Statistical Pattern Recognition

Amirkabir University
of Technology
(Tehran Polytechnic)

Fall 2022

Assignment 4

Overview of Dimensionality Reduction and Supervised Learning

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 do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- ***Don't bother typing!*** You are free to solve by-hand problems on a paper and include their pictures 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. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- ***Appearance matters!*** In each homework, 5 points (out of a possible 100) belong 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 are using Python, you have to use the 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, which must be named '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 is where all assignments as well as course announcements are posted. Homework submissions are also made through Moodle.

- ***Assignment Deadline.*** Please submit your work **before the end of January 30th**.
- ***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 do not hesitate to contact us through the following email addresses: **ali.the.special@gmail.com** and **ebp.mohsen@gmail.com**.

1. Understanding the Behavior of Clustering Techniques (15 Pts.)

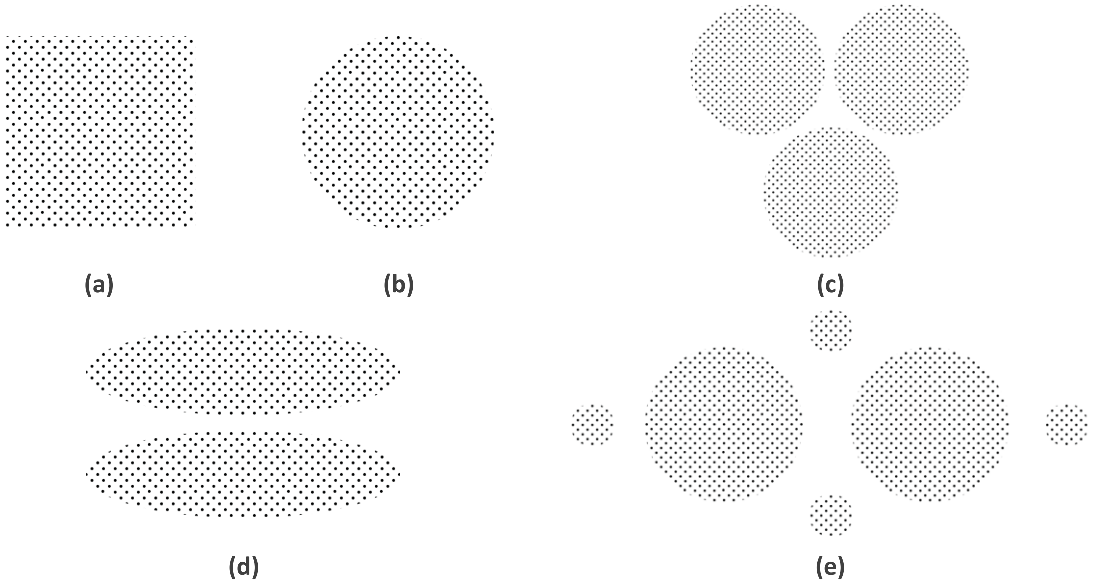
Keywords: *Unsupervised Learning, Clustering Problem, K-Means Clustering*

Another type of machine learning algorithms lie under the concept of **Unsupervised Learning**. These methods make inferences from data using only inputs without referring to known or labelled outputs. **Clustering** – known as an unsupervised method – is the attempt of assigning objects to different groups, or **Clusters**, so that those in the same group are more similar to each other than those in other groups. One of the most popular clustering algorithm is **K-Means**. It keeps *k* **Centroids** that it uses to define clusters. In K-Means, a point is considered to be in a certain cluster if it is closer to that cluster's centroid than any other centroid.

This problem consists of several parts which aim to evaluate your basic understanding of clustering, mainly K-Means method.

First, you are given five different sets of 2-D points in Figure 1, and you are asked to provide a sketch of how K-Means would split them into clusters considering the given number of clusters. You must also indicate approximately where the final centroids would be.

a. $K = 2$
b. $K = 3$
c. $K = 2$
d. $K = 3$
e. $K = 3$



(a) (b) (c)

(d) (e)

Figure 1 Sets of 2-D points provided for the first part of the problem 1

Note 1: Assume that the squared error objective function is being used.

Note 2: If there is more than one possible solution, then please specify for each solution whether it is a global or local minimum.

Note 3: Images in Figure 1 are given to you in "P1" directory attached to this assignment. You can use them in your report.

Next, consider the diagrams in Figure 2.

f. In which one of the two diagrams do the classic clustering techniques, like single linkage, find the patterns represented by door and windows?

g. Specify the limitations that clustering has in detecting the patterns formed by points.

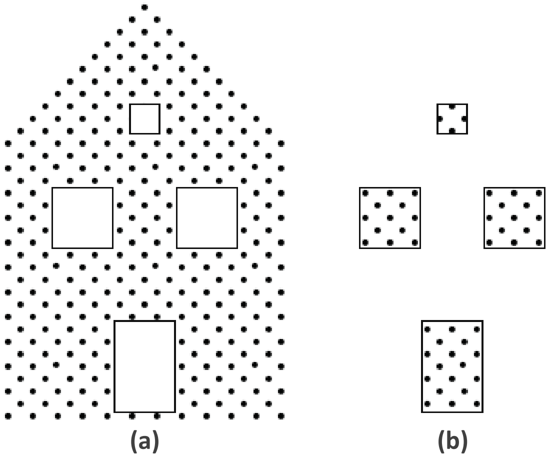


Figure 2 The goal of this clustering problem is to distinguish the main parts, i.e. door and windows

Now, imagine there are five datasets, noted as a, b, c, d and e. The datasets are clustered using two different methods which one of them is K-Means. The distance measure used here is the Euclidean distance. Results are shown in Figure 3.

h. Considering dataset (a) of the Figure 3, determine which result (1 or 2) is more likely to be generated by K-Means method.

i. Repeat the part h for dataset (b).

j. Repeat the part h for dataset (c).

k. Repeat the part h for dataset (d).

l. Repeat the part h for dataset (e).

Hint: You must check out the state when K-Means converges. Centres for each cluster are shown by red circles.

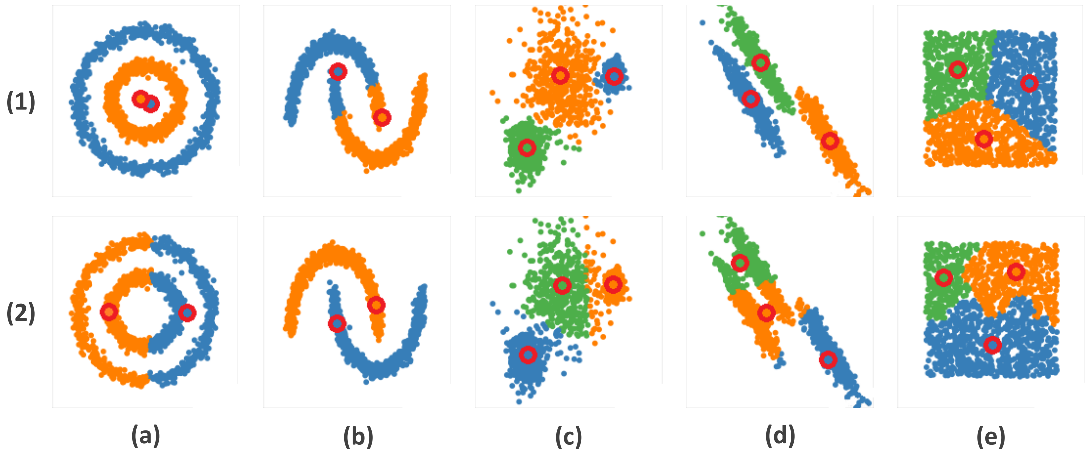


Figure 3 The clustering results of five datasets, each with different data distributions. Final centres are shown by red circles.

Finally, consider the points in Figure 4.

m. Determine all well-separated clusters in the given set of points.

Note: You may use the equivalent image of Figure 4 placed in "P1" directory.

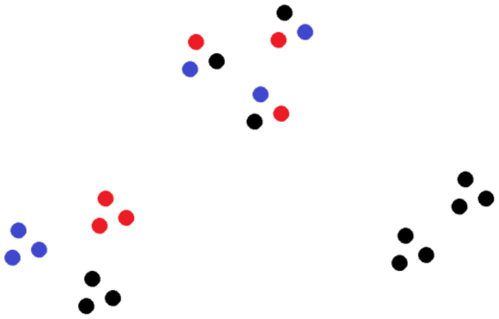


Figure 4 Clustering these points may yield to different results depending on the method and settings considered

2. K-Means and Beyond: Solving Clustering Problems by Hand (14 Pts.)

Keywords: *Unsupervised Learning, Clustering Problem, K-Means, K-Median, K-Medoids, Hierarchical Clustering, Single-Linkage, Complete Linkage, Dendrogram*

K-Means clustering method has several variations, each structurally similar to the main algorithm but with slight differences. **K-Medians**, as an example, calculates the median in order to determine the centroids. However, there is another group of clustering strategies called **Hierarchical Clustering**, which also attempt to partition similar objects into same groups, but by building a hierarchy of clusters, known as **Dendrogram**.

In this part, you are to solve some basic clustering problems. First, consider the following toy dataset consisting of six points:

	x	y
1	1.00	2.00
2	2.00	3.00
3	3.00	1.00
4	4.00	4.00
5	5.00	2.00
6	5.00	4.00

a. Assuming points 1 and 6 as the initial centres of clusters A and B respectively, use K-Means method to determine the final clustering result.

b. Now consider points 3 and 4 as the initial centres of clusters A and B respectively, and repeat the previous part.

c. Compare these two clustering results based on SSE measure and determine which one is better.

Now for hierarchical clustering, assume the following set of points:

	x	y
1	0.15	0.45
2	0.20	0.35
3	0.25	0.70
4	0.45	0.15
5	0.75	0.05
6	0.85	0.20
7	0.90	0.85

d. Draw a sketch of the hierarchical clustering tree (dendrogram) we would obtain for single linkage method, considering Euclidean distance measure.

e. Repeat the previous part for complete linkage method.

3. Clustering Football Players: Is Mbappé More Similar to Ronaldo Than Messi? (20 Pts.)

Keywords: *Clustering Problem, Supervised Learning, K-Means Clustering*

There are thousands of professional football players across the world, each with specific skills and unique playing style. Given their quantitative playing attributes, however, it might be possible to determine players with similar playing styles, and find groups of players with (almost) identical qualities.

To accomplish this task, we make use of *FIFA 23* player ratings. The dataset contains the information corresponding to over 18000 football players, each with 87 various features. Our goal is to investigate how these players can be assigned to distinctive clusters.

First, we assume the players with a rating above 85 (91 players in total), and discard all non-numeric features.

Figure 5 FIFA is the most popular football simulation video game series in the world, in which each football player is given a set of ratings corresponding to different skills, e.g., dribbling.

a. Normalize the data, and apply PCA to them so that the dimensions are reduced to 2.
b. Assuming $k = 5$, perform k-means clustering. Visualize clusters with players' names attached to each point.
c. Is Kylian Mbappé playing style more comparable to that of Lionel Messi or Cristiano Ronaldo? How about Mehdi Taremi?

Now, we wish to see how accurate this approach is in categorizing football players based on their positions. As can be seen in the dataset, each player is given a 'Best Position' attribute, denoting his most preferred position on the pitch.

d. Perform clustering on all the players with $k = 16$ (since there are 16 distinct positions listed), and calculate the clustering accuracy. Which position is clustered more accurately?

4. K-Means as a Powerful Image Processing Tool (18 Pts.)

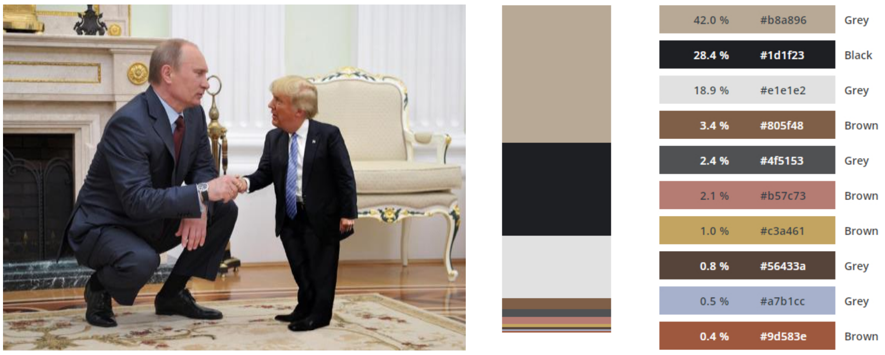
Keywords: *K-Means Clustering, Vector Quantisation, Color Extraction, Image Segmentation, Image Compression*

Despite its simplicity, **K-Means** can be applied to various machine learning tasks. From document classification and data analysis to fraud detection and collaborative filtering, this algorithm still challenges even newly introduced methods in different applications, which many of them are known to be state-of-the-art.

The goal of this problem is to get you more familiar with some of the thing you can do with K-Means in the area of image processing. Given below are three different, but structurally similar image processing tasks and you are required to propose a method to solve them using K-Means.

a. **Color Extraction** is the process of identifying and extracting key colors in images. It gives a better visual understanding of images while providing significant features for other computer vision tasks.

Load the image “trump_tie_1.jpg“. Use K-Means to extract its 3, 5, 7, and 9 main colors. Display these colors properly in separate square shapes.

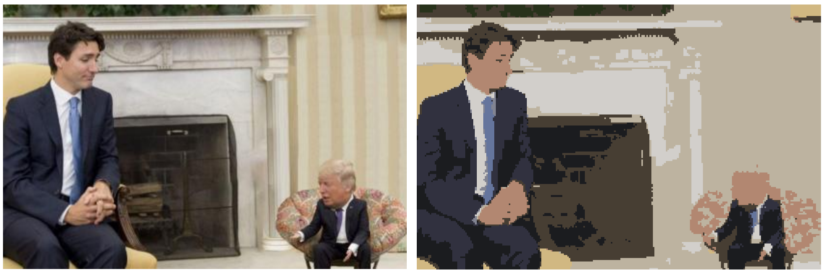


42.0 %	#b8a896	Grey
28.4 %	#1d1f23	Black
18.9 %	#e1e1e2	Grey
3.4 %	#805f48	Brown
2.4 %	#4f5153	Grey
2.1 %	#b57c73	Brown
1.0 %	#c3a461	Brown
0.8 %	#56433a	Grey
0.5 %	#a7b1cc	Grey
0.4 %	#9d583e	Brown

(a) **(b)**

Figure 6 An example image with its extracted main colors, called “palette”, sorted by area they occupy in the input image. The result is obtained using an online tool called TinEye (here) (a) Original image (b) Color extraction results.

b. **Image Segmentation** is a common technique in image processing in which the goal is to divide an image into multiple parts or regions, often based on the characteristics of the pixels in the image.
Load the image “trump_tie_2.jpg“. Use K-Means to divide the given image into 3, 5, 7, and 9 partitions.



(a) **(b)**

Figure 7 Image segmentation applied to an example input image (a) Original image (b) The result of image segmentation

c. **Image Compression** refers to techniques used for minimising the size of an image using the image data which are repeated in the image.
Load the image “trump_tie_3.jpg“. Use K-Means to reduce the size of the input image to %50, %75, %90, and %97 of the original image size (in KB).
Hint: You must find appropriate values for parameter k .

(a) **(b)**

Figure 8 A compression technique has reduced the image size from 211KB to 83KB. Note that the difference is not properly noticeable here (a) Original image (b) Compressed image.

5. Investigating Twitter Reaction to Breaking News (18 Pts.)

Keywords: *Clustering Problem, Text Clustering, K-Means Method, Jaccard Distance*

Twitter is a rich source of data for opinion mining, sentiment analysis and truth discovery. However, one of the biggest problems which many Twitter-based applications encounter with is data redundancy caused by the fact that Twitter users often post similar tweets (e.g. using retweet function) when it comes to popular topics and events. Therefore, clustering similar tweets together would definitely produce more accurate results.

We take into consideration a series of tweets posted during the Boston Marathon Bombing event in April 15, 2013. During this terrorist plot, misinformation spread widely despite efforts by users and experts to correct rumors which were inaccurate.

The Boston Globe
@BostonGlobe

Follow

BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.

2:59 PM - 15 Apr 2013

10,104 RETWEETS 294 FAVORITES

Figure 9 The way rumors and false news were circulated in Twitter during Boston Marathon Bombing has been the subject of many scientific researches and studies

In order to compare different tweets and measure their dissimilarity, we consider *Jaccard distance*. Given two sets A and B, *Jaccard index* is defined as the size of the intersection divided by the size of the union of their samples:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Then the Jaccard distance can be obtained by subtracting Jaccard index from 1:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

To use this metric, a tweet must be considered as a set of words such as $\{a,b,c\}$. Note that this set is unordered, which means $\{a,b,c\}=\{c,a,b\}$. This metric takes values between 0 and 1, and returns smaller values for more similar and larger values for less similar tweets, while it is 0 if the tweets are identical and 1 if they don't have any common word.

a. Use the provided initial centroids and cluster the tweets in $K = 25$ clusters. The output must be a file which contains the clustering results such that each line represents a cluster in the form of: *cluster_id: a list of tweet IDs which belong to this cluster*. Include this file in your report.

b. Design and implement an efficient method to find the K initial centres so that K-Means can generate good clustering results similar to the results you obtained in the previous part.

Note 1: Each element in the initial centroids list is the tweet ID.

Note 2: You are expected to implement Jaccard distance as well as K-Means algorithm by yourself.

6. Some Explanatory Questions (10 Pts.)

Please answer the following questions as clear as possible:

a. Can K-Means with a specific parameter k ever converge to a result which contains more or less than k clusters? Explain.

b. K-Means with typical settings can only return *circular* clusters. However, in many cases clusters are not circular and may appear as lines. Give a method or a distance function to capture both *circular* clusters and *appear-as-line* clusters.

c. In K-Means clustering algorithm, the goal is to minimise the variance of the solution. In general, how does the variance of a partition change as the number of clusters is increased? Justify your answer.

d. K-Means is said to be useful in the problem of outlier detection. Suggest a strategy to perform this task.

e. How is K-Means affected by the curse of dimensionality?

Good Luck!
Ali Abbasi, Mohsen Ebadpour