Machine Learning – Ex2 theoretical

Question 1:

$$\mathrm{IG(S,T)} = \mathrm{H(S)} - \sum_{t \in T} p(t)H(t)$$

$$H(S) = \sum_{c \in Classes} -p(c)\log_2 p(c)$$

Clicked	Size	Position	Sound
F	Big	Top	No
F	Small	Middle	Yes
F	Small	Middle	Yes
T	Small	Bottom	No
T	Big	Bottom	No
F	Big	Top	Yes
T	Big	Bottom	Yes
T	Small	Middle	No
T	Small	Middle	No
F	Big	Top	No

Entropy(S)

- Entropy of clicking, given 10 samples, where 5 are positives.
- $H(S) = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right) = -1\log_2\left(\frac{1}{2}\right) = 1$

Information gain for **Position** split

- $IG(S|Possition) = H(S) - \sum_{Positioin} p(S|Position) \cdot H(S|Position)$
- $H(S|Position = Top)$

  $= -p(T|Position = Top)\log_2\big(p(T|Position = Top)\big) - p(F|Position = Top)\log_2\big(p(F|Position = Top)\big)$

  $= -0\log_2 0 - 1\log_2(1) = 0$
- $H(S|Position = Middle)$

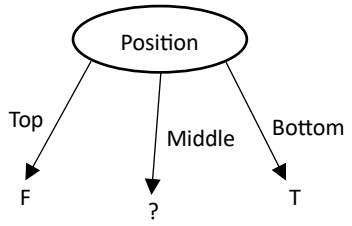
  $= -p(T|Position = Middle)\log_2\big(p(T|Position = Middle)\big) - p(F|Position = Middle)\log_2\big(p(F|Position = Middle)\big)$

  $= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$
- $H(S|Position = Bottom)$

  $= -p(T|Position = Bottom)\log_2\big(p(T|Position = Bottom)\big) - p(F|Position = Bottom)\log_2\big(p(F|Position = Bottom)\big)$

  $= -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - 0\log_2\left(\frac{0}{3}\right) \triangleq 0$
- $p(S|Position = Top) = \frac{3}{10}$
- $p(S|Position = Middle) = \frac{4}{10} = \frac{2}{5}$
- $p(S|Possition = Bottom) = \frac{3}{10}$

- $\Rightarrow IG(S, Position) = 1 - \frac{3}{10} * 0 - \frac{2}{5} * 1 - \frac{3}{10} * 0 = 0.6$



Clicked	Size	Position	Sound
~~F~~	~~Big~~	~~Top~~	~~No~~
F	Small	Middle	Yes
F	Small	Middle	Yes
~~T~~	~~Small~~	~~Bottom~~	~~No~~
~~T~~	~~Big~~	~~Bottom~~	~~No~~
~~F~~	~~Big~~	~~Top~~	~~Yes~~
~~T~~	~~Big~~	~~Bottom~~	~~Yes~~
T	Small	Middle	No
T	Small	Middle	No
~~F~~	~~Big~~	~~Top~~	~~No~~

Check for the next category to split by for the tree:

Entropy(S)

- Entropy of clicking, given 4 samples, where 2 are positives.
- $H(S) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = -1\log_2\left(\frac{1}{2}\right) = 1$

Information gain for **Sound** split

- $IG(S|Sound) = H(S) - \sum_{Sound} p(S|Sound) \cdot H(S|Sound)$
- $H(S|Sound = No)$

  $= -p(T|Sound = Big)\log_2\big(p(T|Sound = No)\big) - p(F|Sound = No)\log_2\big(p(F|Sound = No)\big)$

  $$= -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log\left(\frac{0}{2}\right) \triangleq 0$$
- $H(S|Sound = Yes)$
- $= -p(T|Sound = Yes)\log_2\big(p(T|Sound = Yes)\big) - p(F|Sound = Yes)\log_2\big(p(F|Sound = Yes)\big)$

  $$= -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) \triangleq 0$$
- $p(S|Sound = No) = \frac{2}{4} = \frac{1}{2}$
- $p(S|Sound = Yes) = \frac{2}{4} = \frac{1}{2}$

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

Information gain for **Size** split

- $IG(S|Size) = H(S) - \sum_{Size} p(S|Size) \cdot H(S|Size)$
- $H(S|Size = Big)$

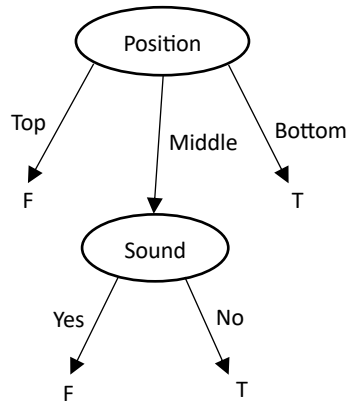
  $= -p(T|Size = Big)\log_2\big(p(T|Size = Big)\big) - p(F|Size = Big)\log_2\big(p(F|Size = Big)\big) = -\frac{0}{0}\log_2\left(\frac{0}{0}\right) - \frac{0}{0}\log_2\left(\frac{0}{0}\right) \triangleq 0$
- $H(S|Size = Small) = -p(T|Size = Small)\log_2\big(p(T|Size = Small)\big) - p(F|Size = Small)\log_2\big(p(F|Size = Small)\big)$

  $$= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$
- $p(S|Size = Big) = \frac{0}{4} = 0$
- $p(S|Size = Small) = \frac{4}{4} = 1$

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

IG(S, Sound) have the largest value between Size and Sound. Therefore, we choose Sound to be the next leaf in the tree.

We got the next tree:



Question 2:

Given the vector: $X = \{big, Middle, No\}$

$P(Clicked) = \frac{5}{10}$

$P(Clicked|X) = P(big|T) * P(middle|T) * P(no|T) * P(Clicked) = \left(\frac{2}{5}\right) * \left(\frac{2}{5}\right) * \left(\frac{4}{5}\right) * \left(\frac{1}{2}\right) = 0.064$

$P(Not\ Clicked|X) = P(big|F) * P(middle|F) * P(no|F) * P(Clicked) = \left(\frac{3}{5}\right) * \left(\frac{2}{5}\right) * \left(\frac{2}{5}\right)\left(\frac{1}{2}\right) = 0.048$

$P(Clicked|X) > P(Not\ Clicked|X) \Rightarrow the\ sample\ will\ predict\ as\ Clicked$

Question 3:

1. The analytical solution for linear regression with Mean Squared Error (MSE) as the distance function involves minimizing the MSE loss function, which calculates the squared difference between the predicted labels and the actual labels. This optimization function aims to minimize the average squared distance between the predicted and actual labels by adjusting the model parameters.
2. If there are many classes (feature value options) for the features, the probability for each class (of a specific feature) will be small, for example the ID feature has a different class (value) for each instance vector. The IG formula is sum of the probability of feature classes multiple by the entropy. Because there are many options values for the classes (of this feature), the probability of each class value will be small, and the total IG will be big. so, in general, the IG might give higher weight for features with a lot of categories\classes (value of a feature)
3. In some cases, we cannot find the minimum\maximum values for the loss\optimization functions. The matrix isn't invertible, or we cannot extract the extreme points from the derivative, or it is too hard to find the extreme points. So, we can use those iterative methods to identify those maximum and minimum points (mostly local minima).

4. In regression problems, a decision tree is used to predict continuous target variables by partitioning the feature space into regions based on the values of input features. The decision tree recursively splits the data into subsets by selecting the feature and split point that minimize the variance of the target variable within each subset. The predicted value for a new data point is then determined by averaging (for example) the target values of the training samples falling within the leaf node corresponding to the data point's feature values.