Name: Jay M Joshi

Student id: 200440993

Assignment: 4

Machine Learning Fall 2020

Q1 Argue that at each iteration of k-means clustering algorithm decreases the following objective $\sum_{j=1}^{k} \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2$ where

$$\mu_j = \frac{1}{n_j} \sum_{i=z_i:j} x_i$$

→ K-means clustering is a simple and elegant approach for partitioning a ~~cluster~~ data set into k distinct overlapping cluster.

→ The main idea behind k-means clustering is that a good clustering is one for which within-cluster variation is as small as possible.

→ Hence we need to minimize inter-cluster distance.

Suppose within-cluster variation for cluster C_k is measure $WC(C_k)$. Hence, we want to solve the problem

$$\underset{C_1, C_2 \ldots C_k}{\text{minimize}} \left\{ \sum_{i=1}^{k} W(C_k) \right\} \quad \text{---------- (1)}$$

→ There are many possible ways of defining within cluster variation. Common way is Euclidean distance

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \quad \text{---(2)}$$

Equation —(2) represents the within cluster variation for kth cluster i.e. sum of all pair wise squared Euclidean distance between the observations. Hence idea of minimizing —(1) is same as minimizing —(2). Combining —(1) and —(2) objective becomes,

$$\underset{C_1, \ldots, C_k}{\text{minimize}} \left\{ \sum_{k=1}^{k} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \quad \text{---(3)}$$

k-means algorithm optimizes above function by following steps:

(1) Randomly assigns center to each of clusters 1 .... k

(2) Iterate following steps until cluster assignment is not changed (remains unchanged)

(a) For each of k clusters compute the cluster centroid. The centroid of kth cluster is vector of p feature means for observations in k th cluster

(b) Assign each observation to cluster whose centroid is closests

→ The above algorithm is guranteed to decrease the following objective

$$\sum_{j=1}^{k} \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2 \quad \text{were} \qquad \text{---(4)}$$

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=j} x_i \quad \text{because,}$$

→ In step 2(a) the centroid of clusters for each feature are constants that minimizes the sum of square deviations i.e. within cluster variation and,

→ In step 2(b) reallocating will only improve —(4)

→ Hence it is observed that during the running course of algorithm, the clusters will continuosly improve until there are no longer changes. Hence the objective in equation —(3) will never increase.

→ when there is no change in the results, a local optimum is reached. So k-means algorithm finds a local optimum rather than a global optimum

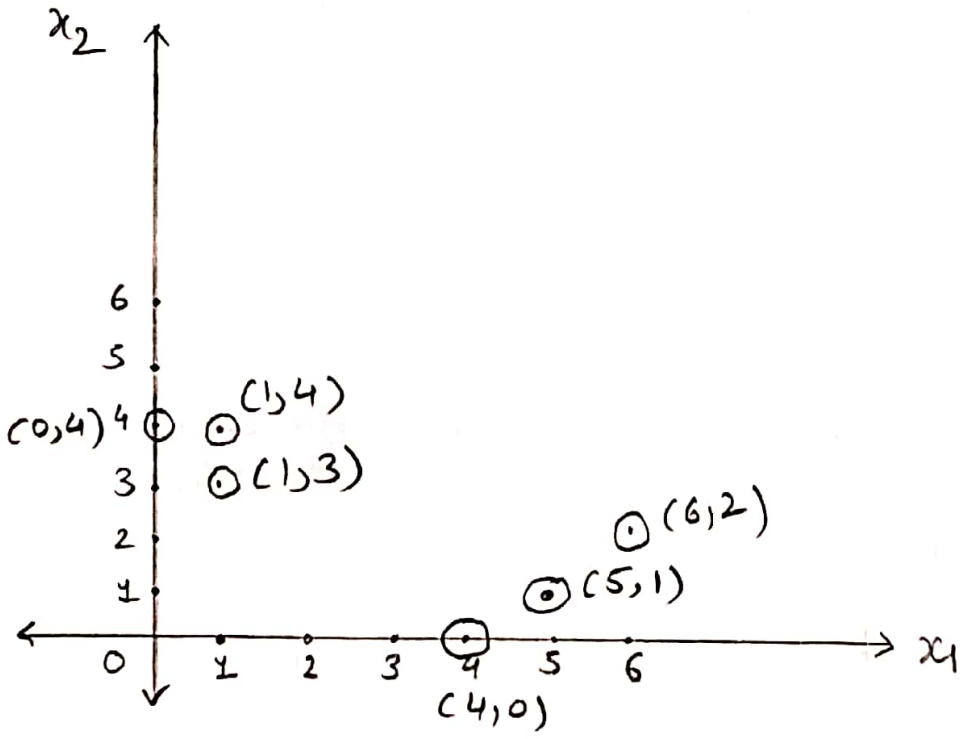
→ Therefore results, will depend upon random initializations since it suffers from local optimum, it is necessary to run algorithm multiple times with different random initializations.

Q2

Given data points:

Observation	x_1	x_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

a) observations are plotted below



(b) Applying k-means algorithm

Step 1: Initializing cluster centers (k=2)

Let $u_1 = (1, 4)$ $u_2 = (5, 1)$

Step 2: Assigning data points to nearest clusters.

we are going to use Manhattan distance to calculate distance between two points

Manhattan distance is given by (two points $P_1(x_1, y_1)$ $P_2(x_2, y_2)$)

$$= |x_1 - x_2| + |y_1 - y_2|$$

→ Hence calculating distance of each data points from two clusters and assigning them to nearest cluster center.

→ Hence, data points:

1 (1, 4)

distance of (1, 4) from $u_1(1, 4)$ is

$|1-1| + |4-4| = 0$

distance of (1, 4) from $u_2(5, 1)$ is

$|1-5| + |4-1| = 4 + 3 = 7$

2 (1, 3)

distance of (1, 3) from $u_1(1, 4)$ is

$|1-1| + |3-4| = 1$

distance of (1, 3) from $u_2(5, 1)$ is

$|1-5| + |3-1| = 4 + 2 = 6$

3 (0,4)

distance of (0,4) from u_1(1,4) is

$|0-1| + |4-4| = 1$

distance of (0,4) from (5,1) is

$|0-5| + |4-1| = 8$

4 (5,1)

distance of (5,1) from (1,4) is

$|5-1| + |1-4| = 4 + 3 = 7$

distance of (5,1) from u_2(5,1) is

$|5-5| + |1-1| = 0$

5 (6,2)

distance of (6,2) from u_1(1,4) is

$|6-1| + |2-4| = 5 + 2 = 7$

distance of (6,2) from u_2(5,1) is

$|6-5| + |2-1| = 2$

6 (4,0)

distance of (4,0) from u_1(1,4) is

$|4-1| + |0-4| = 3 + 4 = 7$

distance of (4,0) from u_2(5,1) is

$|4-5| + |0-1| = 1 + 1 = 2$

Hence after first iteration)

data points	Distance from u_1 & u_2		
	distance from u_1 (1,4)	distance from u_2 (5,1)	Nearest cluster
(1,4)	0	7	u_1
(1,3)	1	6	u_1
(0,4)	1	8	u_1
(5,1)	7	0	u_2
(6,2)	7	2	u_2
(4,0)	7	2	u_2

→ After first iteration

Data points in cluster 1 with $u_1 = (1,4)$

- (1,4)
- (1,3)
- (0,4)

Data points in cluster 2 with $u_2 = (5,1)$

$u_2 = (5,1)$

- (5,1)
- (6,2)
- (4,0)

→ Now cluster centers are updated by taking mean of all points.

Hence,

Updated centroid for cluster 1 $c_1 = \left(\frac{0+1+1}{3}, \frac{4+4+3}{3}\right) = \left(\frac{2}{3}, \frac{11}{3}\right)$

Updated centroid for cluster 2 $c_2 = \left(\frac{6+5+4}{3}, \frac{2+1+0}{3}\right) = (5, 1)$

→ Hence, applying k-means algorithm with $c_1 = (2/3, 11/3)$ and $c_2 = (5, 1)$.

→ calculating distance of data points from each clusters c_1 and c_2

1 (1, 4)

distance of (1, 4) from c_1 $(2/3, 11/3)$ is

$$|1 - \frac{2}{3}| + |4 - \frac{11}{3}| = \frac{1}{3} + \frac{1}{3} = 2/3$$

distance of (1, 4) from c_2 (5, 1) is

$$|1-5| + |4-1| = 7$$

2 (1, 3)

distance of (1, 3) from c_1 $(2/3, 11/3)$ is

$$|1 - \frac{2}{3}| + |3 - \frac{11}{3}| = \frac{1}{3} + \frac{2}{3} = 1$$

distance of (1,3) from u_2 (5,1)

$|1-5| + |3-1| = 6$

3 (0,4)

distance of (0,4) from u_1: $(2/3, 11/3)$ is

$|0-\frac{2}{3}| + |4-\frac{11}{3}| = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$

distance of (0,4) from u_2 (5,1) is

$|0-5| + |4-1| = 8$

4 (5,1)

distance of (5,1) from u_1: $(2/3, 11/3)$ is:

$|5-\frac{2}{3}| + |1-\frac{11}{3}| = \frac{13}{3} + \frac{8}{3} = 7$

distance of (5,1) from u_2: (5,1) is

$|5-5| + |1-1| = 0$

5 (6,2)

distance of (6,2) from u_1: $(2/3, 11/3)$ is

$|6-\frac{2}{3}| + |2-\frac{11}{3}| = \frac{16}{3} + \frac{5}{3} = 7$

distance of (6,2) from u_2 (5,1) is:

$|6-5| + |2-1| = 2$

6 (4,0)

distance of (4,0) from u_4: $(2/3, 11/3)$ is

$$|4 - \frac{2}{3}| + |0 - \frac{11}{3}| = \frac{10}{3} + \frac{11}{3} = 7$$

distance of (4,0) from u_2: (5,1) is

$$|4-5| + |0-1| = 2$$

Hence after second iteration,

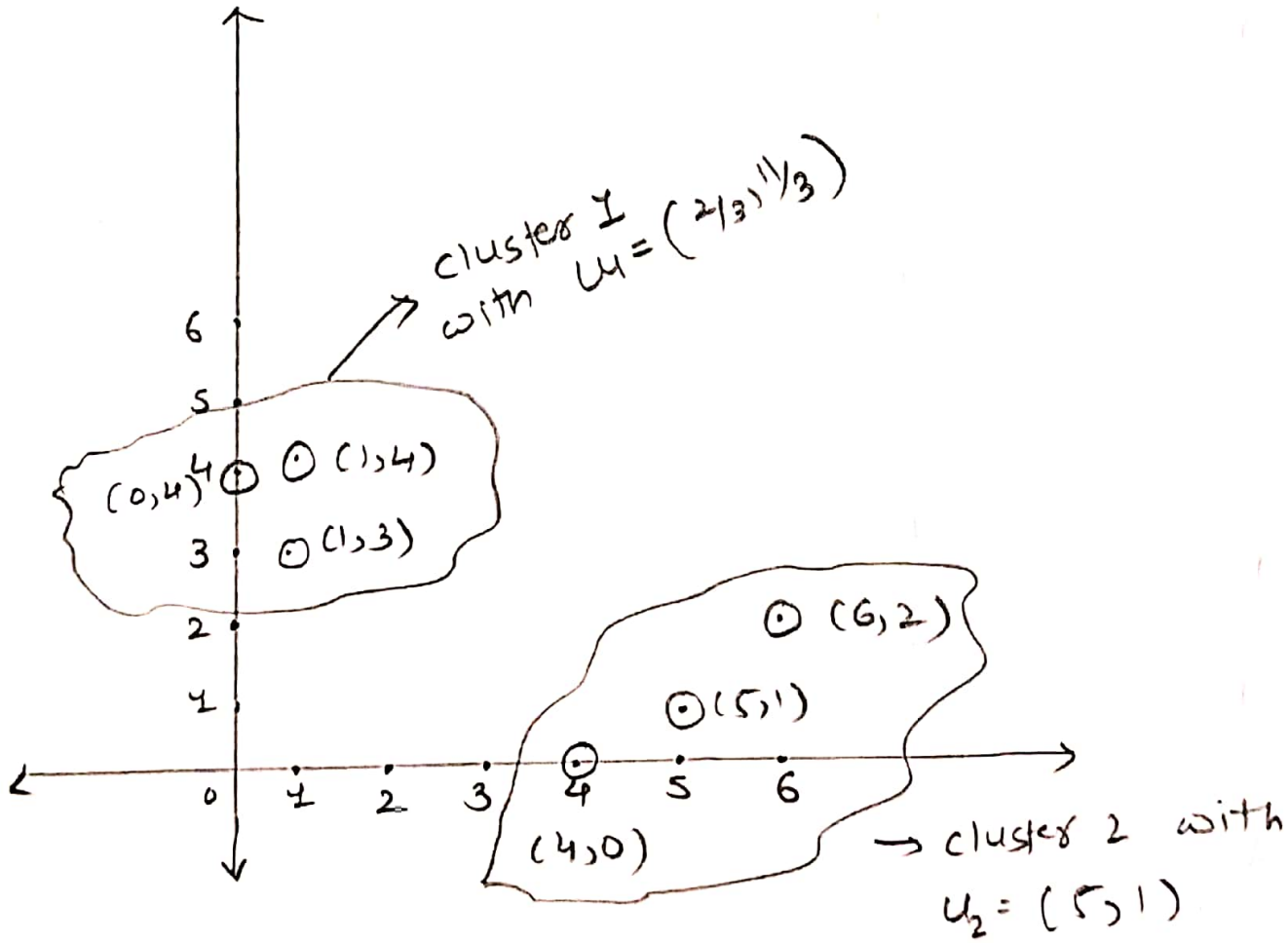
data points	Distance from u_4 & u_2 distance from u_4: $(2/3, 11/3)$	distance from u_2: (5,1)	Nearest cluster
(1,4)	2/3	7	u_4.
(1,3)	1	6	u_4
(0,4)	1	8	u_4
(5,1)	7	0	u_2
(6,2)	7	2	u_2
(4,0)	7	2	u_2

→ After Second iteration,

Data points in cluster 1 with $u_1 = (2/3, 11/3)$	Data points in cluster 2 with $u_2 = (5,1)$
· (1,4)	· (5,1)
· (1,3)	· (6,2)
· (0,4)	· (4,0).

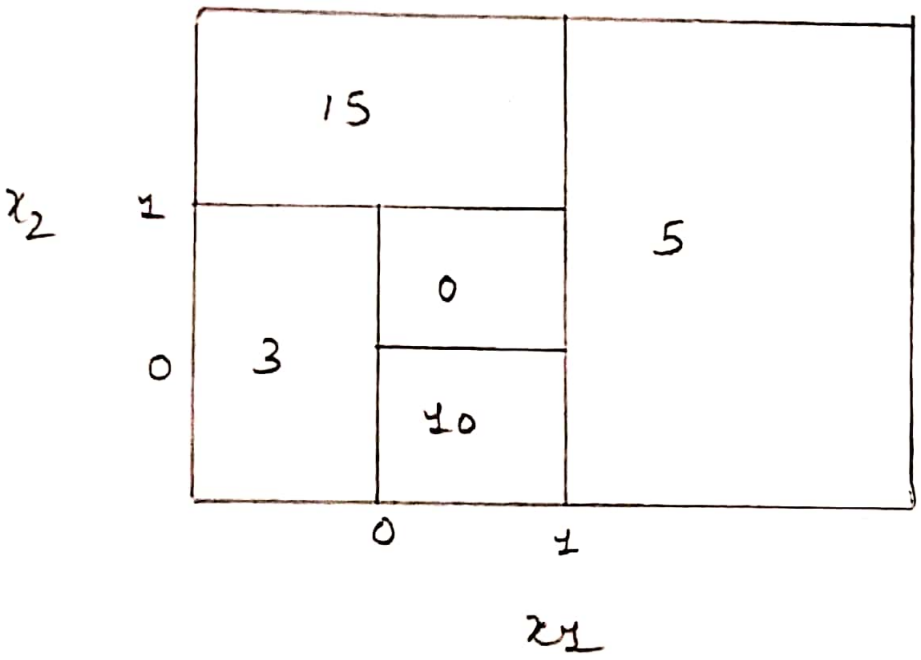
→ Since data points in each individual clusters are not updated from previous iteration, k-means clustering algorithm terminates here.

c)



q3

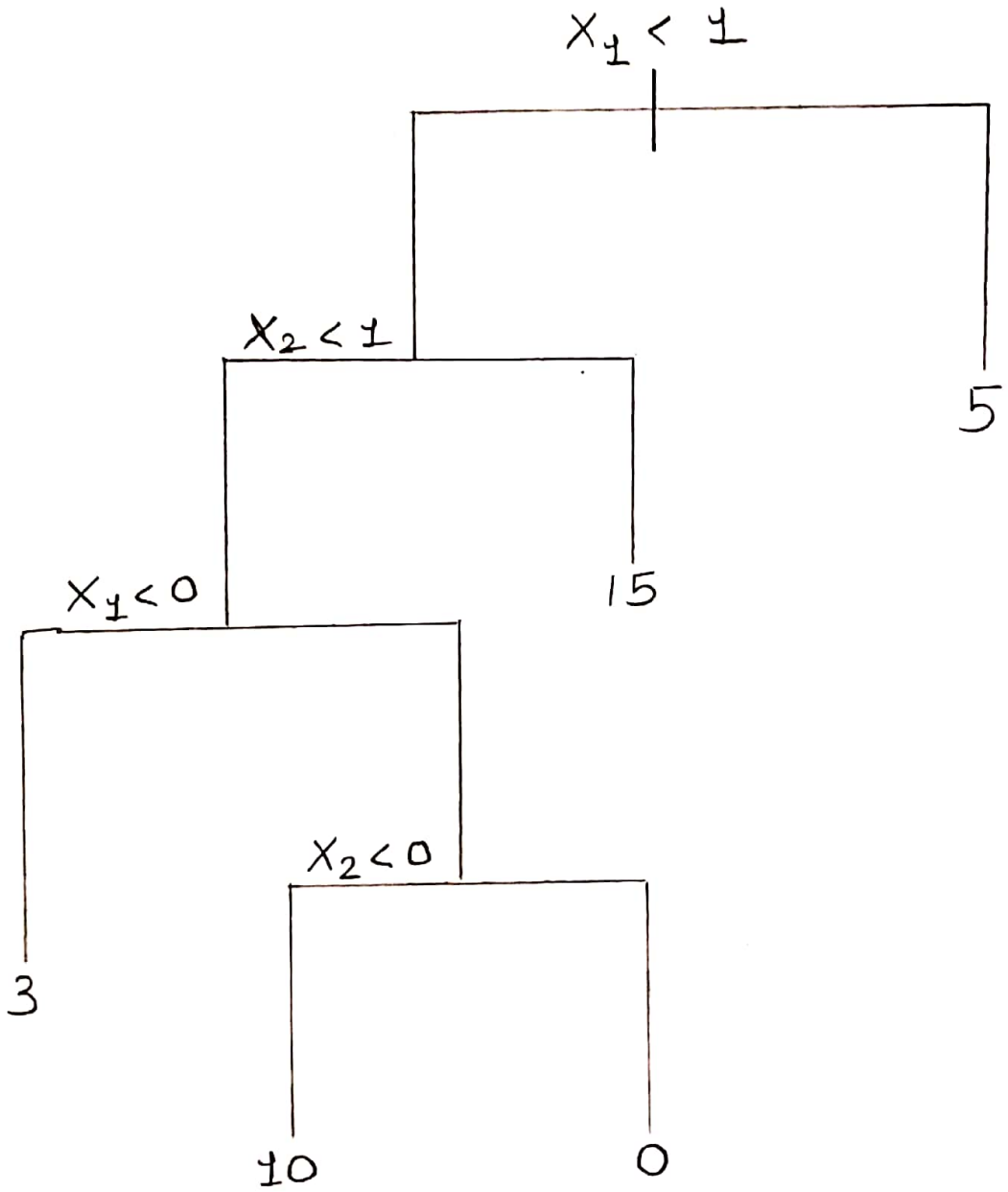
a) Sketch the tree corresponding to partition space as below:



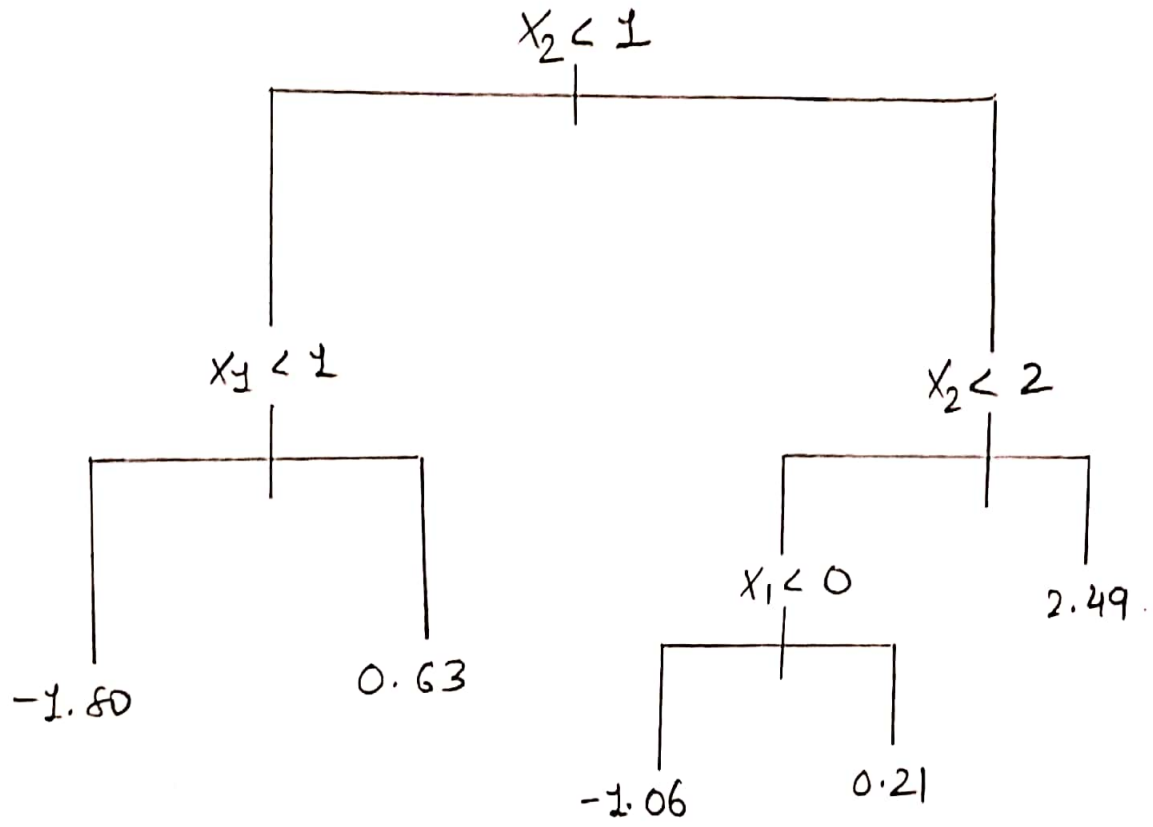
the tree is sketched by following if else loop:

if $x_1 \geq 1$ then 5, else if $x_2 \geq 1$ then 15, else if $x_1 < 0$ then 3, else if $x_2 < 0$ then 10, else 0

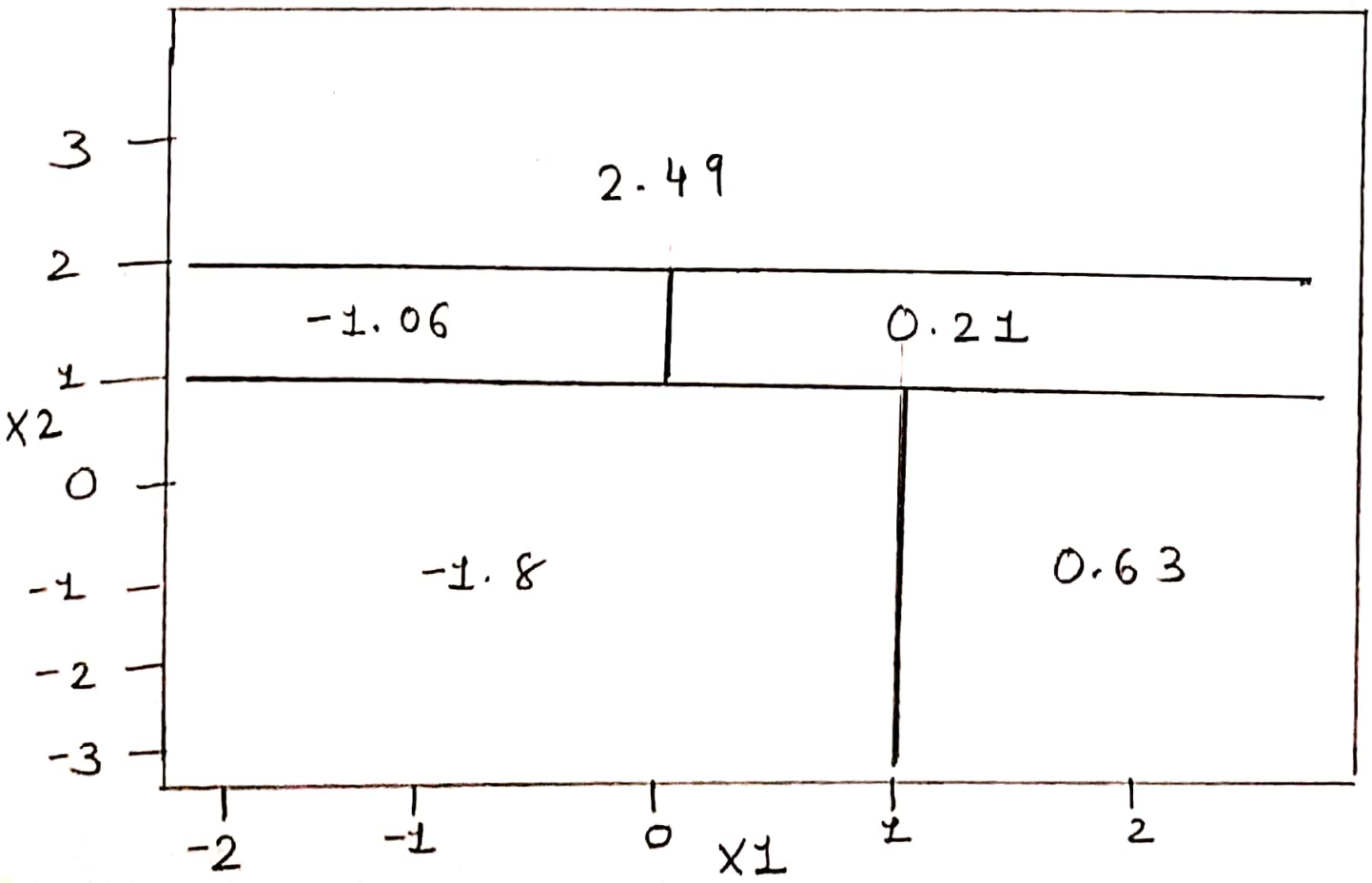
→ The tree constructed from above partition space is as shown below.



(b) the tree is given as



→ Partition space corresponding to above tree is shown below



Q4

Brief summary highlighting the difference and similarity between three ensemble methods Bagging, Random forest and boosting.

Important notes are:

→ Decision trees can easily overfit the training data if we allow growing of trees to its full depth. Hence, decision trees might not generalize well on unseen data. because of overfitting. Hence, the concept of ensemble technique came into existence.

→ Hence in ensemble learning we are combining the set of individual learners. The main purpose of combining individual learners is to force each individual learners to make independent errors. In order to make independent errors, each individual learners can be trained on different subset of data, different algorithms.

→ Different data points for each individual learners can be generated by sampling data with replacement from main population.

→ Each of these data sets can have duplicates.

→ Three main ensemble techniques are summarized below:

1) Random Forest.

→ Random forests are usually one of two top performing algorithms along with boosting in prediction contests.

→ In Random forests individual learners are strong learners i-e they have high variance and low bias. The underlying foundation is to combine all these learners to reduce variance (because averaging complex models can reduce variance)

→ The main steps of algorithm are:

1. Bootstrap sample
2. At each split bootstrap individual features (variables)
3. Grow multiple trees and make prediction as
   - Regression: average all predictions from all trees
   - classification: majority vote among all trees.

Advantages:

→ Maximum accuracy can be achieved with Random forest.

→ Areraging reduces variance, hence we end up lowering both variance & bias.

Disadvantages:

→ Reduced speed (Because we are growing large number of trees)

→ Compromised interpretability for higher accuracy.

2) Bagging.

→ The main idea is to built individual trees by boostrap aggregating.

→ Each individual learners are strong learners. i.e with high variance and low bias.

The main steps of algorithm are.

1) Resampling data points and recalculating predictions.

2) Grow multiple trees, make a final prediction as

- Regression: average all predictions from trees
- classification: majority vote among all trees.

Advantages:

- Accuracy
- we end up lowering both variance and bias.

Disadvantages:

- Speed and interpretability
- High correlation among individual learners.

3) Boosting :

→ The key idea is to combine weak learners with high bias and low variance.

→ The main steps of algorithm are :

1 Initially we built a weak learner by assigning equal probability to each instance in data set.

2 At every next iteration, we will increase the probability of instances which were not classified correctly by previous learners.

3 Each classifier is given weight on basis of errors they make. If weight is high means classifier makes relatively less error than others.

4 Final classifier is linear combination of all learners.

Advantages :-

- Easy to interpret
- Resilient method that curbs overfitting easily
- strong prediction power.

Disadvantages :

- Sequential in nature
- Sensitive to outliers.

Summarizing main differences & similarities among all three algorithm.

Boosting	Bagging	Random forest.
• Individual trees are weak learners with high bias & low variance	• Individual trees are strong learners with high variance and low bias	• Individual trees are strong learners with high variance and low bias
• Individual trees built during algorithm are dependent on each other	• Individual trees built during algorithm are independent of each other	• Individual trees built during algorithm are independent of each other.
• There is correlation, because all features in each individual learners are used to make predictions	• There is correlation, because all features in each individual learners are used to make predictions	- Decorrelates the trees, because random features are selected to grow individual learners.
• Purpose is to decrease bias	• purpose is to decrease variance	• Purpose is to decrease variance
• Base learners are stable learners with low variance	• Base learners are unstable learners with high variance	• Base learners are unstable learners with high variance.
• Bootstrap aggregation of data points.	• Bootstrap aggregation of data points	• Boostrap aggregation of data points + bootstrap features (variable
• Easy to interpret with accuracy	• compromised interpretability plus speed	• Compresed interpretability plus speed.

List of References:

[1] "An introduction to statistical learning with Applications in R" by James, Witten, Hastie and Tibshirani"