Decision Trees

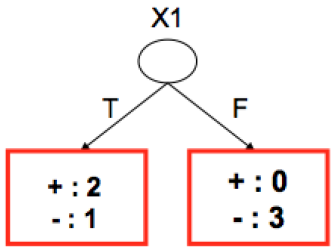
1. (a) Entropy of this collection of training examples:
   $H(S) = -\sum_i^n P(X = x_i) log_2 P(X = x_i)$
   Here, i is Yes or No.

$$H(S) = -\Big(\frac{1}{3} log_2 \frac{1}{3} + \frac{2}{3} log_2 \frac{2}{3}\Big) = 0.9183$$

   (b) Information gain of X_1 relative to the training examples:

$$I(X, X_1) = H(X) - H(X/X_1)$$
$$H(X/X_1) = \sum_{v \in \text{values of } X_1} P(X_1 = v) H(X/X_1 = v)$$

   Given: X_1 is either T or F (values,v):

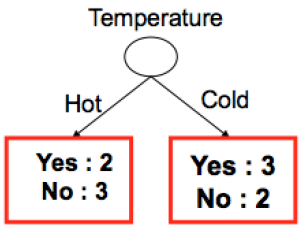


$$H(X/X_1 = T) = -\Big(\frac{2}{3} log_2 \frac{2}{3} + \frac{1}{3} log_2 \frac{1}{3}\Big) = 0.9183$$
$$H(X/X_1 = F) = 0 \text{ (Since all are -)}$$
$$I(X, X_1) = 0.9183 - \Big(\frac{3}{6} * 9183\Big) = 0.4591$$

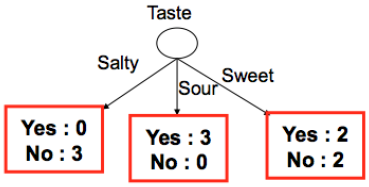
2. $H(X) = 1$ since there are equal number of options for Yes and No(5 each.)

   (a) Information Gain associated with choosing:

   - Temperature:

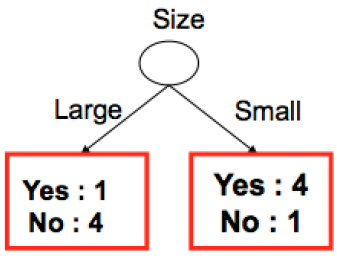


$$H(X/X_1 = hot) = -\Big(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\Big) = 0.9710$$
$$H(X/X_1 = cold) = -\Big(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\Big) = 0.9710$$
$$I(X, X_1) = 1 - \Big(\frac{5}{10} * 0.9710 + \frac{5}{10} * 0.9710\Big) = 0.029$$



- Taste:

$$H(X/X_1 = Salty) = 0$$
$$H(X/X_1 = Sour) = 0$$
$$H(X/X_1 = Sweet) = -\Big(\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\Big) = 1$$
$$I(X, X_1) = 1 - \Big(\frac{4}{10} * 1\Big) = 0.6$$



- Size:

$$H(X/X_1 = Small) = -\Big(\frac{4}{5}log_2\frac{4}{5} + \frac{1}{5}log_2\frac{1}{5}\Big) = 0.7219$$
$$H(X/X_1 = Large) = -\Big(\frac{1}{5}log_2\frac{1}{5} + \frac{4}{5}log_2\frac{4}{5}\Big) = 0.7219$$
$$I(X, X_1) = 1 - \Big(\frac{5}{10} * 0.7219 + \frac{5}{10} * 0.7219\Big) = 0.278$$

(b) Decision tree:

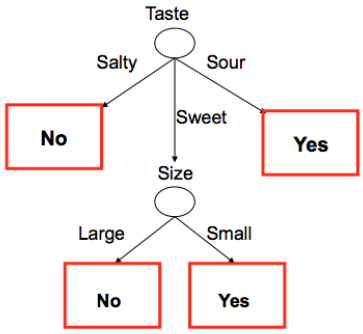
The first split would be by Taste as it has the highest Information gain of 0.6.

Its seen that for Taste=Salty, the decision is always No and for Taste=Sour, the decision is always Yes. So the tree is now split for Taste=Sweet. The choices are as follows:

Table 1

Appealing	**Temperature**	**Size**
No	Cold	Large
No	Cold	Large
Yes	Cold	Small
Yes	Cold	Small

The information gain in the case of Temperature is zero since H(Y/X=Temperature) = 1, whereas the The information gain in the case of Size is 1 since H(Y/X=Temperature) = 0. Thus the Decision Tree would look like the following:



3. (a) The maximum training error on a dataset with m labels can be 1.

   (b) No, both the trees need not have the same number of nodes, since they may have different entropies, thus different splitting criteria.

   (c) Yes, decision trees can perfectly classify any linearly separable dataset (can have arbitrary number of in-numerations for any n-level tree.)