50.007 – Machine Learning

May–August Term, 2022

Homework 2 (23 May 2022)

Question 1. (Subtotal: 25 points)
(not a coding question)

Consider a dataset with 6 data points: $(0,0), (1,1), (1,-1), (3,1), (3,-1), (4,0)$ and answer the following questions:

1.1
Simulate k-means clustering on this dataset with k=2 and the initial two centroids as (0, 0), (1, -1). For each iteration show the two updated centroids, the distances between each data point to each centroid, and the cluster assignment of each instance (data point). Your algorithm should end when your cluster assignment stops changing or your centroid positions stop changing, whichever comes first. (10 points)

1.2
Plot 2D graphs for the k-means simulation process above in **1.1**: starting with the graph of the initial two centroids and the cluster assignment generated by these two centroids. Then make a new 2D graph with updated cluster assignment whenever you moved centroids. Your final graph should illustrate the final centroid positions and the final cluster assignment (6 points).

1.3
By observing the graph above, we can easily see that $(0,0), (1,1), (1,-1)$ should belong to one cluster, and $(3,1), (3,-1), (4,0)$ should belong to the other cluster. Although our initial two centroids are very close to each other, K-means algorithm still manages to find the desired or optimum clustering result (the "global minimum") for this dataset. Think about why this is happening. Then:
1.31 Give an example of a set of data that satisfies the following conditions (6 points)
1) When we run a K-Means clustering on the dataset, the final clustering result is always the same no matter which initial centroids the K-Means algorithm starts with (the centroids are randomly picked from the training dataset).
2) For the purpose of simplicity, your example dataset should have 4 2-dimensional data points, the K-Means algorithm should start with 2 centroids, and each resulting/final cluster should have 2 data points.
1.32 In your own words, explain how you came up with your example dataset (in other words, what principles did you follow when you were creating your example dataset in order to make it satisfy the conditions in **1.31**).(3 points)

Answers:
1.1
• Initial Centroids: $(0,0), (1,-1)$
• Distances between each data point to each centroid:

	$(1,1)$	$(3,1)$	$(3,-1)$	$(4,0)$
$(0,0)$	1.41	3.16	3.16	4
$(1,-1)$	2	2.83	2	3.16

• Initial Cluster Assignment:
cluster 1: $(0,0),(1,1)$. Cluster 2: $(1,-1),(3,1),(3,-1),(4,0)$

• Updated Centroids: (0.5, 0.5), (2.75, -0.25)
• Distances between each data point to each centroid:

	$(0,0)$	$(1,1)$	$(1,-1)$	$(3,1)$	$(3,-1)$	$(4,0)$
$(0.5,0.5)$	0.71	0.71	1.58	2.55	2.92	3.54
$(2.75,-0.25)$	2.76	2.15	1.90	1.27	0.79	1.27

• Updated Cluster Assignment:
cluster 1: $(0,0),(1,1),(1,-1)$. Cluster 2: $(3,1),(3,-1),(4,0)$

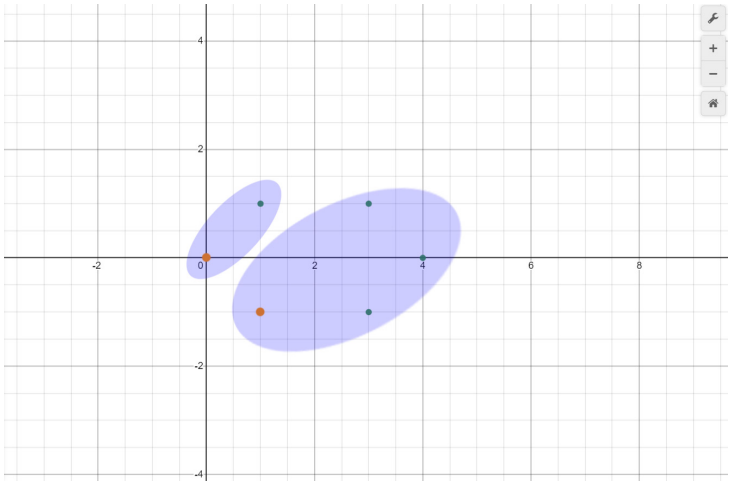
• Updated Centroids: (0.67, 0), (3.33, 0)
• Distances between each data point to each centroid:

	$(0,0)$	$(1,1)$	$(1,-1)$	$(3,1)$	$(3,-1)$	$(4,0)$
$(0.67,0)$	0.67	1.05	1.05	2.54	2.54	3.33
$(3.33,0)$	3.33	2.54	2.54	1.05	1.05	0.67

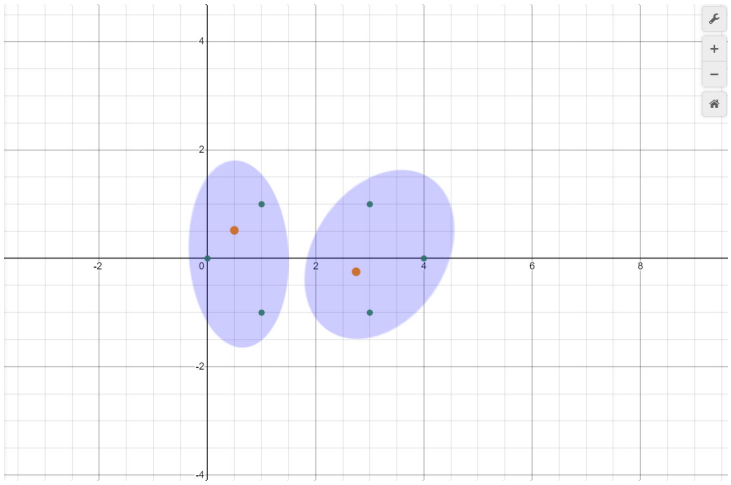
• Updated Cluster Assignment:
cluster 1: $(0,0),(1,1),(1,-1)$. Cluster 2: $(3,1),(3,-1),(4,0)$

• Since cluster assignment no longer changes, K-Means termination condition is met. K-Means stops.

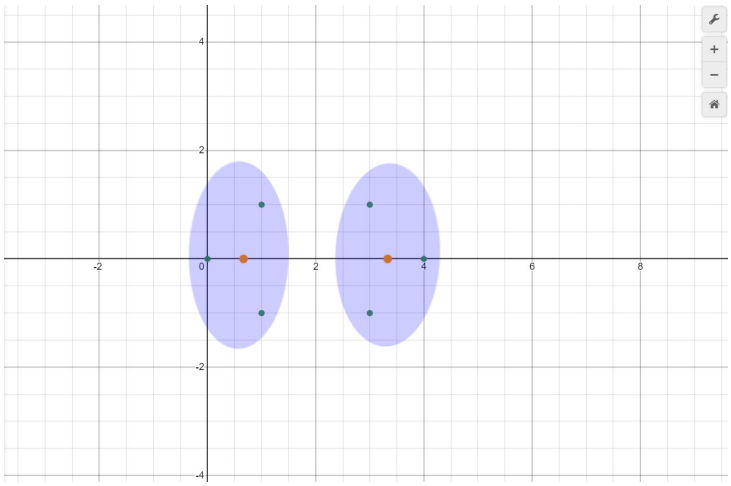
1.2
• The initial centroids and clustering (centroids are in orange color):



• The first update of centroids and clustering:



• The final update of centroids and clustering (this time clustering is not changing, so algorithm stops):



1.3

1.31

• One dataset example that satisfy all the conditions can be: (1,0),(2,0),(102,0),(104,0)

1.32

• In order to satisfy the conditions in **1.31**, we want to make sure that even when the initial two centroids are very close to each other, the updated centroids will become more and more far away from each other.

One way to make sure this will happen is to make sure that the points (instances) that are far away from the initial two centroids are all closer to one of the centroids. In this way, the far-away points (instances) will make the updated centroids far away from each other through the mean calculation procedures.

(This is just one of the principles I can think of. Other reasonable explanations are also acceptable.)

Question 2. (Subtotal: 25 points)
(not a coding question)

Consider a dataset with 4 1-dimensional data points: 1, 2, 3, 3. Their labels are 1, 1, 2, 3. Below is a table of this dataset with labels for each data point:

Data point	Label
1	1
2	1
3	2
3	3

Please answer the following questions:

2.1
Please use the closed form solution of linear regression to find out:
1) the optimized predictor (which is a line) that passes through the origin. (5 points)
2) the optimized predictor (which is a line) that does not pass through the origin. (5 points)

2.2
Plot the two predictors you found in **2.1** (one passes through the origin and the other does not) on a 2D graph. In this graph, x axis is the input data point value and y axis is the predicted value. Then, plots the line y = -0.5 + x on the same graph. Please also plot (1, 1), (2, 1), (3, 2), and (3, 3) on the same graph. (5 points).

2.3
Please calculate the $R_n(\theta)$ (training error) for each of the three lines (predictors) you plotted in **2.2** graph. Which line (predictor) gives us the smallest training error? (10 points)

Answers:
2.1

- 1) $X = \begin{bmatrix}1\\2\\3\\3\end{bmatrix}, X^T = \begin{bmatrix}1 & 2 & 3 & 3\end{bmatrix}, y = \begin{bmatrix}1\\1\\2\\3\end{bmatrix}$

$$so, b = \frac{1}{4}X^T y = \frac{1}{4}\begin{bmatrix}1 & 2 & 3 & 3\end{bmatrix}\begin{bmatrix}1\\1\\2\\3\end{bmatrix} = \frac{1}{4} \times 18 = 4.5$$

$$A = \frac{1}{4}X^T X = \frac{1}{4}\begin{bmatrix}1 & 2 & 3 & 3\end{bmatrix}\begin{bmatrix}1\\2\\3\\3\end{bmatrix} = \frac{1}{4} \times 23 = 5.75$$

Thus, $\theta = A^{-1}b = \frac{1}{5.75} \times 4.5 = 0.783$, and the optimized predictor (which is a line) that passes through the origin is $y = 0.783x$

• 2)$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{bmatrix}, X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 3 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \end{bmatrix}$

so, $b = \frac{1}{4}X^T y = \frac{1}{4} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 7 \\ 18 \end{bmatrix} = \begin{bmatrix} 1.75 \\ 4.5 \end{bmatrix}$

$A = \frac{1}{4}X^T X = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 4 & 9 \\ 9 & 23 \end{bmatrix} = \begin{bmatrix} 1 & 2.25 \\ 2.25 & 5.75 \end{bmatrix}$

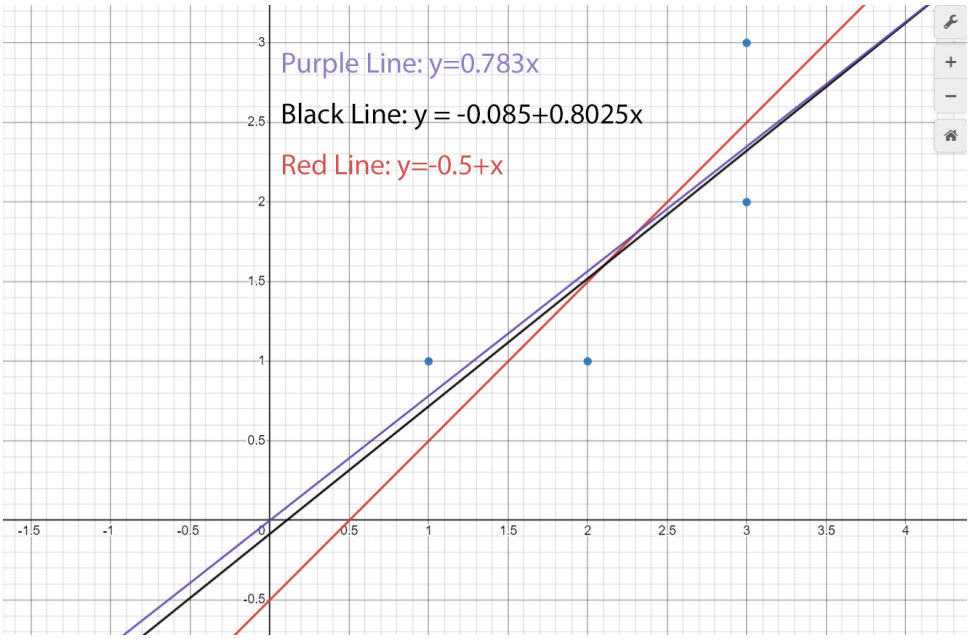
$A^{-1} = \begin{bmatrix} 8.36 & -3.27 \\ -3.27 & 1.45 \end{bmatrix}$

Thus, $\theta = A^{-1}b = \begin{bmatrix} 8.36 & -3.27 \\ -3.27 & 1.45 \end{bmatrix}\begin{bmatrix} 1.75 \\ 4.5 \end{bmatrix} = \begin{bmatrix} -0.085 \\ 0.8025 \end{bmatrix}$

Thus, the optimized predictor (which is a line) that does not pass through the origin is $y = -0.085 + 0.8025x$

2.2

Below is the graph comparing the three lines:



2.3

• For the line $y = 0.783x$, $R_n(\theta) = \frac{1}{4}\sum_{t=1}^{n}\frac{1}{2}(y^t - \theta x^t)^2 = \frac{1}{4}[0.5 \times (1 - 0.783)^2 + 0.5 \times (1 - 1.566)^2 + 0.5 \times (2 - 2.349)^2 + 0.5 \times (3 - 2.349)^2] = \frac{1}{8} \times 0.9130 = 0.114$

• For the line $y = -0.085+0.8025x$, $R_n(\theta) = \frac{1}{4}\sum_{t=1}^{n}\frac{1}{2}(y^t - \theta x^t)^2 = \frac{1}{4}[0.5\times(1 - 0.7175)^2 + 0.5 \times (1 - 1.52)^2 + 0.5 \times (2 - 2.3225)^2 + 0.5 \times (3 - 2.3225)^2] = \frac{1}{8} \times 0.9132 = 0.114$

• For the line $y = -0.5 + x$, $R_n(\theta) = \frac{1}{4}\sum_{t=1}^{n}\frac{1}{2}(y^t - \theta x^t)^2 = \frac{1}{4}[0.5 \times (1-0.5)^2 + 0.5 \times (1-1.5)^2 + 0.5 \times (2-2.5)^2 + 0.5 \times (3-2.5)^2] = 0.5 \times 0.25 = 0.125$

• Thus, both line $y = 0.783x$ and line $y = -0.085 + 0.8025x$ give us the lowest $R_n(\theta)$ value. (Theoretically, $y = -0.085 + 0.8025x$ should give us the lowest $R_n(\theta)$ value. However, since in this case the performance of $y = -0.085 + 0.8025x$ and $y = 0.783x$ are really similar, students could choose either of these two or both of them as the predictor or predictors that give us the lowest $R_n(\theta)$ value).

Question 3. (Subtotal: 25 points)
(coding question: linear regression and ridge regression)
This exercise requires the student to understand the basics of linear regression and ridge regression. Please open the .ipynb file to see details of this question.

Question 4. (Subtotal: 25 points)
(coding question: k-means)
This exercise requires the student to understand the basics of k-means algorithm. Please open the .ipynb file to see details of this question.