

Student Number

Student Name

Date

CS5483 tutorial question 11

Apply C4.5 algorithm to construct a decision tree for purchasing records from the following data after dividing the tuples into two groups according to “age”: one is less than 25, and another is greater than or equal to 25. Show all the steps and calculation for the construction.

Location	Customer Sex	Age	Purchase records
Asia	Male	15	Yes
Asia	Female	23	No
America	Female	20	No
Europe	Male	18	No
Europe	Female	10	No
Asia	Female	40	Yes
Europe	Male	33	Yes
Asia	Male	24	Yes
America	Male	25	Yes
Asia	Female	27	Yes
America	Female	15	Yes
Europe	Male	19	No
Europe	Female	33	No
Asia	Female	35	No
Europe	Male	14	Yes
Asia	Male	29	Yes
America	Male	30	No

Model answer for Question 11

Let S = any set of training case

Let |S| = number of classes in set S

Let Freq (Ci, S) = number of cases in S that belong to class Ci

Info(S) = average amount of information needed to identify the class in S

Gain (X) = information gained by partitioning S according to the test on attribute X

$Info(S) = -\sum \text{freq}(Ci, S) / |S| \times \log_2(\text{freq}(Ci, S) / |S|)$

$Info_X(S) = -\sum |Si| / |S| \times info(Si)$

$Gain(X) = Info(S) - Info_X(S)$

Let S be the training set

$$Info(S) = \frac{-9}{17} \log_2\left(\frac{9}{17}\right) - \frac{8}{17} \log_2\left(\frac{8}{17}\right) = 0.485 + 0.51 = 0.995$$

Case 1: Split table according to different Age

$$Info_{Age}(S) = \frac{9}{17} \left(-\frac{4}{9} \log_2\left(\frac{4}{9}\right) - \frac{5}{9} \log_2\left(\frac{5}{9}\right) \right)$$

$$+ \frac{8}{17} \left(-\frac{5}{8} \log_2\left(\frac{5}{8}\right) - \frac{3}{8} \log_2\left(\frac{3}{8}\right) \right) = 0.274 + 0.248 + 0.199 + 0.248 = 0.969$$

$$Gain(Age) = Info(S) - Info_{Age}(S) = 0.995 - 0.969 = 0.026$$

Case 2: Split table according to different Sex

$$\text{Info}_{\text{Sex}}(\text{S}) = \frac{9}{17} \left(-\frac{6}{9} \text{Log}_2\left(\frac{6}{9}\right) - \frac{3}{9} \text{Log}_2\left(\frac{3}{9}\right) \right)$$

$$+ \frac{8}{17} \left(-\frac{3}{8} \text{Log}_2\left(\frac{3}{8}\right) - \frac{5}{8} \text{Log}_2\left(\frac{5}{8}\right) \right) = 0.206 + 0.277 + 0.248 + 0.199 = 0.93$$

$$\text{Gain}(\text{Sex}) = \text{Info}(\text{S}) - \text{Info}_{\text{Sex}}(\text{S}) = 0.995 - 0.93 = 0.065$$

Case 3: Split table according to Location

$$\text{Info}_{\text{Location}}(\text{S}) = \frac{7}{17} \left(-\frac{5}{7} \text{Log}_2\left(\frac{5}{7}\right) - \frac{2}{7} \text{Log}_2\left(\frac{2}{7}\right) \right)$$

$$+ \frac{4}{17} \left(-\frac{2}{4} \text{Log}_2\left(\frac{2}{4}\right) - \frac{2}{4} \text{Log}_2\left(\frac{2}{4}\right) \right)$$

$$+ \frac{6}{17} \left(-\frac{2}{6} \text{Log}_2\left(\frac{2}{6}\right) - \frac{4}{6} \text{Log}_2\left(\frac{4}{6}\right) \right) = 0.136 + 0.211 + 0.235 + 0.185 + 0.137 = 0.904$$

$$\text{Gain}(\text{Location}) = \text{Info}(\text{S}) - \text{Info}_{\text{Location}}(\text{S}) = 0.995 - 0.904 = 0.091$$

As a result, information gain of Location is greater than information gain of Sex and information gain of Age, and so we must split according to Location

As a result, we split on attribute Location, and the tree will be as follows:

