Artificial Intelligence Lab 5 : Clustering and Classification

1 K-means algorithm [50 Marks]

Q1) Aim is to play around with the K-means (given in Algorithm 1) algorithm by running it for different inputs and different initialisation schemes as below (list is not exhaustive, you are encouraged to find your own combination).

Basic setting: Let n be the total number of points and let groups be from $1, \ldots, m$. Generate $\frac{n}{m}$ observations (from an appropriate distribution) for each group , combine them to form the total data set $(x_i)_{i=1}^n$ containing n observation points.

1-d observations: Pick $m = K = 2$. Observations are sampled from uniform $([-1, 1])$ for group 1, and uniform $[-0.5, 1.5]$ for group 2. Let us run K-means for the following cases:

1. Start from $\bar{x}_1 = -1$, and $\bar{x}_2 = 1.5$. [5 Marks]
2. Start from $\bar{x}_1 = -0.1$, and $\bar{x}_2 = 0.1$. [5 Marks]
3. Start from $\bar{x}_1 = 0$, and $\bar{x}_2 = 1$. [5 Marks]
4. Start from $\bar{x}_1 = 0$, and $\bar{x}_2 = 3$. [5 Marks]

1-d observations: Pick $m = K = 3$. Observations are sampled from uniform $([-1, 1])$ for group 1, and uniform $[-0.5, 1.5]$ for group 2, uniform $[1, 2]$ for group 3. Let us run K-means for the following cases:

1. Show some initialisations where the cluster are found correctly. [5 Marks]
2. Show some initialisations where the cluster are **not** found correctly. [5 Marks]

1-d observations: Pick $m = 3$ $K = 2$. Observations are sampled from uniform $([-1, 1])$ for group 1, and uniform $[-0.5, 1.5]$ for group 2, uniform $[1, 2]$ for group 3. Show how clusters are formed. [5 Marks]

2-d observations: Let there be two groups namely *Kids* and *Adults*. The features are 2-dimensional, i.e., they have $x_i = (x_i(1), x_i(2))$, where $x_i(1)$ stands for height and $x_i(2)$ stands for weight.

1. Choose uniform distribution for height and weight. Use appropriate ranges for the two groups. Choose different K values, and show the output. [10 Marks]
2. Choose normal distribution for height and weight. Use mean and variance in normal distribution for the two groups. Choose different K values, and show the output. [5 Marks]

Note: Use dots for observation data and display the cluster centre with star symbol. Use different colours for different clusters and display them at each iteration. The colours should be based on the nearness to the centres (**not based on the original group form which they were generated**). This way we can visually see at each iteration how the cluster centre changes.

Algorithm 1 K-means Algorithm

Input: Data $(x_i)_{i=1}^n \in \mathbb{R}^d$, and an initial guess for the K centres $\{\bar{x}_1, \ldots, \bar{x}_K\}$
while Centres not converged **do**
  for $k = 1, \ldots, K$ **do**
    Collect C_k the set of points nearby to centre $\bar{x}_k$ as follows:
    $C_k = \{x_i \colon \|x_i - \bar{x}_k\|_2^2 < \|x_i - \bar{x}_{k'}\|_2^2, k' = 1, \ldots, K, k' \neq k\}$
    if $|C_k| > 0$ **then**
      Update centre $\bar{x}_k = \frac{1}{|C_k|} \sum_{i=1}^n x_i \cdot \mathbb{1}_{\{x_i \in C_k\}}$
    end if
  end for
end while

2 K-nearest neighbour algorithm [50 Marks]

There are two states $S = s^1 = Kid, s^2 = Adult$, with $P(Kid)$ and $P(Adult)$ respectively. Sample $s_t \overset{iid}{\sim} P$. The observation $o_t \sim p(\cdot|s_t)$. The observation has two co-ordinates namely height and weight and is given by $o_t = (h_t, w_t)$, where h stands for height and w stands for weight. Let w_t and h_t be independent of each other. Generate n state observation pairs $(o_t, s_t)_{t=1}^n$, and create the dataset $(x_t, y_t)_{t=1}^n$, where $x_i = o_i, \forall i = 1, \ldots, n$ and $y_t = +1$ (if $s_t = Adult$) and $y_t = -1$ (if $s_t = Kid$). Implement the k-nearest neighbour algorithm to classify a new point $x_{new} \in \mathbb{R}^2$ as Kid or Adult

1. First use uniform distribution (of appropriate range) and generate n (say 100) points. Demonstrate that you understand what happens with various k values, and various ranges for the uniform distribution. [25 Marks]
2. Use normal distribution (of appropriate parameters namely mean and variance) and generate n (say 100) points. Demonstrate that you understand what happens with various k values, and various parameters of the normal distribution. [25 Marks]