Decision Tree Algorithm

THEORY

1) It is a machine learning algorithm under supervised learning and classification tasks. It is a special case of the random forest algorithm.
2) Entropy is the measure of unpredictability in the dataset given as:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

3) Information gain is the measure of decrease in entropy after splitting. Our aim is to minimize the final entropy by moving in the direction of highest information gain at each split.

$$IG(S, a) = H(S) - H(S|a) \tag{2}$$

4) Gini index is a measure of impurity or purity of a distribution given as:

$$Gini = 1 - \sum_{j} p_j^2 \tag{3}$$

QUIZ

1) Calculate entropy for given dataset with 2 classes, A = $\frac{1}{10}$ & B = $\frac{9}{10}$. Ans: 0.469
2) Calculate information gain if the above dataset is splitted as,
   a) Split 1: $\frac{1}{4}, \frac{3}{4}$
   b) Split 2: $\frac{3}{6}, \frac{3}{6}$

   Solution:
   Entropy(Split 1) = $\frac{1}{4}\log_2(4) + \frac{3}{4}\log_2(\frac{4}{3}) = 0.811$

   Entropy(Split 2) = $\frac{1}{2}\log_2(2) + \frac{1}{2}\log_2(2) = 1$

   Information Gain = 0.469 - $\frac{4}{10} \times 0.811 + \frac{6}{10}$ = -0.455

3) Which distribution is more preferred:
   a) Gini index = 0.477
   b) Gini index = 0.232