

KDD E MINERAÇÃO DE DADOS

MÉTODOS ESTATÍSTICOS

Prof. Ronaldo R. Goldschmidt

rribeiro@univercidade.br

ronaldo@de9.ime.eb.br

geocities.yahoo.com.br/ronaldo_goldschmidt

CLASSIFICADORES BAYESIANOS

Sejam: X um conjunto de informações, $C_1, C_2, \dots, C_n$ Classes e 0 um objeto / caso.

Um Classificador Bayesiano é um método estatístico que:

- a) Computa $P(C=C_k/0)$ para $K=1, 2, \dots, n$
- b) Seleciona a classe C_k com maior probabilidade

CLASSIFICADORES BAYESIANOS

Cálculo das Probabilidades Condicionais:

$$P(C = C_k | 0) = ?$$

$$P(C = C_k | A_1 = a_1, A_2 = a_2, A_3 = a_3, \dots) = (\text{Subst. 0})$$

$$\frac{P(A_1 = a_1, A_2 = a_2, A_3 = a_3, \dots | C = ck) * P(c = ck)}{P(A_1 = a_1, A_2 = a_2, A_3 = a_3, \dots)} = (\text{Regra Bayes})$$

Denominador comum $\Rightarrow$ pode ser descartado

$$P(A_1 = a_1, A_2 = a_2, A_3 = a_3, \dots | C = ck) * P(C=ck)$$

CLASSIFICADORES BAYESIANOS

Cálculo das Probabilidades Condicionais:

Sup. Independência entre os atributos – Classificador Ingênuo

$P(C = C_k | 0)$ é representada por:

$$P(A_1 = a_1 | C = ck) * P(A_2 = a_2 | C=ck) * \dots * P(C=ck)$$

Onde

$$P(A_i = a_i | C = ck) = \frac{\text{count}(A_i = a_i | C = ck)}{\text{count}(C = ck)}$$

CLASSIFICADORES BAYESIANOS

Exemplo:

Consideremos o seguinte BD: (Jogar Tênis)

DIA	APARÊNCIA	TEMPERATURA	UMIDADE	VENTO	JOGAR TÊNIS
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Moderado	Alta	Fraco	Sim
D5	Chuvoso	Fresco	Normal	Fraco	Sim
D6	Chuvoso	Fresco	Normal	Forte	Não
D7	Nublado	Fresco	Normal	Forte	Sim
D8	Ensolarado	Moderado	Alta	Fraco	Não
D9	Ensolarado	Fresco	Normal	Fraco	Sim
D10	Chuvoso	Moderado	Normal	Fraco	Sim
D11	Ensolarado	Moderado	Normal	Forte	Sim
D12	Nublado	Moderado	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Moderado	Alta	Forte	Não

CLASSIFICADORES BAYESIANOS

Exemplo:

São 2 Classes: Jogar = Sim, Jogar = Não

Os atributos A_i são: Aparência, Temperatura, Umidade e Vento

Pergunta-se: Um dia ensolarado, quente, de alta umidade e vento fraco é adequado para jogar tênis?

CLASSIFICADORES BAYESIANOS

Exemplo:

$$P(\text{Jogar} = \text{Sim} \mid \text{ensolarado, quente, alta umidade, vento fraco}) = 0,0071$$

$$P(\text{Jogar} = \text{Não} \mid \text{ensolarado quente, alta umidade, vento fraco}) = 0,0274$$

Resposta do C.Bayesiano: Jogar = Não

CLASSIFICADORES BAYESIANOS

Exemplo:

Cálculo das Probabilidades:

$$P(\text{Jogar} = \text{Sim} \mid \text{ensolarado, quente, alta umidade, vento fraco}) = (2/9) * (2/9) * (3/9) * (6/9) * (9/14) = 0,0071$$

$$P(\text{Jogar} = \text{Não} \mid \text{ensolarado, quente, alta umidade, vento fraco}) = (3/5) * (2/5) * (4/5) * (2/5) * (5/14) = 0,0274$$

C4.5

- **Baseado no Cálculo de Entropia:** Medida da Teoria da Informação.

Seja $S(A_1, A_2, \dots, A_n, C)$ um conjunto de dados, sendo C o atributo objetivo do problema. $C_1, C_2, \dots, C_k$ são valores de C .

Exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$A_1 - \text{Renda}$
 $A_2 - \text{Despesa}$
 $C - \text{Resposta}$
 $C_1 - AP$
 $C_2 - ES$
 $C_3 - NG$

C4.5

- **Baseado no Cálculo de Entropia:** Medida da Teoria da Informação.

Seja $S(A_1, A_2, \dots, A_n, C)$ um conjunto de dados, sendo C o atributo objetivo do problema. $C_1, C_2, \dots, C_k$ são valores de C .

$$info(S) = - \sum_{j=1}^k \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \text{ bits}$$

$$info_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i) \quad \text{gain}(X) = info(T) - info_x(T)$$

C4.5

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$info(S) = - \sum_{j=1}^k \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \text{ bits}$$

$$info(S) = - (1/4 \log_2 1/4) - (1/4 \log_2 1/4) - (2/4 \log_2 2/4)$$

$$info(S) = - (0.25 * -2) - (0.25 * -2) - (0.5 * -1)$$

$$info(S) = 1.5$$

C4.5

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$info_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i)$$

$$info_{Renda}(T) = 2/4 * [- (1/2 \log_2 1/2) - (1/2 \log_2 1/2) - (0/2 \log_2 0/2)]$$

$$+ 2/4 * [- (2/2 \log_2 2/2) - (0/2 \log_2 0/2) - (0/2 \log_2 0/2)]$$

$$info_{Renda}(T) = 0.5$$

C4.5

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$info_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i)$$

$$\begin{aligned}
 info_{Despesa}(T) &= \frac{2}{4} * [- (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{0}{2} \log_2 \frac{0}{2})] \\
 &\quad + \frac{2}{4} * [- (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{0}{2} \log_2 \frac{0}{2})] \\
 info_{Despesa}(T) &= 1.0
 \end{aligned}$$

C4.5

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$gain(X) = info(T) - info_x(T)$$

$$\begin{aligned}
 info(S) &= 1.5 & info_{Renda}(T) &= 0.5 & gain(Renda) &= 1.0 \\
 \text{Considere } S = T & & info_{Despesa}(T) &= 1.0 & gain(Despesa) &= 0.5 \\
 info(T) &= 1.5 & & & &
 \end{aligned}$$

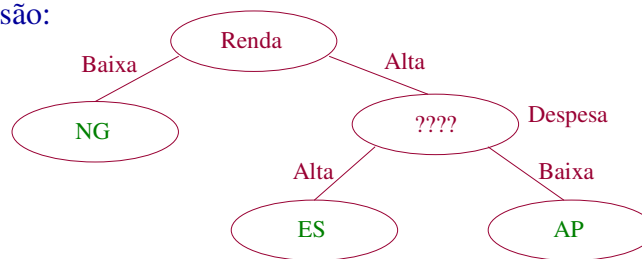
O atributo Renda possui mais informação → deve ser considerado primeiro

C4.5

No exemplo:

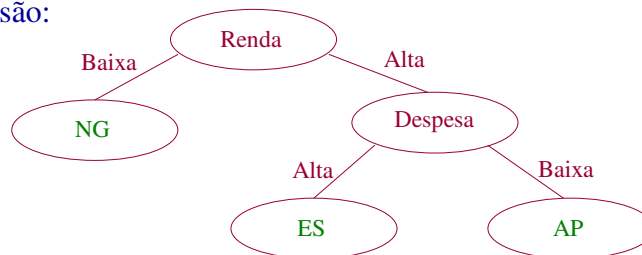
Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

Árvore de Decisão:



C4.5

Árvore de Decisão:



Regras de Produção:

- 1) Se Renda = Baixa Então Resultado = NG
- 2) Se Renda = Alta E Despesa = Alta Então Resultado = ES
- 3) Se Renda = Alta E Despesa = Baixa Então Resultado = AP