SMAI In class Assignment

> Author: Vishal Reddy Mandadi
> Roll Number: 2019101119

Question 1

Information gain for split 1

$$E(parent) = -\sum(p(w_i)log(p(w_i)))$$
$$E(parent) = -1 * 4 * \frac{10}{40} * log_2(10/40) = -log_2(1/4) = 2$$
$$E(child1) = 2 * \frac{-10}{20} log_2(10/20) = 1$$
$$E(child2) = 2 * \frac{-10}{20} log_2(10/20) = 1$$
$$\text{information gain} = i_{parent} - \frac{n_1}{n} i_{child1} - \frac{n_2}{n} i_{child2}$$
$$\implies \text{information gain} = 2 - (\frac{20}{40} * 1) - (\frac{20}{40} * 1) = 2 - 1 = 1$$
$$\therefore \text{information gain for split 1} = 1$$

Information gain for split 2

Assumption: Split 2 isn't considered as the split following split 1. It was mentioned by the professor that we need to consider split 2 as a complete 1st level root split (similarly split 3). This implies, split 2 included the split created by the straight line and the dotted line that continues up from the straight line

$$E(parent) = -\sum(p(w_i)log(p(w_i)))$$
$$E(parent) = -1 * 4 * \frac{10}{40} * log_2(10/40) = -log_2(1/4) = 2$$
$$E(child1) = -1 * \left(\frac{10}{16} * log_2(\frac{10}{16}) + \frac{6}{16} * log_2(\frac{6}{16})\right)$$
$$= 3 - \frac{5}{8} log_2(5) - \frac{3}{8} log_2(3) = 3 - 1.451 - 0.594 = 0.955$$
$$E(child2) = -1 * \left(2 * \frac{10}{24} * log_2(\frac{10}{24}) + \frac{4}{24} * log_2(\frac{4}{24})\right)$$
$$E(child2) = -1(0.833 * (-1.265) + (0.167 * (-2.585))) = 1.485$$
$$\text{information gain} = i_{parent} - \frac{n_1}{n} i_{child1} - \frac{n_2}{n} i_{child2}$$
$$\implies \text{information gain} = 2 - (\frac{16}{40} * 0.955) - (\frac{24}{40} * 1.485) = 2 - 0.382 - 0.891 = 0.727$$
$$\therefore \text{information gain for split 2} = 0.727$$

Information gain for split 3

$$E(parent) = -\sum(p(w_i)log(p(w_i)))$$
$$E(parent) = -1 * 4 * \frac{10}{40} * log_2(10/40) = -log_2(1/4) = 2$$
$$E(child1) = -1 * \left(\frac{10}{14} * log_2(\frac{10}{14}) + \frac{4}{14} * log_2(\frac{4}{14})\right)$$
$$= -1 * (0.714 * (-0.486) + 0.286 * (-1.806)) = 0.864$$
$$E(child2) = -1 * \left(2 * \frac{10}{26} * log_2(\frac{10}{26}) + \frac{6}{26} * log_2(\frac{6}{26})\right)$$
$$E(child2) = -1(2 * 0.385 * (-1.377) + (0.231 * (-2.114))) = 1.549$$
$$\text{information gain} = i_{parent} - \frac{n_1}{n} i_{child1} - \frac{n_2}{n} i_{child2}$$
$$\implies \text{information gain} = 2 - (\frac{14}{40} * 0.864) - (\frac{26}{40} * 1.549) = 2 - 0.302 - 1.007 = 0.691$$
$$\therefore \text{information gain for split 2} = 0.691$$

Therefore, split 1 is most favorable. The order of favorable splits from most to least is

$$split1 > split2 > split3$$

Question 2

a. The problem with the given data is that all possible first splits give 0 information gain, due to which we won't be able to find the most favorable first split. Thus the algorithm gets stuck at the first step itself and might take up a least favorable split as its first choice leading to erroneous classification.

b. This problem can be solved in two ways:

   a. Pre-processing the data using PCA (principle component analysis) - this way we will be able to get the data projected onto a new set of the axis where it can be easily split by decision trees

   b. Using splits that are not parallel to the principle axis, i.e., using a linear combination of features instead of a single feature. The weights for the linear combination can be learned using gradient descent (similar to linear regression). The only downside with this method is that it trains very slowly