Aula 7 – Mineração de Dados Classificação - Parte 2

Profa. Elaine Faria

UFU

Classificadores Bayesianos

- São classificadores estatísticos que classificam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe
- Naive Bayes
  - É um dos mais simples e bem difundidos algoritmos baseados no Teorema de Bayes
  - Supõe como hipótese de trabalho que o efeito do valor de um atributo não-classe é independente dos valores dos outros atributos
    - O valor de um atributo não influencia o valor dos outros

Ideia geral do Naive Bayes

- Consideramos um banco de dados de amostras classificadas em *m* classes distintas

$$C_1, C_2, \ldots C_m$$

- Suponha que X é uma tupla a ser classificada (não está no banco de dados de amostras)
- O classificador vai classificar X numa classe C para a qual a probabilidade condicional P[C|X] á a mais alta possível
- Assim

$$P[C_i|X] > P[C_j|X]$$

para todas as outras classes Cj , Cj ≠ Ci

A probabilidade $P[C_i|X]$ também é chamada **probabilidade posterior**

Probabilidade conjunta

- P(X): probabilidade do evento X ocorrer
- P(C): probabilidade do evento C ocorrer
- P(X & C): probabilidade de X e C ocorrerem
  - Se os eventos forem independentes: P(X & C) = P(X) * P (C) → um evento não afeta a probabilidade de ocorrência do outro
  - Se os eventos não forem independentes: P(X & C) = P(X) * P(C|X)
  - P(C|X) = P(X&C) / P(X) → probabilidade de que C ocorra dado que X ocorreu

Teorema de Bayes

- Note que: $P(X \& C) = P(C \& X)$
- Logo: $P(C|X) * P(X) = P(X|C) * P(C)$
- Teorema de Bayes: $\boxed{P(C|X) = P(X|C) * P(C) / P(X)}$

- $P(X)$ é constante
- Maximizar $P(X|C) * P(C)$
- $P(X|C)$ é chamada de **probabilidade a priori**

Como calcular a probabilidade a priori?

- Suponha $X = (x_1, x_2, ... x_k)$ representa o evento conjunto $x_1 \cap x_2 \cap x_k$. Logo

$$P[X|C] = P[x_1 \cap x_2 \cap .... x_k|C]$$

Supondo a indepedência dos atributos, temos:

$$P[X|C] = P[x_1|C] * P[x_2|C] * ... * P[X_n|C]$$

$P[x_i|C]$ pode ser calculada a partir da base de dados rotulada

Como calcular a probabilidade a priori?

- **Se o atributo Ai é categórico**

$$P[x_i|C] \ = \ \frac{nro\ de\ tuplas\ classificadas\ em\ C\ com\ atributo\ A_i\ =\ x_i}{nro\ de\ tuplas\ classificadas\ em\ C}$$

- **Se o atributo Ai é contínuo**

$$P[x_i|C] \ = \ g(xi, \mu_c, \sigma_c)$$

função de distribuição gaussiana onde μ é a média e σ é o desvio padrão

$$g(xi, \mu_c, \sigma_c) = \frac{1}{\sqrt{2\pi} * \sigma_c} e^{-\frac{(x_i - \mu_c)^2}{2{\sigma_c}^2}}$$

Exemplo

Dada a instância de Teste:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- Precisamos estimar

  P(Evade = Yes | X) and P(Evade = No | X)

  Vamos substituir

  Evade = Yes por Yes, e

  Evade = No por No

Retirado de: Tan P., SteinBack M, Karptane, A. e Kumar V. Introduction to Data Mining, 2n edition, Pearson, 2018

Exemplo

Dada a instância de Teste:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- Usando o teorema de Bayes

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

$$P(No \mid X) = \frac{P(X \mid No)P(No)}{P(X)}$$

Como estimar P(X|Yes) e P(X|No)?

Retirado de: Tan P., SteinBack M, Karptane, A. e Kumar V. Introduction to Data Mining, 2n edition, Pearson, 2018

Exemplo

Dada a instância de Teste:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

P(X | Yes) =
P(Refund = No | Yes) x
P(Divorced | Yes) x
P(Income = 120K | Yes)

P(X | No) =
P(Refund = No | No) x
P(Divorced | No) x
P(Income = 120K | No)

Retirado de: Tan P., SteinBack M, Karptane, A. e Kumar V. Introduction to Data Mining, 2n edition, Pearson, 2018

Exemplo

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- P(y) = fração de instâncias da classe Y

  P(No) = 7/10,
  P(Yes) = 3/10

- Atributos Categóricos - Exemplo

  P(Status=Married|No) = 4/7
  P(Refund=Yes|Yes)=0

- Atributos Contínuos - Exemplo

  Para (Income, Classe=No):

  Se Classe=No

  - média = 110
  - desvio padrão = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Retirado de: Tan P., SteinBack M, Karptane, A. e Kumar V. Introduction to Data Mining, 2n edition, Pearson, 2018

Exemplo

Dada a instância de teste: $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(X | No) = P(Refund=No | No)
  $\times$ P(Divorced | No)
  $\times$ P(Income=120K | No)
  = 4/7 $\times$ 1/7 $\times$ 0.0072 = 0.0006

- P(X | Yes) = P(Refund=No | Yes)
  $\times$ P(Divorced | Yes)
  $\times$ P(Income=120K | Yes)
  = 1 $\times$ 1/3 $\times$ 1.2 $\times$ 10^{-9} = 4 $\times$ 10^{-10}

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Classe = No

Problema da Frequência Zero

- O que acontece se um determinado valor de atributo não aparece na base de treinamento, mas aparece no exemplo de teste?

Problema da Frequência Zero

Considere que o Tid = 7 foi deletado

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Classificador Naïve Bayes

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Dado X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0
P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

Naïve Bayes não será capaz de classificar X em Yes ou No!

Problema da Frequência Zero

- Se uma das probabilidades condicionais é zero, então a expressão inteira torna-se zero
  - É preciso usar outras estimativas de probabilidade condicional
- **Estimativa de probabilidade:**

Original: $P(X_i = c|y) = \frac{n_c}{n}$

Estimativa de Laplace: $P(X_i = c|y) = \frac{n_c + 1}{n + v}$

Estimativa M: $P(X_i = c|y) = \frac{n_c + mp}{n + m}$

n: número de instâncias de treinamento que pertencem à classe y

n_c: número de instâncias com $X_i = c$ and $Y = y$

v: número total de valores do atributo que X_i pode assumir

p: estimativa inicial de $(P(X_i = c|y))$ conhecida a priori

m: hiper-parâmetro para o nível de confiança em p

Valores Ausentes

- Treinamento
  - Excluir exemplo do conjunto de treinamento

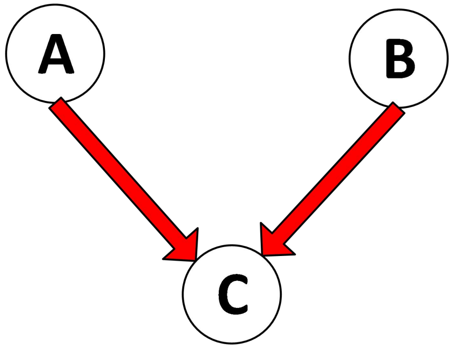
- Classificação
  - Considerar apenas os demais atributos

Naive Bayes - Resumo

- Robusto a ruídos isolados
- Lida com valores ausentes, ignorando a instância durante os cálculos de estimativa de probabilidade
- Robusto a atributos irrelevantes
- Atributos redundantes e correlacionados podem violar a suposição condicional de classe
  - É possível usar outras técnicas, como Bayesian Belief Networks (BBN)

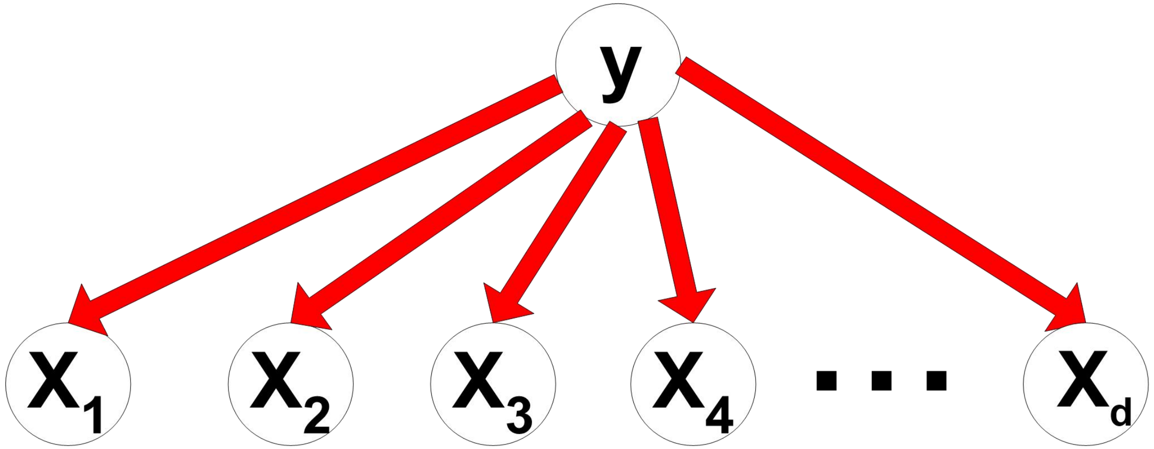
Redes Bayesianas

- É comum existir dependência entre os atributos.
    - Neste caso, pode-se utilizar uma Rede Bayseana de Crença
- Rede Bayesiana de Crença é uma estrutura com duas componentes
    - Um grafo dirigido acíclico onde
        - Cada vértice representa um atributo
        - Os arcos ligando os vértices representam uma dependência entre estes atributos
    - Tabela de Probabilidade Condicional (CPT) para cada atributo Z
        - Associa cada nó ao seu pai



Redes Bayesianas

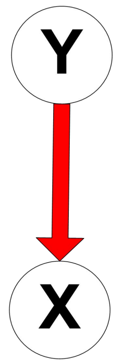
- Suposição do Naive Bayes



Redes Bayesianas

- Se X não tem pai, a tabela contém a probabilidade a priori P(X)

- Se X tem somente um nó pai (Y), a tabela contém a probabilidade condicional P(X|Y)

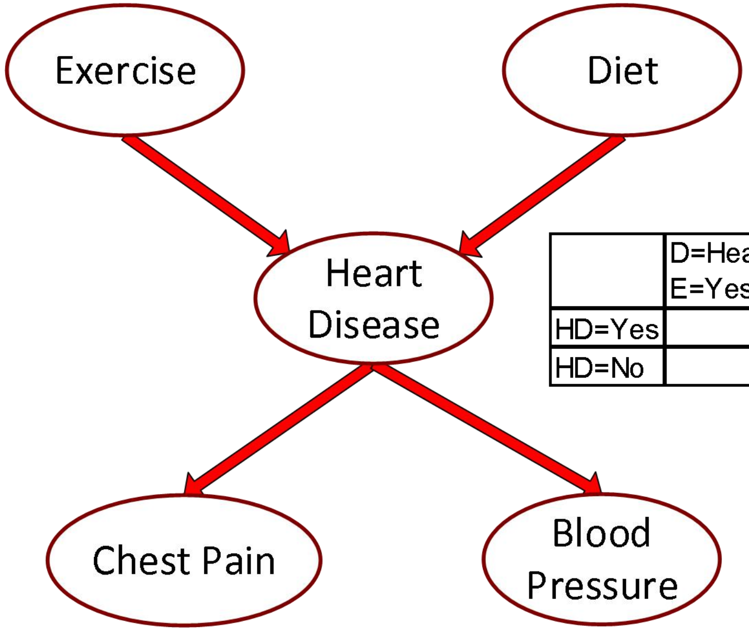


- Se X tem múltiplos nó pai (Y_1, Y_2,…, Y_k), a tabela contém a probabilidade condicional P(X|Y_1, Y_2,…, Y_k)

Redes Bayesianas - Exemplo

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	D=Healthy E=Yes	D=Healthy E=No	D=Unhealthy E=Yes	D=Unhealthy E=No
HD=Yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Redes Bayesianas - Exemplo

- Dado: X = (E=No, D=Yes, CP=Yes, BP=High)
  - Calcular P(HD|E,D,CP,BP)?

- P(HD=Yes| E=No,D=Yes) = 0.55
  P(CP=Yes| HD=Yes) = 0.8
  P(BP=High| HD=Yes) = 0.85
  - P(HD=Yes|E=No,D=Yes,CP=Yes,BP=High)
    $\propto 0.55 \times 0.8 \times 0.85 = 0.374$

- P(HD=No| E=No,D=Yes) = 0.45
  P(CP=Yes| HD=No) = 0.01
  P(BP=High| HD=No) = 0.2
  - P(HD=No|E=No,D=Yes,CP=Yes,BP=High)
    $\propto 0.45 \times 0.01 \times 0.2 = 0.0009$

Classifica X como Yes

Referências

- Katti, F.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. L. F.   Inteligência Artificial: Uma abordagem de Aprendizado de Máquina, LTC, 2011
- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006