Machine Learning Homework 2 :

1) Given data :

A	B	C	Y
0	1	0	Yes
→ 1	0	1	Yes
0	0	0	No
→ 1	0	1	No
0	1	1	No
1	1	0	Yes

a) ⟹ has two instances with same values for the attributes A, B, C but giving different values of outcome Y (Yes & No).

For this reason, a decision tree having 100% accuracy on this training set <u>cannot</u> be drawn.

⟹ Because no form of DT would satisfy both the outcomes in the same structure.

b) A (3+, 3−): 0 → (1+, 2−), 1 → (2+, 1−)

B (3+, 3−): 0 → (1+, 2−), 1 → (2+, 1−)

C (3+, 3−): 0 → (2+, 1−), 1 → (1+, 2−)

$$\text{Entropy}(S) = -P_{\oplus}\log_2 P_{\oplus} - P_{\ominus}\log_2 P_{\ominus}$$

$$= -\frac{3}{6}\log_2 \tfrac{1}{2} - \frac{1}{2}\log_2 \tfrac{1}{2} = 1$$

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$\text{Gain}(S,A) = 1 - \left[\frac{3}{6}E(1,2) + \frac{3}{6}E(2,1)\right] = 0.545$$

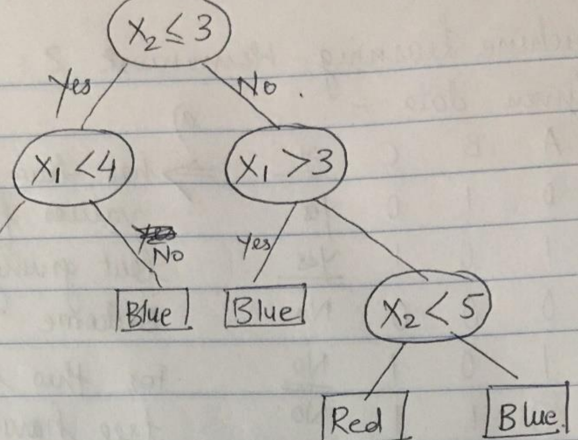
$$\text{Gain}(S,B) = 1 - \left[\frac{3}{6}E(1,2) + \frac{3}{6}E(2,1)\right] = 0.545$$

$$\text{Gain}(S,C) = 1 - \left[\frac{3}{6}E(2,1) + \frac{3}{6}E(1,2)\right] = 0.545$$

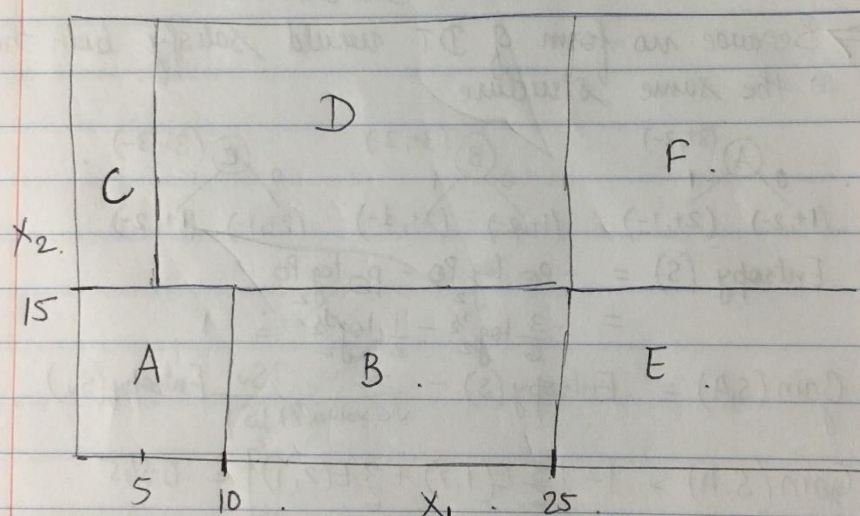
So, A, B, C have <u>same</u> information gain.

Each attribute of A, B & C splits the given training set of 3^+ & 3^- labels into 2 categories which give (1^+ & 2^-) & (2^+ & 1^-) labels. The split for any attribute gives an equal additional knowledge for outcome. That's reflected in the information gain values.

Q.2). Draw the decision tree :



Q.3)



Q.4) In the given data :-

2 instances → # 2 & 3

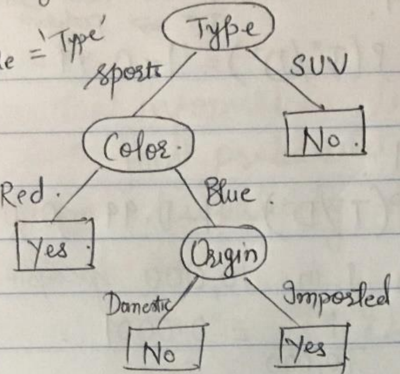
Red ~~Sports~~ Domestic
Red ~~sports~~ Domestic } give a Yes & No as the Buy answer

6 & 7

Blue SUV Imported No
Blue SUV Imported Yes } give a No & Yes for the ~~same set~~ of attribute values.

So if we consider accuracy as the metric of splitting, the best decision tree we can make is

Case I) Root node = 'Type'

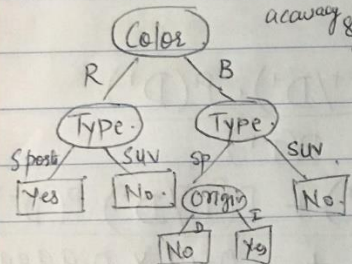


Accuracy result → 8/10.

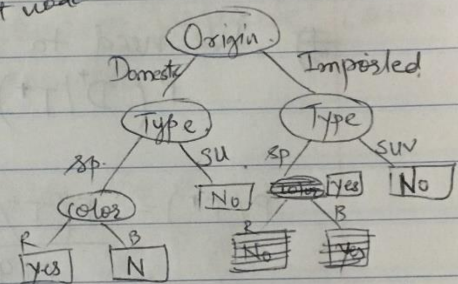
Also, the constructed tree is not generalized. For testing data of form. Red Domestic SUV, no data is available

Case II) Root node = 'Color'.

accuracy 8/10



or Case III root node = 'Origin'.



Overfitting in a decision tree is when the designed tree perfectly fits all (majority) of the training samples. This would affect the accuracy in predicting on non-training samples.

For building the given Decision Tree, for maximum accuracy, we made use of all the data pts. Thus, it ends up with branches with strict rules of sparse data. This affects the accuracy when predicting samples that are not part of the training set.

One of the methods used to address over fitting in decision tree is called pruning. In pruning, we trim off the branches of the tree, such that the overall accuracy is not disturbed.

5) Given probabilities :-

$T^+ \rightarrow$ Test comes +ve.
$D^+ \Rightarrow$ patient has disease
$T^- \rightarrow$ Test comes -ve.
$D^- \Rightarrow$ patient doesn't have disease

a) $P(T^+/D^+) = 0.99.$

$P(T^-/D^+) = 1 - P(T^+/D^+) = 1 - 0.99 = 0.01.$

b) $P(T^-/D^-) = 0.99.$

$P(T^+/D^-) = 1 - P(T^-/D^-) = 1 - 0.99 = 0.01.$

c) disease strikes 1 in 10,000 people.

$\therefore \; P(D^+) = \dfrac{1}{10,000} = 0.0001.$

$P(D^-) = \dfrac{9999}{10,000} = 0.9999.$

We need to compute :

$$P(D^+/T^+) = \frac{P(T^+/D^+) \cdot P(D^+)}{P(T^+)}$$

$$P(T^+) = P(T^+/D^+) \times P(D^+) + P(T^+/D^-) \cdot P(D^-)$$

$$= 0.99 \times 0.0001 + 0.01 \times 0.9999$$

$$= 0.000099 + 0.009999.$$

$$= 0.010098$$

So, $P(D^+/T^+) = \dfrac{0.99 \times 0.0001}{0.010098}$

$$\boxed{P(D^+/T^+) = 0.00980}$$

). ~~Joint PMF & P(X,Y)~~

6). The mutual information between random variables X & Y with joint probability mass function $p(x, y)$ & marginal probability mass functions $p(x)$ & $p(y)$. is defined as :-

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x) \, P(y)}.$$

The mutual information is a measure of the amnt of information that one random variable contains about another random variable.

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x) \, P(y)}.$$

$$= \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x/y)}{P(x)}.$$

$$= -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log P(x) + \sum_{x \in X} \sum_{y \in Y} P(x, y) \log P(x/y).$$

$$= H(X) - H(X|Y).$$

or symmetrically. $H(Y) - H(Y|X)$.

X says as much about Y as Y says about X.

Thus, $\boxed{I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)}$