a ***priori x a posteriori*** de um indivíduo estar doente
- Probabilidade de ocorrência de um evento pode depender da ocorrência de outro: P(A|B)
  - Probabilidade de ocorrência de um evento A depende da ocorrência de um evento B
    - Ex.: Probabilidade de estar doente (A) dado que um exame (B) deu positivo
  - Se eventos são independentes → P(A|B) = P(A)

Probabilidade condicional

- É fácil estimar as probabilidades a priori: pela frequência
  - P(B): prob. do resultado do exame ser positiva
  - P(A): prob. do resultado do paciente estar doente
  - P(B|A): prob. do resultado do exame ser positivo dado que o paciente está doente
- É difícil estimar probabilidade *a posteriori*
  - P(A|B): probabilidade do paciente estar doente dado que um exame deu positivo
  - **Uso do teorema de Bayes**: Permite calcular probabilidade *a posteriori* de um evento

Teorema de Bayes

- Teorema de Bayes:
  - P(A|B) = P(B|A)P(A) / P(B)
  - **Posteriori = (verossimilhança do dado x priori) / evidência**
  - P(B): lei da probabilidade total
    - Evento B pode ter dois possíveis resultados, B_1(B) e B_2(**!** B), que formam uma partição em **Ω**

$$P(A) = P(A \cap B_1) + P(A \cap B_2)$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

Aprendizado Bayesiano

- $P(\mathbf{y_i}|\mathbf{x})$ = probabilidade de **x** pertencer a classe $\mathbf{y_i}$
  - Função de custo 0/1 é minimizada se x é associado a $\mathbf{y_k}$ para o qual $P(\mathbf{y_k}|x)$ é máxima
    - Estimativa MAP (Maximum a Posteriori)
  - Predição

$$y_{MAP} = \arg\max_i P(y_i \mid \mathbf{x})$$

Aprendizado Bayesiano

- Funções que calculam probabilidades $P(\mathbf{y_i} | \mathbf{x})$ são chamadas discriminantes
  - Separam exemplos de classes diferentes
  - Teorema de Bayes provê método para calcula $P(\mathbf{y_i} | \mathbf{x})$

$$P(y_i \mid \mathbf{x}) = \frac{P(y_i)\, P(\mathbf{x} \mid y_i)}{P(\mathbf{x})}$$

Aprendizado Bayesiano

- Teorema de Bayes provê método para calcular $P(\mathbf{y_i}|\mathbf{x})$

$$P(y_i \mid \mathbf{x}) = \frac{P(y_i)\, P(\mathbf{x} \mid y_i)}{P(\mathbf{x})}$$

$P(\mathbf{x})$ pode ser ignorado, pois é o mesmo para todas as classes, não afetando os valores relativos de suas probabilidades

Aprendizado Bayesiano

- Assumindo que as probabilidades a priori P(**yi**) são iguais, o cálculo da hipótese mais provável pode ser simplificado

$$y_{MV} = \arg\max_i P(\mathbf{x}|\, y_i)$$

- P(Dados|hipótese) é chamado verossimilhança (*likelihood*)

Aplicabilidade é reduzida devido ao grande número de exemplos necessários para calcular $P(\mathbf{x}|\, y_i)$ de forma viável

Variações para superar esse problema: diferentes discriminantes

Classification Is to Derive the Maximum Posteriori

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector $\mathbf{X} = (x_1, x_2, \ldots, x_n)$
- Suppose there are *m* classes $C_1, C_2, \ldots, C_m$.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

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

- Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized

Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

- This greatly reduces the computation cost: Only counts the class distribution
- If A_k is categorical, $P(x_k|C_i)$ is the # of tuples in C_i having value x_k for A_k divided by $|C_{i,D}|$ (# of tuples of C_i in D)
- If A_k is continous-valued, $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ

and $P(x_k|C_i)$ is

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

$$P(\mathbf{X}|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

Naïve Bayes Classifier: Training Dataset

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Data to be classified:

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayes Classifier: An Example

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357
- Compute $P(X|C_i)$ for each class
  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4
- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

Avoiding the Zero-Probability Problem

- Naïve Bayesian prediction requires each conditional prob. be **non-zero**. Otherwise, the predicted prob. will be zero

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)
- Use **Laplacian correction** (or Laplacian estimator)
  - *Adding 1 to each case*

    Prob(income = low) = 1/1003

    Prob(income = medium) = 991/1003

    Prob(income = high) = 11/1003
  - The "corrected" prob. estimates are close to their "uncorrected" counterparts

Probabilities for weather data

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						
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9						
Rainy	3/9	2/5	Cool	3/9	1/5								

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

Probabilities for Wather data

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								

- A new day:

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

Weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

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

Probability of class "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Missing values

- Training: instance is not included in frequency count for attribute value-class combination
- Classification: attribute will be omitted from calculation
- Example:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of "yes" = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238

Likelihood of "no" = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343

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

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

Numeric attributes

- Usual assumption: attributes have a *normal* or *Gaussian* probability distribution (given the class)
- The *probability density function* for the normal distribution is defined by two parameters:
  - *Sample mean* μ

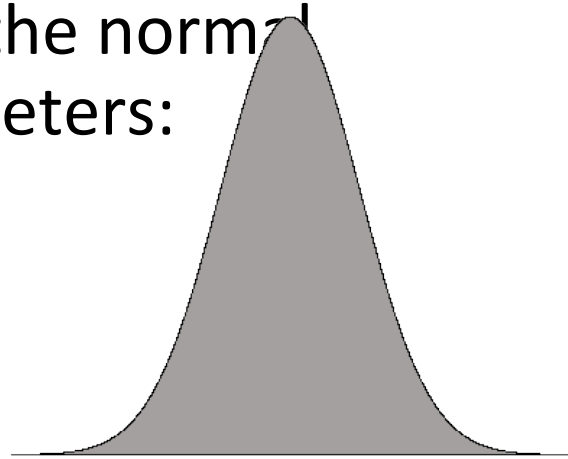
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - *Standard deviation* σ

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

  - Then the density function *f(x) is*

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



Statistics for weather data

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									

- Example density value:

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

Classifying a new day

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of "yes" = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036

Likelihood of "no" = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136

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

P("no") = 0.000136 / (0.000036 + 0. 000136) = 79.1%

- Missing values during training are not included in calculation of mean and standard deviation

Naïve Bayes Classifier: Comments

- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables
    - E.g., hospitals: patients: Profile: age, family history, etc.
      Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
    - Dependencies among these cannot be modeled by Naïve Bayes Classifier
- How to deal with these dependencies? Bayesian Belief Networks (Chapter 9)