Curso de Data Mining

**Sandra de Amo**

# Classificadores Bayesianos

Classificadores Bayesianos são classificadores estatísticos que classificam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe. Produz resultados rapidamente, de grande correção quando aplicados a grandes volumes de dados, comparáveis aos resultados produzidos por árvores de decisão e redes neurais.

## 0.1 Classificadores Bayesianos Simples

Os classificadores Bayesianos Simples supõem como hipótese de trabalho que o efeito do valor de um atributo não-classe é independente dos valores dos outros atributos. Isto é, o valor de um atributo não influencia o valor dos outros. Esta hipótese tem como objetivo facilitar os cálculos envolvidos na tarefa de classificação.

Por exemplo, suponhamos os atributos Idade, Profissão, Renda. É claro que estes atributos não são independentes uns dos outros. Uma pessoa com profissão Médico tem maior probabilidade de ter uma renda alta do que um porteiro. Uma pessoa com idade superior a 40 tem maior probabilidade de ter uma renda alta do que alguém com menos de 25 anos. Assim, vemos que os valores dos atributos dependem uns dos outros. Por outro lado, é claro que os atributos Gênero, Cidade, Idade são independentes uns dos outros.

Um *Classificador Bayesiano Ingênuo ou Simples* funciona da seguinte maneira:

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 $P[C|X]$ é a mais alta possível. Repare que os valores dos atributos de X podem ser encarados como um *evento conjunto*. Assim, se os atributos do banco de dados são Idade, Profissão e Renda e $X = (30..40, Professor, Alta)$, então $X$ pode ser vista como o evento Idade = 30..40, Profissão = Professor e Renda = Alta. $X$ será classificada na classe $C$ se a probabilidade condicional de $C$ acontecer dado que $X$ acontece é maior do que a probabilidade de qualquer outra classe $C'$ acontecer dado que $X$ acontece.

Assim, a tupla $X$ será classificada na classe $C_i$ se

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

para todas as outras classes $C_j$, $C_j \neq C_i$. Esta probabilidade $P[C_i|X]$ também é chamada *probabilidade posterior*.

### Como calcular as probabilidades posteriores

O Teorema de Bayes fornece uma maneira de calcular $P[C_i|X]$. Sabemos que :

$$P[X \cap C] = P[X|C] * P[C] = P[C|X] * P[X]$$

Logo:

$$P[C|X] = \frac{P[X|C] * P[C]}{P[X]} \qquad \text{(Teorema de Bayes)}$$

Como $P[X]$ é uma constante (pois $X$ está fixo), a fim de maximizar $P[C|X]$ precisamos maximizar o numerador $P[X|C] * P[C]$. Se as probabilidades $P[C]$ não estão disponíveis para cada classe $C$, supõe-se que elas são idênticas, isto é, $P[C] = \frac{1}{m}$. Em seguida, resta somente maximizar $P[X|C]$. Caso as probabilidades $P[C]$ sejam conhecidas, maximiza-se o produto $P[X|C] * P[C]$.

Como é calculada a probabilidade condicional $P[X|C]$, também chamada *probabilidade a priori* ? Suponha que $X = (x_1, x_2, \ldots, x_k)$. $X$ representa o evento conjunto $x_1 \cap x_2 \cap \ldots \cap x_k$. Logo,

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

Supondo a hipótese da *independência dos atributos* discutida acima, temos que:

$$P[x_1 \cap x_2 \cap \ldots \cap x_k|C] = P[x_1|C] * P[x_2|C] * \ldots * P[x_k|C]$$

As probabilidades $P[x_i|C]$ podem ser calculadas a partir da base de amostras da seguinte maneira:

- Se o atributo $A_i$ é categórico:

$$P[x_i|C] = \frac{\text{n}^o \text{ de tuplas classificadas em } C \text{ com atributo } A_i = x_i}{\text{n}^o \text{ de tuplas classificadas em } C}$$

- Se o atributo $A_i$ é contínuo (não-categórico):

$$P[x_i|C] = g(x_i, \mu_C, \sigma_C) = \text{função de distribuição de Gauss}$$

onde $\mu_C$ = média, $\sigma_C$ = desvio padrão.

A distribuição de Gauss é dada por :

$$g(x_i, \mu_C, \sigma_C) = \frac{1}{\sqrt{2\Pi} * \sigma_C} e^{-}(\frac{x_i - \mu_C)^2}{2\sigma_C^2}$$

**Exemplo de uso**

Consideremos o seguinte banco de dados (o atributo classe é Compra-Computador):

| ID | Idade | Renda | Estudante | Crédito | Compra-Computador |
|---|---|---|---|---|---|
| 1 | $\leq 30$ | Alta | não | bom | não |
| 2 | $\leq 30$ | Alta | não | bom | não |
| 3 | 31..40 | Alta | não | bom | sim |
| 4 | $> 40$ | Média | não | bom | sim |
| 5 | $> 40$ | Baixa | sim | bom | sim |
| 6 | $> 40$ | Baixa | sim | excelente | não |
| 7 | 31..40 | Baixa | sim | excelente | sim |
| 8 | $\leq 30$ | Média | não | bom | não |
| 9 | $\leq 30$ | Baixa | sim | bom | sim |
| 10 | $> 40$ | Média | sim | bom | sim |
| 11 | $\leq 30$ | Média | sim | excelente | sim |
| 12 | 31..40 | Média | não | excelente | sim |
| 13 | 31..40 | Alta | sim | bom | sim |
| 14 | $> 40$ | Média | não | excelente | não |

A classe $C_1$ corresponde a Compra-Computador = 'sim' e a classe $C_2$ corresponde a Compra-Computador = 'não'. A tupla desconhecida que queremos classificar é :

$$X = (Idade = \leq 30, Renda = Media, Estudante = sim, Credito = bom)$$

Precisamos maximizar $P[X|C_i]P[C_i]$ para $i = 1, 2$. As probabilidades $P[C_i]$ podem ser calculadas baseando-se no banco de dados de amostras:

$$P[C_1] = \frac{9}{14} = 0.643$$

$$P[C_2] = \frac{5}{14} = 0.357$$

Para calcular $P[X|C_i]$, para $i = 1, 2$, calculamos as seguintes probabilidades:

$$
\begin{aligned}
P[Idade \leq 30 | CompraComp = sim] &= \frac{2}{9} = 0.222 \\
P[Idade \leq 30 | CompraComp = nao] &= \frac{3}{5} = 0.6 \\
P[Renda = Media | CompraComp = sim] &= \frac{4}{9} = 0.444 \\
P[Renda = Media | CompraComp = nao] &= \frac{2}{5} = 0.4 \\
P[Estudante = sim | CompraComp = sim] &= \frac{6}{9} = 0.667 \\
P[Estudante = sim | CompraComp = nao] &= \frac{1}{5} = 0.2 \\
P[Credito = bom | CompraComp = sim] &= \frac{6}{9} = 0.667 \\
P[Credito = bom | CompraComp = nao] &= \frac{2}{5} = 0.4
\end{aligned}
$$

Utilizando as probabilidades acima, temos:

$$
\begin{aligned}
P[X | CompraComp = sim] &= 0.222 * 0.444 * 0.667 * 0.667 = 0.044 \\
P[X | CompraComp = nao] &= 0.6 * 0.4 * 0.2 * 0.4 = 0.019 \\
P[X | CompraComp = sim] * P[CompraComp = sim] &= 0.044 * 0.643 = 0.028 \\
P[X | CompraComp = nao] * P[CompraComp = nao] &= 0.019 * 0.357 = 0.007
\end{aligned}
$$

Desta maneira, o classificador Bayesiano prediz que a tupla $X$ é classificada na classe Compra-Computador = 'sim'.

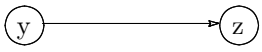
## 0.2 Redes Bayseanas de Crença

Quando a hipótese da independência entre os atributos se verifica, então o classificador Bayesiano simples (ou ingênuo) é o que tem a melhor performance em termos de resultados corretos, com relação a outros classificadores. Entretanto, na prática, é comum existir dependência entre os atributos. Neste caso, utilizamos uma *Rede Bayseana de Crença* como método classificador.

### O que é uma Rede Bayseana de Crença ?

Uma Rede Bayseana de Crença é uma estrutura com duas componentes:

1. Um grafo dirigido acíclico onde cada vértice representa um atributo e os arcos ligando os vértices representam uma dependência entre estes atributos:




z depende de y

y = pai

z = filho

2. A segunda componente definindo uma Rede Bayseana de Crença consiste de uma *Tabela de Probabilidade Condicional* (CPT) **para cada atributo** $Z$. A probabilidade condicional associada ao atributo $Z$ é a probabilidade $P[Z|pais(Z)]$, onde $pais(Z)$ é o conjunto dos atributos que são pais de $Z$. A tabela CPT para um atributo Z, tem o seguinte formato : as linhas correspondem aos possíveis valores de $Z$, as colunas correspondem às combinações de valores possíveis dos pais($Z$). Na linha $i$, coluna $j$, temos a probabilidade condicional de $Z$ ter o valor da linha $i$ e seus pais terem os valores especificados na coluna $j$.

**Classificação utilizando Rede Bayseana de Crença**

O processo de classificação utilizando Redes Bayseanas de Crença tem como **input** um banco de dados de amostras e uma Rede Bayseana (com seus dois componentes especificados acima). Um dos vértices da Rede Bayseana é selecionado como sendo o atributo classe. Podem existir diversos atributos classe numa Rede Bayseana. Caso só tenhamos um único atributo classe assumindo os valores $C_1, ..., C_m$, o **output** será a distribuição de probabilidade $P[C_1|X], ..., P[C_m|X]$. Caso existam diversos atributos classe, por exemplo $Classe_1$ e $Classe_2$, assumindo os valores $C_1^1, ..., C_m^1$ (para a $Classe_1$) e $C_1^2, ..., C_l^2$ (para a $Classe_2$), o algoritmo retornará a a distribuição de probabilidade $P[C_1^1|X], ..., P[C_m^1|X]$,$P[C_1^2|X], ..., P[C_l^2|X]$.

O processo de classificação é análogo ao classificador Bayesiano simples: procura-se maximizar a probabilidade condicional posterior $P[C_i|X]$. Segundo o Teorema de Bayes:

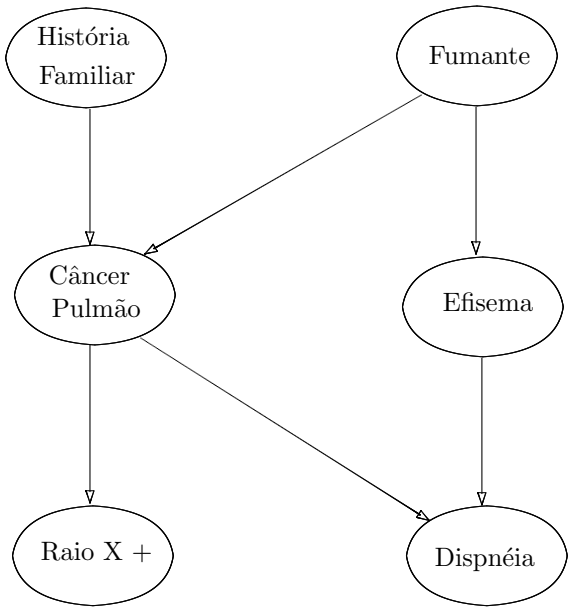
$$P[C_i|X] = \frac{P[X|C_i] * P[C_i]}{P[X]}$$

Logo, precisamos maximizar $P[X|C_i] * P[C_i]$, já que $P[X]$ é fixo. Agora, não podemos mais utilizar a hipótese simplicadora da independência dos atributos no cálculo de $P[X|C_i]$. O produto $P[X|C_i] * P[C_i]$ é a probabilidade do evento conjunto $X \cap C_i$. Supondo $X = (x_1, ..., x_k)$ (como já dissemos, a tupla $X$ é vista como um evento conjunto $x_1 \cap x_2 \cap ...x_k$) temos que :

$$P[X|C_i]*P[C_i] = P[X \cap C_i] = P[x_1|pais(x_1)]*P[x_2|pais(x_2)]*...*P[x_k|pais(x_k)]*P[C_i|pais(C_i)]$$

**Exemplo de uso**

Ilustraremos o processo de classificação utilizando Redes Bayseanas de Crença através de um exemplo. Consideremos a Rede Bayseana cuja primeira componente é o seguinte grafo:




O único atributo classe é CancerPulmão. A tabela CPT para este atributo é a seguinte:

| | HF=1,F=1 | HF=1,F=0 | HF=0,F=1 | HF=0,F=0 |
|---|---|---|---|---|
| CP = 1 | 0.8 | 0.5 | 0.7 | 0.1 |
| CP = 0 | 0.2 | 0.5 | 0.3 | 0.9 |

A tabela CPT para o atributo Efisema é :

| | F=1 | F=0 |
|---|---|---|
| E = 1 | 0.03 | 0.2 |
| E = 0 | 0.97 | 0.8 |

A tabela CPT para o atributo RaioX+ é :

| | CP=1 | CP=0 |
|---|---|---|
| RX+ = 1 | 0.9 | 0.02 |
| RX+ = 0 | 0.1 | 0.98 |

A tabela CPT para o atributo Dispnéia é :

| | E=1,CP=1 | E=1,CP=0 | E=0,CP=1 | E=0,CP=0 |
|---|---|---|---|---|
| D = 1 | 0.99 | 0.2 | 0.3 | 0.01 |
| D = 0 | 0.01 | 0.8 | 0.7 | 0.99 |

Estamos utilizando as seguintes abreviações: CP = CâncerPulmão, HF = História Familiar e F = Fumante, D = Dispnéia, RX+ = RaioX+, Efisema = E.

Suponhamos a seguinte tupla $X = (HF = 1, F = 1, E = 0, RaioX+ = 1, Dispneia = 0)$ a ser classificada. Queremos maximizar $P[X|CP] * P[CP]$. Sabemos que :

$P[X|CP = i] * P[CP = i] = P[HF = 1] * P[F = 1] * P[CP = i|HF = 1, F = 1] *$
$P[E = 0|F = 1] * P[RX+ = 1|CP = i] * P[D = 0|CP = i, E = 0]$

Para maximizar a expressão acima, precisamos maximizar

$$P[CP = i|HF = 1, F = 1] * P[RX+ = 0|CP = i] * P[D = 0|CP = i, E = 1]$$

já que os demais elementos envolvidos no produto acima são constantes ($P[HF = 1], P[F = 1], P[E = 0|F = 1]$).

**Exercício**: Utilizando as tabelas CPT para cada um dos atributos RX+, CP, E, D, determine qual o valor de CP que maximiza o produto $P[CP|HF = 1, F = 1] * P[RX+ = 1|CP = i] * P[D = 0|CP = i, E = 0]$. Esta será a classe na qual será classificada a tupla $X$.

# References


[1] D. Heckerman. Bayesian Networks for Knowledge Discovery. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy Editors. Advances in Knowledge Discovery and Data Mining, pages 273-305. MIT Press, 1996.

[2] Pedro Domingos, Michael Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Proc. International Conference on Machine Learning, 1996, pages 105-112, Morgan Kaufmann.