1. Decision Tree (15 points)

a. Assume that Position is chosen for the root of the decision tree. What is the information gain associated with this attribute?

$$IG(S, Position) = H(S) - \sum_{Position} P(S|Position)H(S|Position)$$

$$H(S) = -P(1)\log(P(1)) - P(0)\log(P(0)) = -\frac{5}{10}\log\left(\frac{5}{10}\right) - \frac{5}{10}\log\left(\frac{5}{10}\right) = 1$$

$$H(S|Position = Top) = -P(T|Position = Top)\log(P(T|Position = Top)) - P(F|Position = Top)\log(P(F|Position = Top)) = -0\log(0) - 1\log(1) = 0$$

$$H(S|Position = Middlel) = -P(T|Position = Middlel)\log(P(T|Position = Middlel)) - P(F|Position = Middlel)\log(P(F|Position = Middlel)) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$H(S|Position = Bottom) = -P(T|Position = Bottom)\log(P(T|Position = Bottom)) - P(F|Position = Bottom)\log(P(F|Position = Bottom)) = -1\log(1) - 0\log(0) = 0$$

$$IG = 1 - \frac{3}{10}\cdot 0 - \frac{4}{10}\cdot 1 - \frac{3}{10}\cdot 0 = 0.6$$

b. Draw the full decision tree learned from this data (without any pruning)

If we examine the information when the tree splits based on the position where position = top and all clicked values are false, this split will lead us straight to FALSE. Additionally, when the tree splits based on the position where position = bottom and all clicked values are true, this split will lead us straight to True.

$$H(S) = -P(1)\log(P(1)) - P(0)\log(P(0)) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$$

$$H(S|Size = Small) = -P(T|size = Small)\log(P(T|Size = Small)) - P(F|Size = Small)\log(P(F|Size = Small)) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$H(S|Size = Big) = -P(T|size = Big)\log(P(T|Size = Big)) - P(F|Size = Big)\log(P(F|Size = Big)) = 0$$

$$IG(S, Size) = 1 - 1\cdot 1 - 0 = 0$$

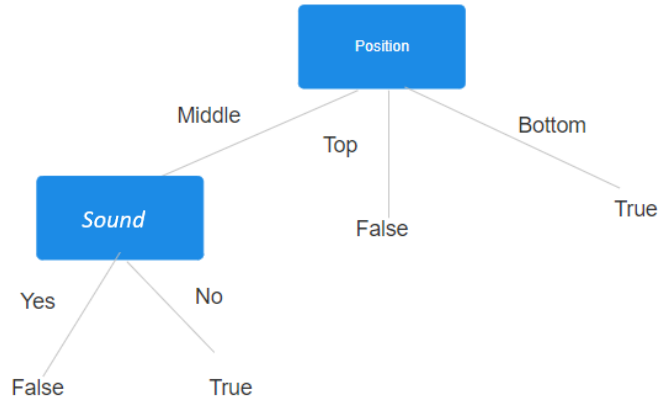
$$H(S|Sound = Yes) = -P(T|sound = Yes)\log(P(T|Sound = Yes)) - P(F|Sound = Yes)\log(P(F|Sound = Yes)) = -0\log(0) - 1\log(1) = 0$$

$$H(S|Sound = No) = -P(T|sound = No)\log(P(T|Sound = No)) - P(F|Sound = No)\log(P(F|Sound = No)) = -1\log(1) - 0\log(0) = 0$$

$$IG(S, Sound) = 1 - \frac{1}{2}\cdot 0 - \frac{1}{2}\cdot 0 = 1$$

We will split the decision tree based on the feature 'sound' because the Gini Index (GI) is higher.

When examining the information, when the tree splits based on the condition position = middle and Sound = Yes , all clicked values are false, so this split will lead us straight to FALSE. Similarly, when the tree splits based on the condition position = middle and Sound = No, all clicked values are true, so this split will lead us straight to TRUE.



2. Naive Base

For the same data Using Naïve Base what is the prediction of the new Sample (big,Middle,No) .

$P(T|Size = big, Position = Middle, Sound = No) =$

$$\frac{P(T)\cdot P(size = big|T)\cdot P(Position = Middle|T)\cdot P(Sound = No|T)}{\cancel{P(Size = big, Position = Middle, Sound = No)}} = \frac{5}{10}\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{4}{5} = 0.064$$

$P(F|Size = big, Position = Middle, Sound = No) =$

$$\frac{P(F)\cdot P(size = big|F)\cdot P(Position = Middle|F)\cdot P(Sound = No|F)}{\cancel{P(Size = big, Position = Middle, Sound = No)}} = \frac{5}{10}\cdot\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{2}{5} = 0.048$$

We deleted the denominator because it is the same in both calculations; thus, there is no need to calculate it to compare the probabilities.

$0.064 > 0.048$ *The prediction is True*

3. Understanding

a. Describe the analytical solution for linear regression with MSE as a distance function.

A simple linear regression with one independent variable. The linear equation is given by:

$$y = \beta_0 + \beta_1 x$$

The Mean Squared Error (MSE) for a set of data points (x_i, y_i) for N points is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

The analytical solution for finding the optimal coefficients β_0 *and* β_1 involves minimizing the MSE. This can be achieved by taking partial derivatives of the MSE with respect to β_0 *and* β_1 and setting them equal to zero.

b. What is the problem with information gain? Describe any solution for it

The problem with Information Gain is that it tends to favor attributes with many categories, giving them an advantage during decision tree construction. This can lead to biased tree structures.

A solution for this problem : Information Gain Ratio.

Information Gain Ratio biases the decision tree against considering attributes with many distinct values.

c. Why do we use Gradient Descent or Newton Raphson for Linear Regression?

Both Gradient Descent and Newton-Raphson are optimization algorithms employed in linear regression to iteratively minimize the cost function and find optimal parameter values, albeit differing in their approaches and computational characteristics.

d. Explain how a Decision tree is used for regression problems.

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.