Data Mining (CSE542)

Homework 09

ID: ____ Name: _조원석__ Date: _2023-06-04_

Task-1: Given Table 19.3, construct a decision tree. Use information gain as the split point evaluation measure. Next, classify the point (Age=27, Car=Vintage).

Table 19.3. Data : Age is numeric and Car is categorical. Risk gives the class label for each point: high (H) or low (L)

	Age	Car	**Risk**
$\mathbf{x}_1^T$	25	Sports	L
$\mathbf{x}_2^T$	20	Vintage	H
$\mathbf{x}_3^T$	25	Sports	L
$\mathbf{x}_4^T$	45	SUV	H
$\mathbf{x}_5^T$	20	Sports	H
$\mathbf{x}_6^T$	25	SUV	H

$$P_L = \frac{2}{6} = \frac{1}{3},\ H(D) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = -(0.390 - 0.528) = 0.918$$

Age ≤ 22.5 & Age ≤ 35 : they are closen to be the mid-points between the distinct values, namely, 20, 25 and 45, that we observe for Age.

(a) Age ≤ 22.5

D_L includes only the points x_2 and x_5, whereas D_R comprises the remaining points: $x_1, x_3, x_4, and\ x_6$. For D_L this yields P_L = 0 and P_H=1, whereas for D_R we have $P_L = \frac{2}{4}\ and\ P_H = \frac{2}{4}$
The weighted entropy is then

$$H(D_L, D_R) = \frac{2}{6}H(D_L) + \frac{4}{6}H(D_R) = -\frac{2}{6}(0) - \frac{4}{6}\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = \left(-\frac{2}{3}log_2\frac{1}{2}\right) = -0.67$$

This yields an information gain of 0.918 – 0.67 = 0.248

(b) In a similar manner we can compute the weighted entropy for Age ≤ 35.
For $D_R = \{x_4\}$and$D_L has\ the\ remaining\ points.$
So that H(D_L) $= \frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5} = 0.971$ and H(D_R) $= 0$
The split entropy is then H(D_L, D_R) $= \frac{5}{6}(0.971) = 0.809$
The information gain is 0.915-0.809 = 0.106, which is not as high as for Age <= 22.5

Next We evaluate all possible splits for Car. Nogte that categorical data, in general, yields $(2^{v} - 1)/2$ possible splits, where v is the set of possible values for the attribute.
This can be reduced to O(v) by using a greedy split selection approach. For Car the possible values are { Sports(S), Vintage(V), SUV(U)}, which yields the following three distinct splits:

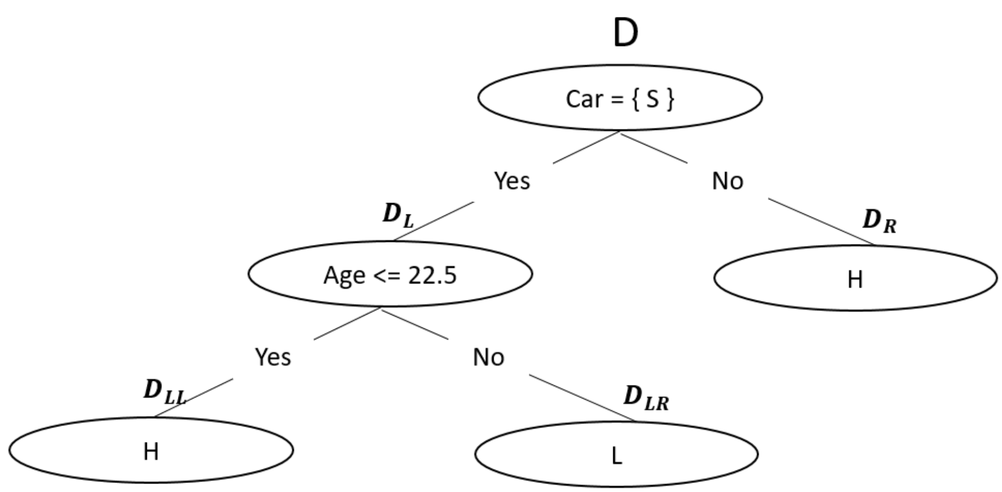
Car $\in$	**Car $\notin$**
{ S }	**{ V, U }**
{ V }	**{S, U }**
{ U }	**{ S, V }**

Note that the split Car $\in$ { V, U} is essentially the same as the split Car $\in$ { S }, the only difference being that the decision has being that the decision has been "reversed". It is therefore not a distinct split, and we do not consider such splits. Next we evaluate the three categorical; splits as follows:

(a) For the split Car $\in$ { S }, D_L = { x_1 , x_3 , x_5 }, and D_R ={ x_2 , x_4 , x_6 }. For D_L , this yields P_L = 2/3 and P_H = 1/3 , and for D_R and for P_L = 0 and P_H = 1. The weighted entropy of the split is then $H(D_L, D_R)=(3/6)H(D_H)+(3/6)H(D_R)= -(3/6)((1/3)\log_2(1/3)+(2/3)\log_2 (2/3)-(3/6)(0)=0.459$

(b) For Car $\in$ { V }, we get the same information gain as for Age <= 35, i.e., 0.106

(c) For Car $\in$ { U }, the gain is the same as for Age <= 22.5, i.e., 0.248.
Among all the possible split points for both Age and Car, the one with the highest information gain is Car $\in$ { S }, which is chosen as the best split decision at the root of the decision tree, as show in below that.



We therefore make this split and recursively call the decision tree algorithm on each new subset

D_L = { x_1 , x_3 , x_5 }, and D_R ={ x_2 , x_4 , x_6 }.

Notice that for D_R all points are already labeled as high risk (H). Since the partition is already pure, we make it a leaf node, labeled as H. On the other hand, D_L is not completely pure, so we consider partitioning it further. Since all points in D_L have Car $\in$ { S }, we cannot use Car to further distinguish the points. Further, for Age, Age <= 22.5 is the only possible split to consider. Note that the entropy of D_L is given as

$$H(D_L) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.918$$

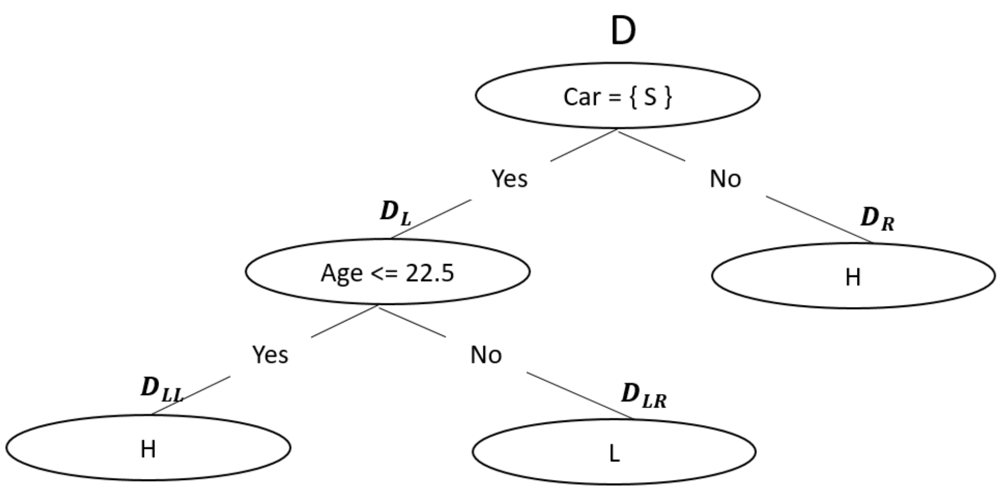
For Age <= 22.5, D_{LL} = { x_1 , x_3 }, whereas D_{LR} = { x_5 }.

For D_{LL} we get P_L= 1 and P_H= 0, and for D_{LR} , we get P_L = 0 and P_H = 1.

The weighted entropy is then

$$H(D_{LL} = \{\,1,3\,\}, D_{LR} = \{\,5\,\}) = \frac{2}{3}H(D_{LL}) + \frac{1}{3}H(D_{LR}) = -\frac{2}{3}(0) - \frac{1}{3}(0) = 0$$

Thus the information gain is 0.918 – 0 = 0.918. In this example, this is the only possible split decision. After D_L is split, we obtain the two new leaves D_{LL} , which is labeled as low-risk(L), and D_{LR} , which is labeled as high-risk(H). The full decision tree is shown in below



One of the advantages of decision trees, is that each path from the root to a leaf can be written as a rule. For our example tree above, we obtain the following three rules

1) **R_1 : if Car $\in$ { S } and Age <= 22.5, then Rist = H**
2) **R_2 : if Car $\in$ { S } and Age > 22.5, then Rist = L**
3) **R_3 : if Car $\notin$ { S } , then Rist = H**

This is one of the strengths of decision trees, namely the ability to aid understanding of the model via simple rules presented to the user. Once a decision tree model has been built, it can be used to classify new points. For example, for the test point Age =27, and Car =Vintage, we can classify the point by applying the set of decisions starting at the root. First we check whether Car $\in$\{S\}. Since this test will be false, we go to the right branch, and since it is a leaf, we predict the class to be H

Task-2: Given the dataset in Table 19.4. Show which decision will be chosen at the root of the decision tree using information gain, Gini index, and CART, measures.

Table 19.4.

Instance	a_1	a_2	a_3	Class
1	T	T	5.0	Y
2	T	T	7.0	Y
3	T	F	8.0	N
4	F	F	3.0	Y
5	F	T	7.0	N
6	F	T	4.0	N
7	F	F	5.0	N
8	T	F	6.0	Y
9	F	T	1.0	N

The entropy for the whole dataset is given as

$$H(\mathbf{D}) = \frac{4}{9} log_{10} \frac{4}{9} - \frac{5}{9} log_{10} \frac{5}{9} = 0.2983$$

The Gini index for the whole dataset is:

$$G(\mathbf{D}) = 1 - (\frac{4^2}{9^2} + \frac{5^2}{9^2}) = 0.4938$$

Consider the split for attribute a_1 ,which has only one possible split, namely a_1 $\in$ T

The split entropy is given as:

$$H(D_L, D_R) = \frac{4}{9}\left(-\frac{1}{4} log_{10} \frac{1}{4} - \frac{3}{4} log_{10} \frac{3}{4}\right) + \frac{5}{9}\left(-\frac{1}{5} log_{10} \frac{1}{5} - \frac{4}{5} log_{10} \frac{4}{5}\right) = 0.2293$$

Thus the gain is 0.2983 – 0.2293 – 0.0690

The gini for the split is

$$G(D_L, D_R) = \frac{4}{9}\left(1 - \frac{1^2}{4^2} - \frac{3^2}{4^2}\right) + \frac{5}{9}\left(1 - \frac{1^2}{5^2} - \frac{4^2}{5^2}\right) = 0.3444$$

The CART measure for the split is given as:

$$\boldsymbol{CART(D_L, D_R) = 2 * \frac{4}{9} * \frac{5}{9} * \left(\left|\frac{3}{4} - \frac{1}{5}\right| + \left|\frac{1}{4} - \frac{4}{5}\right|\right) = 0.5432}$$

Likewise for attribute a$_2$, we have only one split, and the split entropy is:

$$\boldsymbol{H(D_L, D_R) = \frac{5}{9}\left(-\frac{2}{5}log_{10}\frac{2}{5} - \frac{3}{5}log_{10}\frac{3}{5}\right) + \frac{4}{9}\left(-\frac{1}{2}log_{10}\frac{1}{2} - \frac{1}{2}log_{10}\frac{1}{2}\right) = 0.2962}$$

With gain 0.2983 – 0.2962 = 0.0021.

The gini for the split is

$$\boldsymbol{G(D_L, D_R) = \frac{4}{9}\left(1 - \frac{1^2}{2^2} - \frac{1^2}{2^2}\right) + \frac{5}{9}\left(1 - \frac{2^2}{5^2} - \frac{3^2}{5^2}\right) = 0.4889}$$

The CART measure for the split is given as:

$$\boldsymbol{CART(D_L, D_R) = 2 * \frac{4}{9} * \frac{5}{9} * \left(\left|\frac{1}{2} - \frac{2}{5}\right| + \left|\frac{1}{2} - \frac{3}{5}\right|\right) = 0.0988}$$

For attribute a$_3$ there are several numeric split points, namely a$_3$ < 3.0, a$_3$ <4.0, a$_3$ < 5.0, a$_3$ < 6.0, a$_3$ <7.0, a$_3$ < 8.0, The split entropy for each of these cases is as follows:

(a) For a$_3$ < 3.0 we have

$$\boldsymbol{H(D_L, D_R) = 0 + \frac{8}{9}\left(-\frac{1}{2}log_{10}\frac{1}{2} - \frac{1}{2}log_{10}\frac{1}{2}\right) = 0.2676}$$

The gain is 0.2983 – 0.2676 = 0.0307.

The gini for the split is

$$\boldsymbol{G(D_L, D_R) = \frac{8}{9}\left(1 - \frac{1^2}{2} - \frac{1^2}{2}\right) = 0.4444}$$

The CART measure for the split is given as:

$$\boldsymbol{CART(D_L, D_R) = 2 * \frac{1}{9} * \frac{8}{9} * \left(\left|1 - \frac{1}{2}\right| + \left|0 - \frac{1}{2}\right|\right) = 0.1975}$$

(b) For a_3 < 4.0 we have

$$\boldsymbol{H(D_L, D_R) = \frac{2}{9}\left(-2 * \frac{1}{2} log_{10}\frac{1}{2}\right) + \frac{7}{9}\left(-\frac{3}{7} log_{10}\frac{3}{7} - \frac{4}{7} log_{10}\frac{4}{7}\right) = 0.2976}$$

The gain is 0.2983 – 0.2976 = 0.0007.
The gini for the split is

$$\boldsymbol{G(D_L, D_R) = \frac{2}{9}\left(1 - \frac{1^2}{2^2} - \frac{1^2}{2^2}\right) = 0.4921}$$

The CART measure for the split is given as:

$$\boldsymbol{CART(D_L, D_R) = 2 * \frac{2}{9} * \frac{7}{9} * \left(\left|\frac{1}{2} - \frac{3}{7}\right| + \left|\frac{1}{2} - \frac{4}{7}\right|\right) = 0.0494}$$

(c) For a_3 < 5.0 we have

$$\boldsymbol{H(D_L, D_R) = \frac{3}{9}\left(-\frac{3}{9} log_{10}\frac{3}{9} - \frac{6}{9} log_{10}\frac{6}{9}\right) + \frac{6}{9}\left(-2 * \frac{1}{2} log_{10}\frac{1}{2}\right) = 0.2928}$$

The gain is 0.2983 – 0.2928 = 0.0055.
The gini for the split is

$$\boldsymbol{G(D_L, D_R) = \frac{1}{3}\left(1 - \frac{1^2}{3^2} - \frac{2^2}{3^2}\right) + \frac{2}{3}\left(1 - \frac{1^2}{2^2} - \frac{1^2}{2^2}\right) = 0.4815}$$

The CART measure for the split is given as:

$$\boldsymbol{CART(D_L, D_R) = 2 * \frac{1}{3} * \frac{2}{3} * \left(\left|\frac{1}{3} - \frac{1}{2}\right| + \left|\frac{2}{3} - \frac{1}{2}\right|\right) = 0.1481}$$

(d) For a_3 < 6.0 we have

$$\boldsymbol{H(D_L, D_R) = \frac{5}{9}\left(-\frac{2}{5} log_{10}\frac{2}{5} - \frac{3}{5} log_{10}\frac{3}{5}\right) + \frac{4}{9}\left(-2 * \frac{1}{2} log_{10}\frac{1}{2}\right) = 0.2962}$$

The gain is 0.2983 – 0.2962 = 0.0021.
The gini for the split is

$$\boldsymbol{G(D_L, D_R) = \frac{5}{9}\left(1 - \frac{2^2}{5^2} - \frac{3^2}{5^2}\right) + \frac{4}{9}\left(1 - \frac{1^2}{2^2} - \frac{1^2}{2^2}\right) = 0.4889}$$

The CART measure for the split is given as:

$$\boldsymbol{CART(D_L, D_R) = 2 * \frac{5}{9} * \frac{4}{9} * \left(\left|\frac{1}{2} - \frac{2}{5}\right| + \left|\frac{1}{2} - \frac{3}{5}\right|\right) = 0.0988}$$

(e) For $a_3 < 7.0$ we have

$$H(D_L, D_R) = \frac{6}{9}\left(-2 * \frac{1}{2} log_{10}\frac{1}{2}\right) + \frac{3}{9}\left(-\frac{3}{9} log_{10}\frac{3}{9} - \frac{6}{9} log_{10}\frac{6}{9}\right) = 0.2928$$

The gain is 0.2983 – 0.2928 = 0.0055.
The gini for the split is

$$G(D_L, D_R) = \frac{6}{9}\left(1 - \frac{1^2}{2^2} - \frac{1^2}{2^2}\right) + \frac{3}{9}\left(1 - \frac{1^2}{3^2} - \frac{2^2}{3^2}\right) = 0.4815$$

The CART measure for the split is given as:

$$CART(D_L, D_R) = 2 * \frac{2}{3} * \frac{1}{3} * \left(\left|\frac{1}{3} - \frac{1}{2}\right| + \left|\frac{2}{3} - \frac{1}{2}\right|\right) = 0.1481$$

(f) For $a_3 < 8.0$ we have

$$H(D_L, D_R) = \frac{8}{9}\left(-\frac{1}{2} log_{10}\frac{1}{2} - \frac{1}{2} log_{10}\frac{1}{2}\right) + 0 = 0.2976$$

The gain is 0.2983 – 0.2976 = 0.0307.
The gini for the split is

$$G(D_L, D_R) = \frac{8}{9}\left(1 - \frac{1^2}{2^2} - \frac{1^2}{2^2}\right) = 0.4444$$

The CART measure for the split is given as:

$$CART(D_L, D_R) = 2 * \frac{1}{9} * \frac{8}{9} * \left(\left|0 - \frac{1}{2}\right| + \left|1 - \frac{1}{2}\right|\right) = 0.1975$$

So the best split for all three measures is $a_1 \in \{ T \}$

It has the highest gain (0.069), the lowest Gini value (0.3444), and the highest CART measure (0.5432).