Exercício INE5644 - Data Mining
Classificação - Árvores de Decisão

Bruno Marques do Nascimento*

23 de Abril de 2018

Contexto:

Considere o seguinte conjunto de treinamento, em que cada exemplo é definido por três atributos (A,B,C) e a classe X.

Tabela 1 – Conjunto de treinamento

	A	**B**	**C**	**X**
X1	1	1	3	P
X2	1	2	4	P
X3	2	2	4	N
X4	2	1	4	N

Sabendo que:

- $\boldsymbol{Entropia(S) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-)}$
- $\boldsymbol{Ganho(S, A) = Entropia(S) - \sum((|Sv|/|S|) * Entropia(Sv))}$, onde:
  - **Ganho(S, A)**: ganho do atributo A sobre o conjunto S
  - **Sv**: subconjunto de S para um valor do atributo A.
  - **|Sv|**: número de elementos de Sv.
  - **|S|**: número de elementos de S.

Tabela 2 – Dados fornecidos

$\log_2 1$	=	0
$\log_2 0.5$	=	-1
$\log_2 0.25$	=	-2
$\log_2 0.75$	=	-0.415
$\log_2 0.333$	=	-1.585
$\log_2 0.667$	=	-0.585

*brunomn95@gmail.com - Universidade Federal de Santa Catarina - Matrícula: 15104098

Perguntas:

a) Qual a incerteza (entropia) associada ao conjunto de treinamento inicial?

$$
\begin{aligned}
Entropia(S) &= -(\tfrac{2}{4} \times \log_2 \tfrac{2}{4}) - (\tfrac{2}{4} \times \log_2 \tfrac{2}{4}) \\
&= -(0.5 \times \log_2 0.5) - (0.5 \times \log_2 0.5) \\
&= -(0.5 \times -1) - (0.5 \times -1) \\
&= -(-0.5) - (-0.5) \\
&= 0.5 + 0.5 \\
&= 1
\end{aligned}
$$

$$
\begin{aligned}
Entropia(A = 1) &= -(\tfrac{2}{2} \times \log_2 \tfrac{2}{2}) - (\tfrac{0}{2} \times \log_2 \tfrac{0}{2}) \\
&= -(1 \times \log_2 1) - \cancel{(0 \times \log_2 \tfrac{0}{2})} \\
&= -(1 \times 0) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
Entropia(A = 2) &= -(\tfrac{0}{2} \times \log_2 \tfrac{0}{2}) - (\tfrac{2}{2} \times \log_2 \tfrac{2}{2}) \\
&= -\cancel{(0 \times \log_2 \tfrac{0}{2})} - (1 \times \log_2 1) \\
&= -(1 \times 0) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
Entropia(B = 1) &= -(\tfrac{1}{2} \times \log_2 \tfrac{1}{2}) - (\tfrac{1}{2} \times \log_2 \tfrac{1}{2}) \\
&= -(0.5 \times \log_2 0.5) - (0.5 \times \log_2 0.5) \\
&= -(0.5 \times -1) - (0.5 \times -1) \\
&= 1
\end{aligned}
$$

$$
\begin{aligned}
Entropia(B = 2) &= -(\tfrac{1}{2} \times \log_2 \tfrac{1}{2}) - (\tfrac{1}{2} \times \log_2 \tfrac{1}{2}) \\
&= -(0.5 \times \log_2 0.5) - (0.5 \times \log_2 0.5) \\
&= -(0.5 \times -1) - (0.5 \times -1) \\
&= 1
\end{aligned}
$$

$$
\begin{aligned}
Entropia(C = 3) &= -(\tfrac{1}{1} \times \log_2 \tfrac{1}{1}) - (\tfrac{0}{1} \times \log_2 \tfrac{0}{1}) \\
&= -(1 \times \log_2 1) - \cancel{(0 \times \log_2 \tfrac{0}{1})} \\
&= -(1 \times 0) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
Entropia(C = 4) &= -(\tfrac{1}{3} \times \log_2 \tfrac{1}{3}) - (\tfrac{2}{3} \times \log_2 \tfrac{2}{3}) \\
&= -(0.333 \times \log_2 0.333) - (0.667 \times \log_2 0.667) \\
&= -(0.333 \times -1.585) - (0.667 \times -0.585) \\
&= 0.918
\end{aligned}
$$

b) Qual o Ganho de Informação para cada um dos atributos?

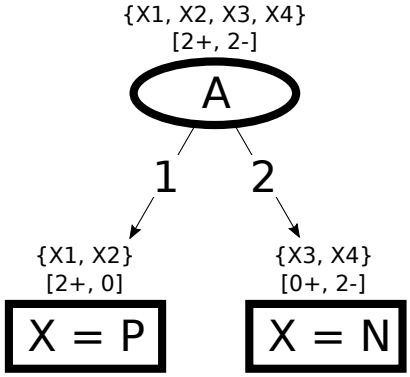
$$\begin{aligned} Ganho(S, A) &= 1 - ((\tfrac{2}{4}) \times 0) - ((\tfrac{2}{4}) \times 0) \\ &= 1 \end{aligned}$$

$$\begin{aligned} Ganho(S, B) &= 1 - ((\tfrac{2}{4}) \times 1) - ((\tfrac{2}{4}) \times 1) \\ &= 0 \end{aligned}$$

$$\begin{aligned} Ganho(S, C) &= 1 - ((\tfrac{1}{4}) \times 0) - ((\tfrac{3}{4}) \times 0.918) \\ &= 0.3115 \end{aligned}$$

c) Face a este resultado, qual seria a árvore de decisão obtida para este conjunto de treinamento, construída de acordo com o critério de maximização do ganho de informação?

Figura 1 – Árvore de decisão



Referências

SAVARIS, A. *Slides de Aula: Classificação: Conceitos básicos e Árvores de Decisão - Parte 1.* [S.l.], 2018. Nenhuma citação no texto.