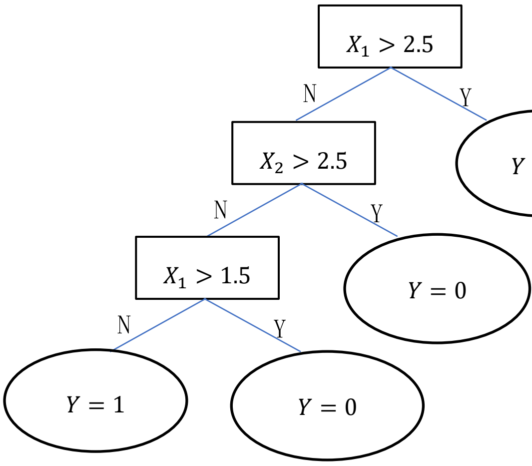
Reference Solution to Problem Set 4 - Q1

1. Classification and Regression Tree
   a) As long as any two data points have same covariates, we can always split the data set in a way that every covariate goes to the same leaf, so it leads to the same dependent variable.



   b) Gini index of the tree:

$$G_1 = 1 - \left(\frac{30}{30+10}\right)^2 - \left(\frac{10}{30+10}\right)^2 = 1 - \frac{9}{16} - \frac{1}{16} = \frac{3}{8} \quad ,$$

$$G_2 = 1 - \left(\frac{15}{15+25}\right)^2 - \left(\frac{25}{15+25}\right)^2 = 1 - \frac{9}{64} - \frac{25}{64} = \frac{15}{32} \quad ,$$

$$G_3 = 1 - \left(\frac{5}{5+25}\right)^2 - \left(\frac{25}{5+25}\right)^2 = 1 - \frac{1}{36} - \frac{25}{36} = \frac{5}{18} \quad ,$$

$$G_4 = 1 - \left(\frac{45}{45+20}\right)^2 - \left(\frac{20}{45+20}\right)^2 = 1 - \frac{81}{169} - \frac{16}{169} = \frac{72}{169} \quad ,$$

$$G_5 = 1 - \left(\frac{10}{10+25}\right)^2 - \left(\frac{25}{10+25}\right)^2 = 1 - \frac{4}{49} - \frac{25}{49} = \frac{20}{49} \quad ,$$

$$G = \frac{110}{110+100}\left[\frac{40}{40+70}G_1 + \frac{70}{40+70}\left(\frac{40}{40+30}G_2 + \frac{30}{40+30}G_3\right)\right]$$
$$+ \frac{100}{110+100}\left[\frac{65}{65+35}G_4 + \frac{35}{65+35}G_5\right] = 0.4003$$

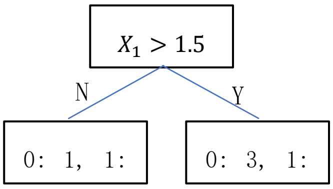
Gini index of data set:

$$G = 1 - \left(\frac{105}{105+105}\right)^2 - \left(\frac{105}{105+105}\right)^2 = 1 - 0.5^2 - 0.5^2 = 0.5$$

c) The Gini index of the data set is

$$G = 1 - \left(\frac{4}{4+5}\right)^2 - \left(\frac{5}{4+5}\right)^2 = 1 - \frac{16}{81} - \frac{25}{81} = \frac{40}{81}$$

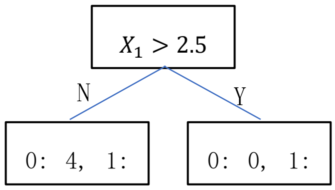
In order to choose the first splitting point, we choose the cutting point with the least Gini index. In total, there four possible ways to separate the dataset:



$$G_L = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$G_R = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

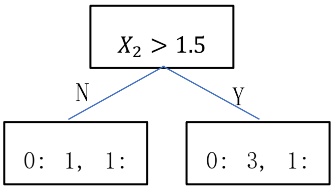
$$G_T = \frac{1}{3} * \frac{4}{9} + \frac{2}{3} * \frac{1}{2} = \frac{13}{27}$$



$$G_L = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$G_R = 1 - (0)^2 - (1)^2 = 0$$

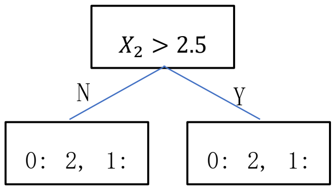
$$G_T = \frac{1}{3} * 0 + \frac{2}{3} * \frac{4}{9} = \frac{8}{27}$$



$$G_L = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$G_R = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$G_T = \frac{1}{3} * \frac{4}{9} + \frac{2}{3} * \frac{1}{2} = \frac{13}{27}$$



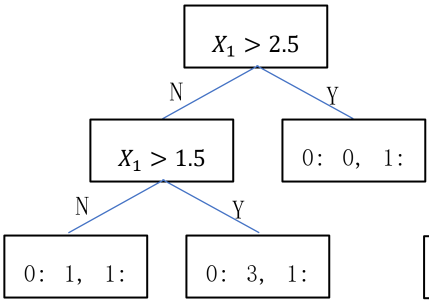
$$G_L = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

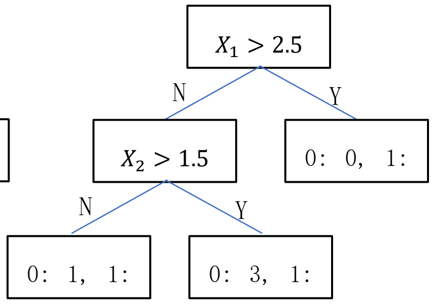
$$G_R = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

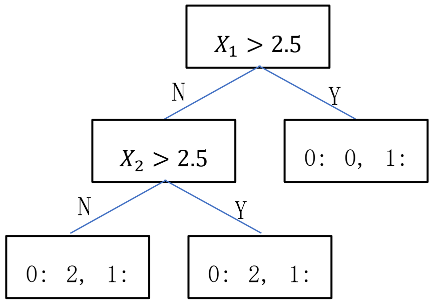
$$G_T = \frac{1}{3} * \frac{4}{9} + \frac{2}{3} * \frac{4}{9} = \frac{12}{27}$$

Since the $X_1 > 2.5$ produces the smallest tree Gini index, we choose $X_1 > 2.5$ to be the first split of the CART tree. With this split, Gini index decreases from $\frac{40}{81}$ to $\frac{8}{27}$ by $\frac{16}{81}$, which is larger than $c_p = 0.05$.

Next, we build the second layer of the tree. There are three remaining possibilities to split the tree shown below.







$$G_{LL} = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$G_{LR} = 1 - 1 - 0 = 0$$

$$G_L = \frac{1}{2} * \frac{4}{9} + 0 = \frac{2}{9}$$

$$G_T = \frac{2}{3} * \frac{2}{9} + 0 = \frac{4}{27}$$

$$G_{LL} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$G_{LR} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{3}{8}$$

$$G_L = \frac{1}{3} * \frac{1}{2} + \frac{2}{3} * \frac{3}{8} = \frac{5}{12}$$

$$G_T = \frac{2}{3} * \frac{5}{12} + 0 = \frac{5}{18}$$

$$G_{LL} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$G_{LR} = 1 - 1 - 0 = 0$$

$$G_L = \frac{2}{3} * \frac{1}{2} + 0 = \frac{1}{3}$$

$$G_T = \frac{2}{3} * \frac{1}{3} + 0 = \frac{2}{9}$$

The first choice gives the smallest Gini index of the tree, so we choose $X_1 > 1.5$ as the second split. So, we finished the building of the tree with max depth 2, and it looks like the first tree above.

d)

i. If $X = 2.2$, its predicted outcome is the mean of L_2, which is

$$Y = \frac{2.3 + 3.5 + 1.7 + 3.2 + 0.8}{5} = 2.3$$

ii. For L_1, $Y_1 = \frac{13.4+12.1+15.3+14.8+11.7}{5} = 13.46$; For L_2, $Y_2 = 2.3$.

$$\begin{aligned} SSE &= (13.4 - 13.46)^2 + (12.1 - 13.46)^2 + (15.3 - 13.46)^2 \\ &\quad + (14.8 - 13.46)^2 + (11.7 - 13.46)^2 + (2.3 - 2.3)^2 \\ &\quad + (3.5 - 2.3)^2 + (1.7 - 2.3)^2 + (3.2 - 2.3)^2 + (0.8 - 2.3)^2 \\ &= 14.992 \end{aligned}$$

For the whole training set, the mean is

$$\bar{Y} = \frac{13.4 + 12.1 + 15.3 + 14.8 + 11.7 + 2.3 + 3.5 + 1.7 + 3.2 + 0.8}{10} = 7.88$$

$$SST = (13.4 - 7.88)^2 + (12.1 - 7.88)^2 + (15.3 - 7.88)^2 + (14.8 - 7.88)^2 + (11.7 - 7.88)^2 + (2.3 - 7.88)^2 + (3.5 - 7.88)^2 + (1.7 - 7.88)^2 + (3.2 - 7.88)^2 + (0.8 - 7.88)^2 = 326.356$$

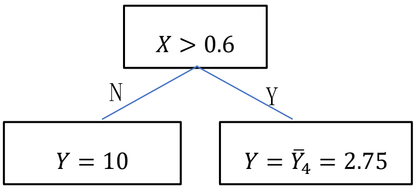
In this given data set, the in-sample performance is better than the baseline model.

e) For the five data points, the mean is

$$\bar{Y} = \frac{10 + 8 - 4 + 3 + 4}{5} = 4.2$$

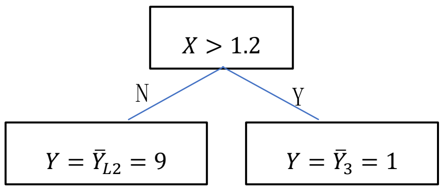
$$SST = (10 - 4.2)^2 + (8 - 4.2)^2 + (-4 - 4.2)^2 + (3 - 4.2)^2 + (4 - 4.2)^2 = 116.8$$

There are four possibilities to split the data firstly.



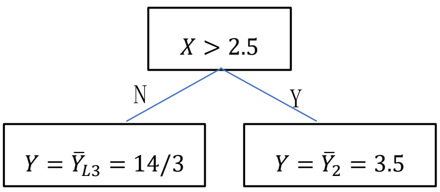
$$\text{SSE} = 0 + \sum_{i=2}^{5} (Y_i - \bar{Y}_4)^2 = 74.75$$

$$R^2 = 1 - \frac{74.75}{116.8} = 0.360$$



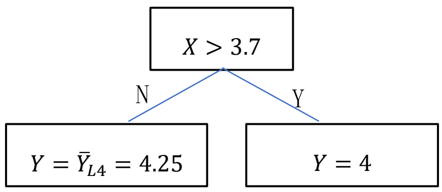
$$\text{SSE} = \sum_{i=1}^{2} (Y_i - \bar{Y}_{L2})^2 + \sum_{i=3}^{5} (Y_i - \bar{Y}_3)^2 = 40$$

$$R^2 = 1 - \frac{40}{116.8} = 0.658$$



$$\text{SSE} = \sum_{i=1}^{3} (Y_i - \bar{Y}_{L3})^2 + \sum_{i=4}^{5} (Y_i - \bar{Y}_2)^2 = 115.17$$

$$R^2 = 1 - \frac{115.17}{116.8} = 0.014$$



$$\text{SSE} = \sum_{i=1}^{4} (Y_i - \bar{Y}_{L4})^2 + 0 = 116.75$$

$$R^2 = 1 - \frac{116.75}{116.8} = 0.0004$$

Since $X > 1.2$ generates the largest R^2, we choose $X > 1.2$ as the split. The associated R^2 is 0.658.

f) The scales/units of the covariates do not matter, since we can adjust the unit/scale of the axis to keep the relative relationship, which generates the same CART tree. From another perspective, for a classification tree, it only cares about the number of different outcome in each leaf to grow the tree without considering the absolute values, and for a regression tree, R^2 will cancel out the effect of unit, so both of them have no influence on the scales/units of the covariates.