EE2211 Tutorial 11

Question 1: The K-means clustering method uses the target labels for calculating the distances from the cluster centroids for clustering.

a) True
b) False

Question 2: The fuzzy C-means algorithm groups the data items such that an item can exist in multiple clusters.

a) True
b) False

Question 3: How can you prevent a clustering algorithm from getting stuck in bad local optima?

a) Set the same seed value for each run
b) Use the bottom ranked samples for initialization
c) Use the top ranked samples for initialization
d) All the above
e) None of the above

Question 4: Consider the following data points: $\boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\boldsymbol{z} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The k-means algorithm is initialized with centers at $\boldsymbol{x}$ and $\boldsymbol{y}$. Upon convergence, the two centres will be at

a) $\boldsymbol{x}$ and $\boldsymbol{z}$
b) $\boldsymbol{x}$ and $\boldsymbol{y}$
c) $\boldsymbol{y}$ and the midpoint of $\boldsymbol{y}$ and $\boldsymbol{z}$
d) $\boldsymbol{z}$ and the midpoint of $\boldsymbol{x}$ and $\boldsymbol{y}$
e) None of the above

Question 5: Consider the following 8 data points: $\boldsymbol{x}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\boldsymbol{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \boldsymbol{x}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{x}_4 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{x}_5 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \boldsymbol{x}_6 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \boldsymbol{x}_7 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$ and $\boldsymbol{x}_8 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$. The k-means algorithm is initialized with centers at $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 3 \\ 0 \end{bmatrix}$. The first center after convergence is $\boldsymbol{c}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$. The second centre after convergence is $\boldsymbol{c}_2 = \begin{bmatrix} blank1 \\ blank2 \end{bmatrix}$ (up to 1 decimal place)

(K-means Implementation on 2D data)

Question 6:

Generate three clusters of data using the following codes.

```
## Import necessary libraries
import random as rd
import numpy as np # linear algebra
from matplotlib import pyplot as plt
## Generate data
## Set three centers, the model should predict similar results
center_1 = np.array([2,2])
center_2 = np.array([4,4])
center_3 = np.array([6,1])
## Generate random data and center it to the three centers
data_1 = np.random.randn(200, 2) + center_1
data_2 = np.random.randn(200,2) + center_2
data_3 = np.random.randn(200,2) + center_3
data = np.concatenate((data_1, data_2, data_3), axis = 0)
plt.scatter(data[:,0], data[:,1], s=7)
```

(i) Implement the Naïve K-means (the basic/standard algorithm shown in lecture) clustering algorithm to find the 3 cluster centroids. Classify the data based on the three centroids found and illustrate the results using a plot (e.g., mark the 3 clusters of data points using different colours).

(ii) Change the number of clusters K to 5 and classify the data points again with a plot illustration.

(K-means Classification of iris data, 4D input features)

Question 7:

Load the iris data "`from sklearn.datasets import load_iris`". Assume that the class labels are not given. Use the Naïve K-means clustering algorithm to group all the data based on K=3. How accurate is the result of clustering comparing with the known labels?