Homework 4 Corrections

November 7, 2021

Jose Carlos Munoz

3.10)
Each Node is describe by the attribue number, Number of child and splitting value.To find the cost function of both of the Decision tree we use this formula.

$$|h| = 2 * [\log_2 N] - 1 + 2 * [\log_2 A + C] - 1 + 2 * [\log_2 n + BD] - 1 \tag{1}$$

Where A is the number of Attributes, C is the number of classes and n is the unkown Sample size, N is the number of nodes, and BD is the branching degree. For both trees the number of attributes is 16 and 3 for the number of classes
Now we find the cost function for decision Tree A, where Nodes = 5, and a BD = 2

$$\begin{aligned} |h| &= 2 * [\log_2 5] - 1 + 2 * [\log_2 16 + 3] - 1 + 2 * [\log_2 n + 2] - 1 \\ |h| &= 2 * 3 - 1 + 2 * 5 - 1 + 2 * [\log_2 n + 2] - 1 \\ |h| &= 6 - 1 + 10 - 1 + 2 * [\log_2 n + 2] - 1 \\ |h| &= 13 + 2 * [\log_2 n + 2] \end{aligned} \tag{a}$$

Now we find the cost function for decision Tree b, where N = 9,BD = 3.

$$\begin{aligned} |h| &= 2 * [\log_2 9] - 1 + 2 * [\log_2 16 + 3] - 1 + 2 * [\log_2 n + 3] - 1 \\ |h| &= 2 * 4 - 1 + 2 * 5 - 1 + 2 * [\log_2 n + 3] - 1 \\ |h| &= 8 - 1 + 10 - 1 + 2 * [\log_2 n + 3] - 1 \\ |h| &= 15 + 2 * [\log_2 n + 3] \end{aligned} \tag{b}$$

Using the MDL paradigm we need to find a $L_S(h) + \sqrt{\frac{\log_2\left(\frac{2}{\delta}\right)+|h|}{2n}}$. The better decision tree is the one that gives us the lowest value. m is found as the sample size which is 200, δ is given as .99. and the $L_S(h)$ is the error devided by the sample size of the Decision tree. $|h|$ is the encoding length in which we solved above.
For Decision tree A we get

$$\begin{aligned} |h| &= 13 + 2 * [\log_2 200 + 3] \\ |h| &= 13 + 2 * 8 \\ |h| &= 29 \\ L_S(h) &= \frac{7}{200} \\ \delta &= 0.99 \\ &= \frac{7}{200} + \sqrt{\frac{\log_2\left(\frac{2}{.99}\right) + 29}{2 * 200}} \\ &= \frac{7}{200} + \sqrt{\frac{1.01449 + 29}{2 * 200}} \\ &= 0.293927 \end{aligned} \tag{a}$$

For Decision tree B we get

$$
\begin{aligned}
|h| &= 15 + 2 * [\log_2 200 + 3] \\
|h| &= 15 + 2 * 8 \\
|h| &= 31 \\
L_S(h) &= \frac{4}{200} \\
\delta &= 0.99 \\
&= \frac{4}{200} + \sqrt{\frac{\log_2\left(\frac{2}{.99}\right) + 31}{2 * 200}} \\
&= \frac{4}{200} + \sqrt{\frac{1.01449 + 31}{2 * 200}} \\
&= 0.3029067
\end{aligned}
\tag{b}
$$

From the results we can conclude that Decision Tree A is the best one of the two.