INE5644 - Data Mining
Exercício - Árvore de decisão - Parte 2

Bruno Marques do Nascimento*

04 de Maio de 2018

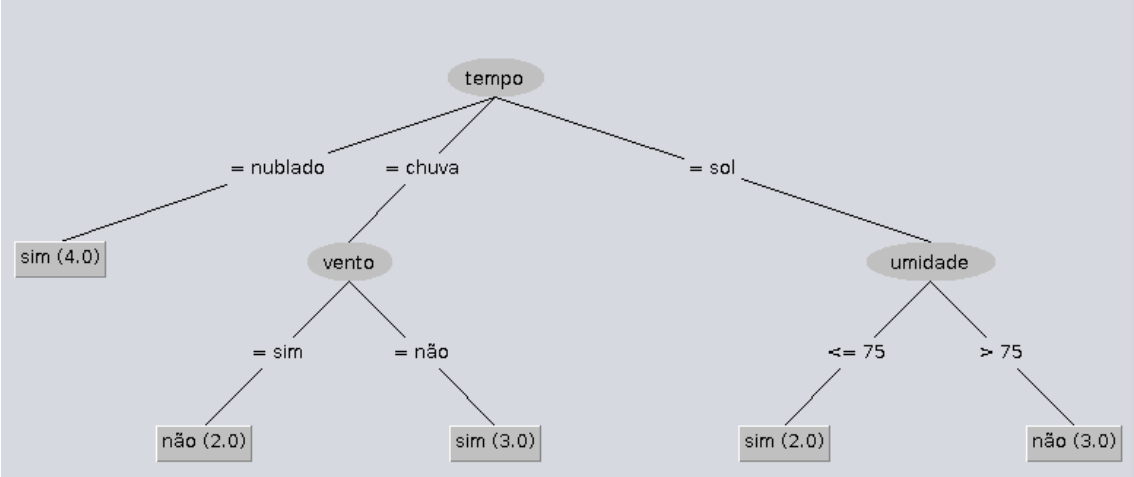
Respostas:

Exercício slide 1

Tabela 1 – Praticar ou não praticar esportes?

tempo	temperatura	umidade	vento	jogar
nublado	64	65	sim	sim
chuva	65	70	sim	não
chuva	68	80	não	sim
sol	69	70	não	sim
chuva	70	96	não	sim
chuva	71	91	sim	não
sol	72	95	não	não
nublado	72	90	sim	sim
chuva	75	80	não	sim
sol	75	70	sim	sim
sol	80	90	sim	não
nublado	81	75	não	sim
nublado	83	86	não	sim
sol	85	85	não	não

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



Comparando a árvore acima (gerada no weka), com a presente nos slides constata-se que ambas são iguais.

Exercício slide 2

Tabela 2 – Exercício slide

Peso	Idade	Sexo	Classe
35	criança	Fem	1
50	criança	Masc	1
60	adulto	Fem	1
70	jovem	Masc	2
75	jovem	Masc	2
80	adulto	Masc	2
85	adulto	Masc	2

Através da ordenação é possível ver a presença de 1 limiar de valor **65**, gerando assim dois intervalos de decisão, os de `Peso` $<$ `65` e `Peso` $\geq$ `65`.

$$\begin{aligned} Entropia(S) &= -(\tfrac{3}{7} \times \log_2 \tfrac{3}{7}) - (\tfrac{4}{7} \times \log_2 \tfrac{4}{7}) \\ &= -(0.43 \times \log_2 0.43) - (0.57 \times \log_2 0.57) \\ &= 0.986 \end{aligned}$$

$$\begin{aligned} Entropia(Peso < 65) &= -(\tfrac{3}{3} \times \log_2 \tfrac{3}{3}) - (\tfrac{0}{3} \times \log_2 \tfrac{0}{3}) \\ &= -(1 \times \log_2 1) - \cancel{(0 \times \log_2 \tfrac{0}{2})} \\ &= -(1 \times 0) \\ &= 0 \end{aligned}$$

$$
\begin{aligned}
Entropia(Peso \geq 65) &= -(\tfrac{0}{4} \times \log_2 \tfrac{0}{4}) - (\tfrac{4}{4} \times \log_2 \tfrac{4}{4}) \\
&= -\cancel{(0 \times \log_2 \tfrac{0}{2})} - (1 \times \log_2 1) \\
&= -(1 \times 0) \\
&= 0 \\
Entropia(Idade = crianca) &= -(\tfrac{2}{2} \times \log_2 \tfrac{2}{2}) - (\tfrac{0}{2} \times \log_2 \tfrac{0}{2}) \\
&= -(1 \times \log_2 1) - \cancel{(0 \times \log_2 \tfrac{0}{2})} \\
&= -(1 \times 0) \\
&= 0
\end{aligned}
$$

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

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

$$
\begin{aligned}
Entropia(Sexo = Masc) &= -(\tfrac{1}{5} \times \log_2 \tfrac{1}{5}) - (\tfrac{4}{5} \times \log_2 \tfrac{4}{5}) \\
&= -(0.2 \times \log_2 0.2) - (0.8 \times \log_2 0.8) \\
&= -(0.2 \times -2.32) - (0.8 \times -0.322) \\
&= 0.722
\end{aligned}
$$

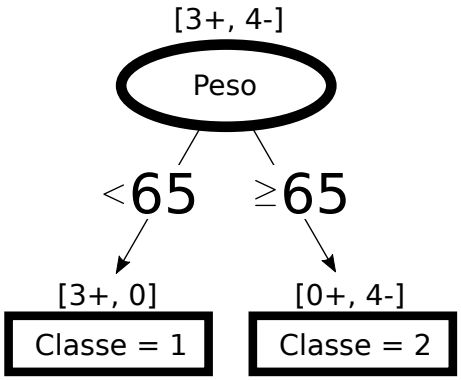
$$
\begin{aligned}
Entropia(Sexo = Fem) &= -(\tfrac{2}{2} \times \log_2 \tfrac{2}{2}) - (\tfrac{0}{2} \times \log_2 \tfrac{0}{2}) \\
&= -(1 \times \log_2 1) - \cancel{(0 \times \log_2 \tfrac{0}{2})} \\
&= -(1 \times 0) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
Ganho(S, Peso) &= 0.986 - ((\tfrac{3}{7}) \times 0) + (\tfrac{4}{7}) \times 0)) \\
&= 0.986
\end{aligned}
$$

$$
\begin{aligned}
Ganho(S, Idade) &= 0.986 - ((\tfrac{2}{7}) \times 0) + (\tfrac{2}{7}) \times 0) + (\tfrac{3}{7}) \times 0.918)) \\
&= 0.986 - 0.393 \\
&= 0.593
\end{aligned}
$$

$$
\begin{aligned}
Ganho(S, Sexo) &= 0.986 - ((\tfrac{5}{7}) \times 0.722) + (\tfrac{2}{7}) \times 0)) \\
&= 0.986 - 0.516 \\
&= 0.47
\end{aligned}
$$

Com os ganhos calculados, gera-se a árvore.



Exercício folha

a) Qual o objetivo de se executar poda (*prunning*) em árvores de decisão?

Existem dois objetivos principais, maximizar o desempenho da árvore de decisão em questão sem perder seu poder de decisão e reduzir seu tamanho físico, ou seja, tornar a árvore de decisão mais leve, irá ocupar menos espaço em disco. Um outro objetivo é evitar que os erros e ruídos presentes em ramificações muitos específicas da árvore atrapalhe na decisão tomada.

b) Quais tipos de poda (*prunning*) estão previstos na literatura?

Na literatura encontram-se diversos tipo de poda para árvore de decisão. Eles estão classificados em dois grandes grupos, os métodos `pré-poda` e `pós-poda`. Os métodos pré-poda são realizado durante a construção da árvore de decisão, ou seja, na sua construção um nodo pode parar de ser ramificado e transformado em nodo folha se os critérios para isto forem satisfeito. Já a pós-poda acontece após a construção completa da árvore, aonde a ramificação abaixo de um nodo é removida e este nodo passa a ser um nodo folha, representando a classe de maior representatividade na ramificação removida. Alguns métodos conhecidos que se destacam, são: *Cost Complexity Pruning*, *Reduced Error Pruning*, *Minimum Error Pruning (MEP)*, *Pessimistic Pruning*, *ErrorBased Pruning (EBP)*, *Minimum Description Length (MDL) Pruning*, *Mininum Message Length (MML) Pruning*, *Critical Value Pruning (CVP)*, *OPT* e *OPT-2*, conforme Zuben e Attux (2010).

c) Para o *dataset* de exemplo, qual o impacto da poda (*prunning*) nas árvores resultado? Quais são as diferenças perceptíveis entre as árvores geradas?

O impacto gerado pela poda foi extremamente significativo, a árvore gerada com o algoritmo sem utilizar poda tinha um tamanho de 7976 nodos com 6812 nodos folha, após a utilização do algoritmo a árvore passou a ter um tamanho de 710 com 564 nodos folha, uma redução no tamanho da árvore e na quantidade de nodos folha de 91%. A principal diferença percebida, é a ausência de ramificações existentes a partir de certos nodo, que é exatamente

o que a poda realiza. Na Figura 1, podemos observar de maneira clara a poda que acontece no nodo $capital\text{-}gain \leq 3942$.

Figura 1 – Captura de tela

Referências

SAVARIS, A. *Aula 20 - Classificação - Conceitos básicos e Árvores de Decisão - Parte 2*. [S.l.], 2018. Acesso em: 03 maio 2018. Nenhuma citação no texto.

SAVARIS, A. *Aula 20 - Classificação - Conceitos básicos e Árvores de Decisão - Parte 2 - Prática*. [S.l.], 2018. Acesso em: 24 abril 2018. Nenhuma citação no texto.

WITTEN, I. H. et al. *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. 4th. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016. ISBN 0128042915, 9780128042915. Nenhuma citação no texto.

ZUBEN, F. J. V.; ATTUX, R. R. F. *Notas de Aula - Árvores de Decisão*. [S.l.], 2010. Disponível em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf>. Acesso em: 03 maio 2018. Citado na página 4.