Homework 1

Part I: Basic Concepts

1. There are two vectors, x_1 and x_2

$x_1 = \begin{bmatrix}1\\2\end{bmatrix}$ and $x_2 = \begin{bmatrix}10\\18\end{bmatrix}$

What is the distance between x_1 and x_2 ? (Please show the steps of the calculations)

(1) if the distance measure is based on L2 norm (a.k.a Euclidean norm)

(2) if the distance measure is based on L1 norm

(3) if the distance measure is based on L∞ norm (a.k.a infinity norm)

In class, we studied the customer segmentation example and tried to find the most valuable customers who have good income but low spend. There are two feature components $x = \begin{bmatrix}income\\spend\end{bmatrix}$ in this application. Assume that we use a clustering method similar to k-mean, and this method could use any type of vector norms as distance measure. Then, does the L∞ norm-based distance measure make sense for this application?

2. We define a scalar valued function of a vector variable

$$f(x) = x^T Ax$$

Here, x is a column vector, x^T is the transpose of x, and A is a symmetric matrix

To simplify this question, let's assume x has only two elements $x = \begin{bmatrix}\alpha\\\beta\end{bmatrix}$, and $A = \begin{bmatrix}a & c\\c & b\end{bmatrix}$

The derivative of f with respect to x is a vector defined by $\frac{df}{dx} = \begin{bmatrix}\frac{df}{d\alpha}\\\frac{df}{d\beta}\end{bmatrix}$

Show that $\frac{df}{dx} = 2Ax$

Hint: calculate $f(x)$, $2Ax$, $\frac{df}{d\alpha}$ and $\frac{df}{d\beta}$

K-means clustering (10 points)

3. Briefly describe the two key steps in one iteration of the k-means algorithm. (1 point)

4. What is the distance measure used in k-means (implemented in sk-learn)? (1 point)

5. The k-means algorithm can always converge in a finite number of iterations. Why? (1 point)

6. The clustering result of k-means could be random. Why? (1 point)

7. The minimum value of the loss function is zero for any dataset. What is the clustering result when the loss function is zero? – assuming that the dataset has millions of different samples. (1 point)

Note: for questions 3,4,5,6,7, you only need to write a few words (bullet points) for each one.

8. find the optimal centers by following the steps below when cluster labels are given. (5 points)

The loss function is $L = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K}\alpha_{(n,i)}\|x_n - c_i\|^2$ as defined in the lecture notes.

First, we calculate $\frac{\partial L}{\partial c_k}$, where k could be 1, 2, 3, …, K.

$$\frac{\partial L}{\partial c_k} = \frac{\partial\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K}\alpha_{(n,i)}\|x_n - c_i\|^2\right]}{\partial c_k} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial\left[\sum_{i=1}^{K}\alpha_{(n,i)}\|x_n - c_i\|^2\right]}{\partial c_k} = \frac{1}{N}\sum_{n=1}^{N}\boxed{\text{(A)}}\frac{\partial\left[\|x_n - c_k\|^2\right]}{\partial c_k}$$

$$\|x_n - c_k\|^2 = (x_n - c_k)^T(x_n - c_k) = x_n^T x_n + c_n^T c_n - \boxed{\text{(B)}}$$

$$\frac{\partial\left[\|x_n - c_k\|^2\right]}{\partial c_k} = \boxed{\text{(C)}}$$

Thus, $\frac{\partial L}{\partial c_k} = \frac{1}{N}\sum_{n=1}^{N}\boxed{\text{(D)}}$

Then, we set $\frac{\partial L}{\partial c_k} = 0$, and we obtain the optimal center $c_k = \frac{1}{N_k}\sum_{n=1}^{N}\alpha_{(n,k)}x_n$, where $N_k = \sum_{n=1}^{N}\boxed{\text{(E)}}$

What are (A), (B), (C), (D), and (E) in the above equations ?

Note: (E) is a variable, not a word or sentence

Part 2: Programming

Complete the tasks in the files:

H1P2T1_kmeans.ipynb

If you want to get some bonus points, try this task:

H1P2T2_kmeans_compression.ipynb

Grading: the number of points

	Undergraduate Student	Graduate Student
Basic Concepts	10	10
K-means clustering	10	10
H1P2T1	30	30
H1P2T2	5 (bonus)	5 (bonus)
Total number of points	50 + 5	50 + 5

The following rules are used for every homework assignment.

Each homework assignment is an individual assignment, NOT a group assignment.

For part 1: You may use MS-word to write the answers, convert the file to PDF, and upload it to Blackboard. You may write the answers on a piece of paper, take a photo using your cell phone, and upload the picture to Blackboard. **Make sure that your handwriting is human/TA-readable**, otherwise you may lose points.

For part 2: complete the ipynb files. Do NOT convert ipynb files to pdf or py or another format.

Upload your files to Blackboard and do not miss any files.

Before you submit homework files, make sure you run each and every code cell of your program files. If a code cell is supposed to generate some output (e.g., figure or text) and nothing shows up below the cell because you forget to run the cell, you will lose the points of that cell.

Do NOT use ChatGPT and Copilot.