Statistical Machine Learning: Coursework 4

General Instructions

- Work in groups of at least one and at most two.
- Include your student number and your name on your assignment.
- In your solutions, you should present your R output or snippets of R code as you deem appropriate. Make sure to present your results clearly.
- You should submit your work as a single PDF file on Blackboard by Thursday the 13^{th} March at 12 noon.

Additional Information

- There are **40 marks** available in total. Your mark out of 40 will be scaled by 2.5 to give the mark out of 100 that you see on Blackboard.

Problem 1: Theory (20 marks)

In this half of the assignment, we shall investigate some theoretical aspects of the K-Means approach to Clustering, unpacking some of the details which underpin the derivation of both the K-Means objective function and Lloyd's Algorithm.

(i) **(5 marks)** Let $X = \{x_1, x_2, \cdots, x_n\} \subseteq \mathbb{R}^p$. With $\|\cdot\|$ denoting the standard Euclidean norm,

(a) Prove that for any $m_1 \in \mathbb{R}^p$, there holds the inequality

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \|x_i - x_j\| \leq 2 \cdot N \cdot \sum_{\ell=1}^{N} \|x_\ell - m_1\|.$$

(b) Defining $m_2 \in \mathbb{R}^p$ to be the sample mean of the points in X, prove the *equality*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \|x_i - x_j\|^2 = 2 \cdot N \cdot \sum_{\ell=1}^{N} \|x_\ell - m_2\|^2.$$

(ii) **(5 marks)** Let K be a natural number, fix a vector $c \in \mathbb{R}^K$, and consider the minimisation problem

$$\begin{aligned} &\text{minimise } \sum_{k=1}^{K} c_k z_k \\ &\text{such that } z \in \text{OneHot}\,(K)\,. \end{aligned}$$

Without a loss of generality, we assume that the entries of c are ordered as $c_1 \leq c_2 \leq \cdots \leq c_K$. Identify the full set of solutions to this problem, carefully justifying why this set includes all possible solutions and nothing else. (Hint: you may wish to consider the set $\mathcal{K} = \{k : 1 \leq k \leq K, \text{ and } c_k = c_1\}$.)

(iii) **(10 marks total)** Suppose that we have obtained a solution to the K-Means Problem, with $K \geq 3$, i.e. we decompose X into

$$X = C_1 \sqcup C_2 \sqcup \cdots \sqcup C_K$$

where for $1 \leq k \leq K$, each C_k is a subset of X containing $N_k \geq 1$ members, with cluster mean μ_k, i.e.

$$\frac{1}{N_k} \sum_{x_i \in C_k} x_i = \mu_k.$$

Assume also that you have also computed the within-cluster sum-of-squares for each cluster, i.e. for each k, you have access to

$$V_k := \sum_{x_i \in C_k} \|x_i - \mu_k\|^2.$$

Given our clustering into K components, we could try to use it to build a nice initialisation for finding a good clustering into $(K-1)$ components by strategically 'merging' two similar clusters.

(a) **(5 marks)** Let $1 \leq k \neq \ell \leq K$, and let C_k, C_ℓ denote the corresponding clusters. Consider then *merging* these clusters, i.e. defining $C'_{k,\ell} = C_k \sqcup C_\ell$. In terms of $\{N_k, \mu_k, V_k, N_\ell, \mu_\ell, V_\ell\}$, write down (with justification) simple formulas for

(α) $\mu'_{k,\ell}$:= the mean of the cluster $C'_{k,\ell}$, and

(β) $V'_{k,\ell}$:= the within-cluster sum-of-squares for the cluster $C'_{k,\ell}$.

(b) **(5 marks)** Upon merging the clusters k and ℓ, the total value of the K-Means objective will increase by some value $\Delta_{k,\ell} \geq 0$. Using your answer to part (a),

(α) Estimate the computational cost of computing $\Delta_{k,\ell}$ for a single pair (k, ℓ), and hence

(β) Estimate the computational cost of finding the 'optimal' pair of clusters to merge, in terms of increasing the objective function as little as possible.

Both answers should be presented in Big-Oh notation, and should depend only on the values of (N, K, p). (You may wish to revisit the 'Big-Oh Notation' supplementary notes which are available on the 'Lecture Notes' section of the unit Blackboard page).

Problem 2: Practical (20 marks)

In this half of the assignment, we will apply K-Means Clustering to some data, and examine some of its behaviours. The dataset we shall use for this task, `moons.csv`, can be found on Blackboard, alongside this document.

(i) **(10 marks)** Load in `moons.csv`, and use the native `kmeans` function in R to cluster the data into $K = \{2, 3, 4, 5\}$ components, according to each of the following loss functions:

$$\begin{aligned}
\text{Loss}^{(1)}(Z, \mu) &:= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \left| x_{i,1} - \mu_k \right|^2 \\
\text{Loss}^{(2)}(Z, \mu) &:= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \left| x_{i,2} - \mu_k \right|^2 \\
\text{Loss}^{(3)}(Z, \mu) &:= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \left\| x_i - \mu_k \right\|^2 \\
\text{Loss}^{(4)}(Z, \mu) &:= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \left\| g(x_i) - \mu_k \right\|^2,
\end{aligned}$$

where $g : \mathbb{R}^2 \to \mathbb{R}^2$ is the function given by

$$g(x_1, x_2) = \left(\sqrt{\left(x_1 + \frac{1}{2} \right)^2 + x_2^2}, \sqrt{\left(x_1 - \frac{1}{2} \right)^2 + x_2^2} \right),$$

and the vectors z_i are subject to the usual one-hot constraints. You may find it useful to plot the results of the clusterings which you obtain, colouring the points of each cluster differently.

(ii) **(5 marks)** Write a function `cluster_eval` which, for a given data set X and a range of natural numbers $K_1 \leq K_2$,

(α) Computes the K-Means clusterings for each $K_1 \leq K \leq K_2$, and

(β) Given this collection of clusterings, plots the values of the K-Means loss function at the optimal clusterings, as a function of K.

Use your implementation of `cluster_eval` to compare the clusterings which you obtained in Part (i) of this problem. Do you find that your different clusterings agree on what the 'correct' number of clusters in this data are?

For this question, you may still make use of the native `kmeans` function to compute the clusterings themselves, but you should write the rest of the function *from scratch*. The function itself does not need to output the clusterings, only the plot of the loss function values.

(*iii*) **(5 marks)** Load in R's `quakes` dataset. This dataset stores information about 1000 seismic events (i.e. earthquakes and similar) which have occurred in the vicinity of the island of Fiji over the past 60 or so years. The first three columns (`lat`, `long`, `depth`) record the geographical location of the event, the fourth (`mag`) records the strength of the event (as measured on the Richter scale), and the the fifth (`station`) records how many stations documented the event in question.

In the context of *this question*, we are interested in whether the locations at which these seismic events take place are *spatially localised*, i.e. if they cluster around specific areas of the island.

(α) Run a standard K-Means clustering on the full data set for $K = \{2, 3, 4, 5\}$, and plot the outcome, focusing on the `lat` and `long` dimensions.

(β) Does the clustering which you obtain in this way look spatially reasonable? Propose an alternative approach to obtaining a clustering which is more coherent (still using K-Means).