Assignment 4

Wrestling with Data in an Unsupervised Fashion

Homeworks Guidelines and Policies

- ***What you must hand in.*** It is expected that the students submit an assignment report (HW4_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW4_[student_id].zip).
- ***Pay attention to problem types.*** Some problems are required to be solved *by hand* (shown by the ✎ icon), and some need to be implemented (shown by the icon).
  Please don't use implementation tools when it is asked to solve the problem by hand, otherwise you'll be penalized and lose some points.
- ***Don't bother typing!*** You are free to solve by-hand problems on a paper and include picture of them in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- ***Reports are critical.*** Your work will be evaluated mostly by the quality of your report. Don't forget to explain what you have done, and provide enough discussions when it's needed.
- ***Appearance matters!*** In each homework, 5 points (out of a possible 100) belongs to compactness, expressiveness and neatness of your report and codes.
- ***Python is also allowable.*** By default, we assume you implement your codes in MATLAB. If you're using Python, you have to use equivalent functions when it is asked to use specific MATLAB functions.
- ***Be neat and tidy!*** Your codes must be separated for each question, and for each part. For example, you have to create a separate .m file for part b. of question 3. Please name it like p3b.m.
- ***Use bonus points to improve your score.*** Problems with bonus points are marked by the ⭐ icon. These problems usually include uncovered related topics or those that are only mentioned briefly in the class.
- ***Moodle access is essential.*** Make sure you have access to Moodle because that's where all assignments as well as course announcements are posted on. Homework submissions are also done through Moodle.

- ***Assignment Deadline.*** Please submit your work **before the end of February 12th**.
- ***Delay policy.*** During the semester, students are given 7 free late days which they can use them in their own ways. Afterwards there will be a 25% penalty for every late day, and no more than three late days will be accepted.
- ***Collaboration policy.*** We encourage students to work together, share their findings, and utilize all the resources available. However you are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- ***Any questions?*** If there is any question, please don't hesitate to contact us through the following email address: **ali.the.special@gmail.com** and **fardin.aiar@gmail.com**.

1. Practicing the Fundamentals of Clustering (15 Pts.)

Keywords: *Clustering Problem, Unsupervised Learning, K-Means Clustering, Hierarchical Clustering, Dendrogram*

Clustering algorithms are probably the most common examples of **Unsupervised Learning** methods, in which the goal is to group a set of objects in such a way that objects in the same group, i.e. **Cluster**, are more similar according to a certain metric. One of the most popular methods of clustering is **K-Means**, where the clustering process of an observation is done based on the nearest clusters mean. **Hierarchical Clustering**, on the other hand, is a different method of clustering which tries to build a hierarchy of clusters, usually presented in a **Dendrogram**.

In this problem, we are going to walk through the clustering process by solving some examples by hand. First, assume the dataset X as follows:

$$X = \begin{bmatrix} 0.4 & 1.5 & 1.6 & 2.2 & 2.4 & 2.8 & 3.4 & 3.8 & 4.0 & 4.2 \\ 4.5 & 0.8 & 1.4 & 4.2 & 1.0 & 0.2 & 1.8 & 3.6 & 2.3 & 2.9 \end{bmatrix}^T$$

You are asked to perform a K-Means clustering on the given dataset, where k is set to 3 and the centres of 3 clusters are initialised as $\mu_1 = (2.2,2.8)$ (red), $\mu_2 = (2.0,2.0)$ (blue), $\mu_3 = (1.2,3.0)$ (green). Use the Euclidean distance as the distance function.

a. What would be the centre of the cluster 1 (red) after the first iteration?
b. What would be the centre of the cluster 2 (blue) after the second iteration?
c. What would be the centre of the cluster 3 (green) when the algorithm converges?
d. How many iterations are required for the algorithm to converge?

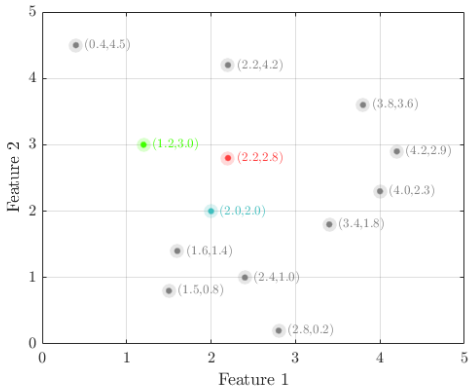

Figure 1 Scatter plot of the dataset points (black) as well as the initial centre points (colored).

Now, assume you apply K-Means with $k = 2$ clusters on the following set of 1-D points: $\{-3, 1, 8\}$

e. Considering the fact that K-Means can get stuck in local optima, describe the precise conditions on the initialisation of $\mu_1 \in \mathbb{R}$ and $\mu_2 \in \mathbb{R}$ such that running K-Means will yield the global optimum of the objective function. Assume $\mu_1 \leq \mu_2$, and if in the first step of K-Means, no points are assigned to a cluster j, then in the next step μ_j is set to ∞.
Hint: To get some intuition, try running K-Means with several different initialisations.

Finally, assume the given similarity matrix below:

	P₁	**P₂**	**P₃**	**P₄**	**P₅**	**P₆**
P₁	1.0	0.52	0.47	0.52	0.31	0.37
P₂	0.52	1.0	0.39	0.53	0.47	0.46
P₃	0.47	0.39	1.0	0.28	0.40	0.44
P₄	0.52	0.53	0.28	1.0	0.50	0.38
P₅	0.31	0.47	0.40	0.50	1.0	0.44
P₆	0.37	0.46	0.44	0.38	0.44	1.0

f. Perform a single linkage hierarchical clustering.
g. Perform a complete linkage hierarchical clustering.
h. Perform an average linkage hierarchical clustering.

2. Breast Cancer Prevention using K-Means Algorithm (18 Pts.)

Keywords: *Unsupervised Learning, Clustering Problem, K-Means Method*

Breast cancer is known as the most frequent type of cancer worldwide, contributing 12.5% of the total number of new cases detected in 2020. Despite being a lethal disease, it can be diagnosed at early stages by regular check-ups and experiments. One effective technique for recognizing breast cancer is named Fine Needle Aspiration (FNA), in which a syringe is used to take a very small tissue sample, from which an image is taken. Then a clinician attempts to isolate individual cells in each image to obtain 30 characteristics, such as size, shape, and texture. Here, our goal is to use K-Means to cluster these features to diagnose breast cancer based on the samples of Wisconsin Diagnostic Breast Cancer dataset.

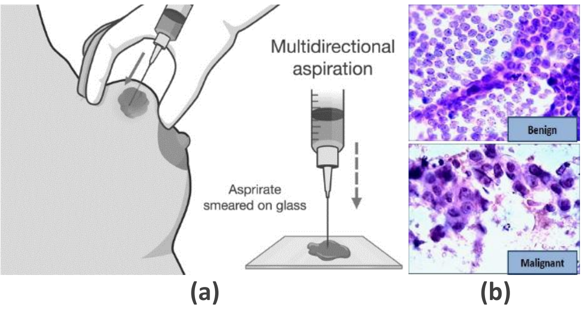

Figure 2 The process of Fine Needle Aspiration using ultrasound. (a) Taking sample from breast tissue. (b) Two benign and malignant breast samples obtained by FNA.

a. Implement a function `C = kmeanscluster(X,k,mu,tol,maxIter)`, in which `X` is the data matrix, `k` represents the number of clusters, `mu` denotes the matrix of initial centres, `tol` indicates the tolerance parameter for convergence, and `maxIter` is the maximum number of iterations. The clustering result is also stored in matrix `C`. Use your function to cluster the data, and find the accuracy of the clustering.

b. Run your function 5 times using different starting points, and calculate the accuracy of each case. What were your observations? Explain.

c. Run your function using the provided initial centres (init_mu.mat), in which each column represents one of the initial centres, and report the accuracy of the clustering.

d. What happens if you initialize with the *true* centres, obtained after the *true* clustering?

e. Can you achieve better accuracy using another unsupervised learning method? What about a supervised one? Explain.

3. More Into Clustering: K-Medoids, Elbow Methods, and Beyond (20 Pts.)

Keywords: *Unsupervised Learning, Clustering Problem, K-Means, K-Medoids, Hierarchical Clustering, Dendrogram, Elbow Method, Purity Measure, Normalised Mutual Information (NMI)*

Following the previous problem, here you are going to investigate **Clustering** more in detail. You are also going to get familiar with several new clustering concepts. In the first part, you are working with a randomly generated two-dimensional dataset “rand.mat”, shown in Figure 3.

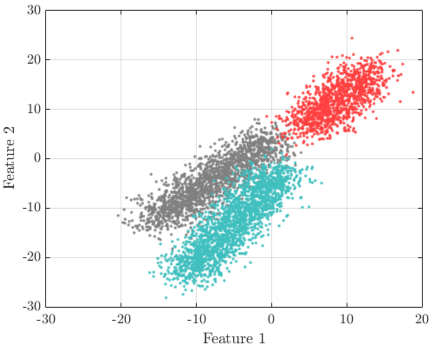

Figure 3 The dataset contains 5000 samples in three different categories.

a. Perform clustering on the given data by applying K-Means method with parameter k set to 2, 3, 5 and 7. In each case, plot the clustered data and highlight each cluster with a specific color. Also, indicate the final centres with a unique shape on your plot.

b. In more complicated datasets, finding a proper value for k is often challenging. There are various methods one can use to determine a good value for k, such as **Elbow Method**, which is probably the most known method in this area. Run the algorithm for values of $k = 10,\dots,1$, and plot the k values against the **Sum of Squared Errors** (**SSE**). How can this plot help you finding a good value for k? What would be that value?

c. **K-Medoids** is one of the many variations of K-Means, which is fundamentally similar to it, with only slight differences. Repeat part a. for the case that K-Medoids clustering is applied.

Next, read the dataset “gene_ex.csv“, which contains the data on 40 tissue samples with measurements on 1000 genes. The first 20 samples are from healthy patients, while the second 20 are from diseased patients.

d. Perform hierarchical clustering on the samples, once using Euclidean and once using correlation-based distance. Plot the dendrogram. Do the genes separate samples into two groups? Do your results depend on the type on linkage used?

e. We would like to know which genes differ the most across two groups. Try to answer this question.

Finally, you are to evaluate the clustering performance of K-Means versus hierarchical methods on the famous two-spiral dataset. It contains 2000 samples, each with sample coordinate points and the corresponding cluster label, and can be found in the file “two_spiral.mat“.

f. Compare K-Means method with a hierarchical clustering method in terms of **Purity** and **Normalised Mutual Information** (**NMI**) measures. For hierarchical clustering, use a bottom up way and for cluster merging use both single linkage and complete linkage methods.
Note: You are free to use ready-to-use codes (like this) for clustering evaluation measures.

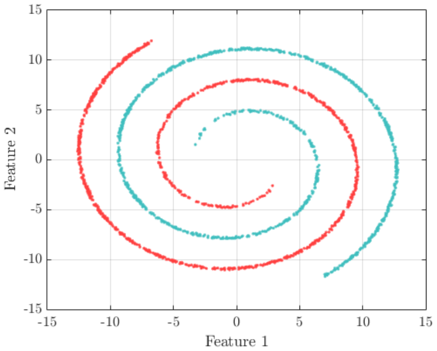

Figure 4 The distribution of the samples in two-spiral dataset.

Note: Please use your own implementation of clustering algorithms.

Recommended MATLAB functions: `csvread(), dist(), cor(), dendrogram()`

4. How Good People Can Draw? (32 Pts.)

Keywords: *Dimensionality Reduction, Clustering Problem, Principal Component Analysis (PCA), K-Means Method*

QuickDraw is an online game developed by Google in which the players are challenged to draw a picture of an object, then attempts to guess what the drawings represent using a neural network artificial intelligence. The AI is capable of learning from each drawing and therefore increasing its accuracy in the future.

Figure 5 In QuickDraw, player has to draw the given objects as much accurate as possible

Using the drawings made by millions of players, a dataset with 50 million drawings from 345 objects has been created. Here, a subset of this gigantic dataset, QuickDraw10, is going to be used for an image clustering task.

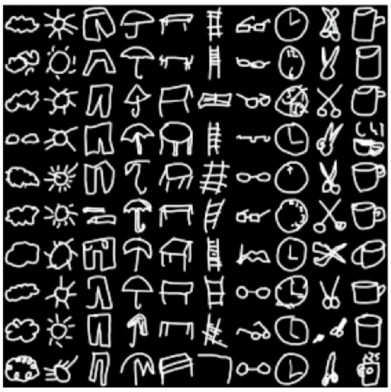

Figure 6 QuickDraw10 is a subset of QuickDraw dataset, a commonly used alternative for the MNIST dataset.

a. Perform PCA on the dataset to reduce its dimensions. Display the top 20 eigenvalues as well as the representation of the samples projected onto their first two and three principal components.

b. Now use LDA to project the samples onto their first and second linear discriminants, and display the results.

c. Perform K-Means on the data using their first two principal components. Select random initial centroids and set $K = 3, 7, 10$. Display the result of clustering with final centres highlighted, and compare the results together and with the plot obtained in part (a).

d. Repeat part (c) considering the initial centroids as the mean of the samples such that for $K = 3$, initial centroids are the mean of samples of classes {1,3,5,7}, {2,4}, and {6,8,9,10}, for $K = 7$, initial centroids are the mean of samples of classes {1,2,5}, {3}, {4}, {6}, {7}, {9} and {8,10}, and for $K = 10$, initial centroids are the mean of sample of each of the 10 classes separately. Display and compare the results with those of the previous parts. Also comment on your observations.

e. As can be seen, clustering the data using their first two principal components doesn't produce satisfactory results. We are going to increase the clustering accuracy by considering more principal components. By trial and error find the number of principal components which is enough to capture 0.95 of the data variance. Reconstruct three arbitrary samples using these principal components and compare the results with their corresponding original images.

f. Use K-Means to divide data into 10 clusters. Set the required parameters appropriately. Display 10 samples of each cluster randomly (100 in total) and comment on the results.

g. Draw 10 bar graphs for each cluster, each representing the distribution of the samples of different classes (in percentage). Compare different clusters using these bar graphs.

h. Finally, visualize the clustering by keeping only 2 and 3 features and displaying the corresponding scatter plots.

5. Some Explanatory Questions (10 Pts.)

Please answer the following questions as clear as possible:

a. How can we relate clustering algorithms with instance-based learning algorithms like K-NN?

b. What influences can noise and outliers have on the performance of the K-Means?

c. Is it possible for the K-Means algorithm to revisit a state? Explain how your answer can be used to prove that the K-Means converges in a finite number of iterations.

d. Assume a clustering task using K-Means with a specific parameter k and samples number n. Under what conditions does the algorithm converge to a result with more than k clusters? What about a result with less than k clusters?

e. Under what circumstances can hierarchical clustering algorithms outperform K-Means algorithm? What about the reverse?

Good Luck!
Ali Abbasi, Fardin Ayar