Assignment 4

The ~~Curse~~ Blessing of Dimensionality!

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 10th**.
- ***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 me through **ali.the.special@gmail.com**.

1. PCA vs. LDA: Comparison of Two Different Linear Projections (6 Pts.)

Keywords: *Dimensionality Reduction, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA)*

Although **PCA** and **LDA** are both based on linear transformations, one is considered to be supervised whereas the other is unsupervised. Therefore, they return different feature subspaces. Here, we are going to investigate this fact by considering two different scatter plots taken from the Penguin dataset. You are required to perform the following tasks on them both.

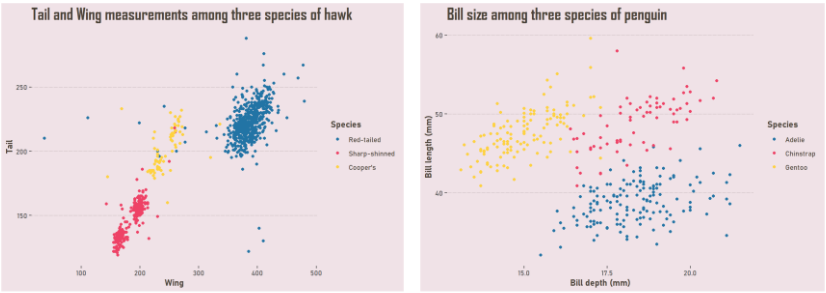

Figure 1 Two scatter plots of different features of the samples in Penguins dataset

a. Draw the first principal component direction assuming samples of class blue and red.
b. Draw the first Fisher’s linear discriminant direction assuming samples of class blue and red.
c. Repeat part a. considering samples of class red and yellow.
d. Repeat part b. considering samples of red and yellow.
e. Repeat part a. considering all samples.
f. Repeat part b. considering all samples.

Note: Graphs displayed in Figure 1 are provided along with this assignment.

2. Understanding the Behavior of Clustering Techniques (8 Pts.)

Keywords: *Clustering Problem, K-Means Clustering, Hierarchical 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 2, 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 will be. If there is more than one possible solution, then please specify for each solution whether it is a global or local minimum.

a. K = 3
b. K = 2
c. K = 3
d. K = 4
e. K = 3

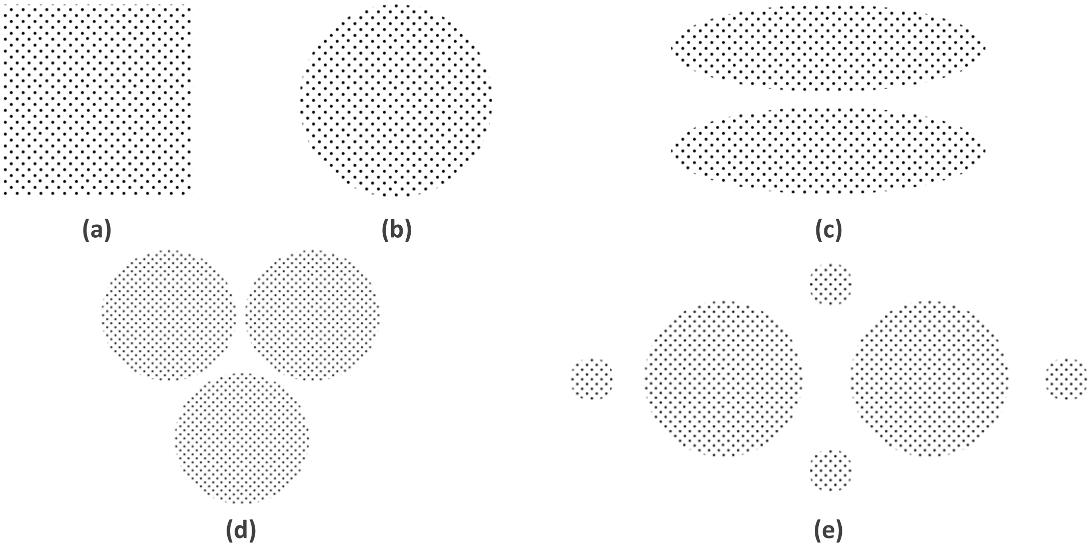

Figure 2 Sets of 2-D points provided for the first part of the problem

Finally, consider the diagrams in Figure 3.

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.

Note: Images in both figures are given so that you can use them in your report.

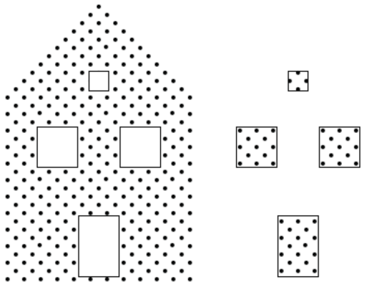

Figure 3 The goal of this clustering problem is to distinguish the main parts, i.e. door and windows

3. Why Naples-Style Pizzas Are So Delicious? (12 Pts.)

Keywords: *Dimensionality Reduction, Principal Component Analysis (PCA)*

Naples-style pizza (also known as *Neapolitan pizza*) is one of the most well-known and popular types of pizza in the world. Basically made with tomatoes and mozzarella cheese, this type of pizza is a Traditional Speciality Guaranteed (TSG) product in Europe and the art of its making is included on UNESCO's list of intangible cultural heritage.

Figure 4 Naples-style pizzas are famous for their unique hand-made doughs (a) Pizza Margherita, a variation of Neapolitan pizza (b) The process of preparing dough for Neapolitan pizza

The process of preparing Neapolitan pizza dough is a key factor which makes this kind of pizza so special. This lead to a series of researches made by food technicians in order to compare traditionally made pizza doughs from Naples with doughs from other areas. While these two groups of doughs tasted significantly different, the researchers weren't able to find specific characteristics to distinguish between doughs made in Naples and those which were made elsewhere. Hence they collected a multivariate dataset to further investigate the matter.

The dataset contains different characteristics measured from doughs used in six different restaurants, with the first four being famous Naples restaurants and the last two are from restaurants in other Italian cities. For each dough, two mechanical tests (measuring pressure load and deformation volume) and four microbiological/chemical tests (counting the bacteria in the dough, counting the yeast, measuring pH and measuring total titratable acidity) were conducted.

a. Calculate the first three principal components of the given data. Are they sufficient to describe most of the variation in the dataset?
b. Use the principal components in different ways to evaluate the normality of the data.
c. Plot two and three-dimensional scatter plots of the first three principal components. Based on the plots, are the first three PCs enough to discriminate between doughs from Naples and doughs from other areas? Does it comply with your conclusion in part (a)?

Note: You are required to use your own implementation of PCA in this problem.

4. Interpretation of 1000 Genomes Project Results (18 Pts.)

Keywords: *Dimensionality Reduction, Principal Component Analysis*

The 1000 Genomes Project (abbreviated as *1KGP*) is by far the most sophisticated and detailed attempt to list different human genetic variations. Started from 2008, the project was conducted by various research teams from institutes around the world, and finished its pilot phase in 2010 by announcing the sequencing of 1092 recorded genomes in a paper published by Nature.

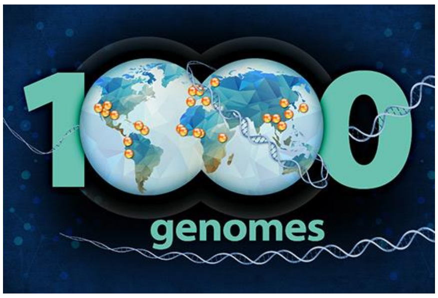

Figure 5 Naples-style pizzas are famous for their unique hand-made doughs (a) Pizza

The goal of this problem is to use **Principal Component Analysis** to scrutinize the results obtained in this project. You are given a dataset published in 1000 genomes website, which contains 995 lines each representing an individual. The first three columns are the individuals unique identifier, sex (1 = male, 2 = female) and the population he/she belongs to (see here for decoding), respectively. The subsequent 10101 columns of each line are a subsample of nucleobases from the individual's genome.

Since PCA needs a real-valued matrix, you must first convert the data from the text file of nucleobases to a real-valued matrix. To do so, convert the genetic data into a binary matrix Y such that $Y_{i,j} = 0$ if the i^{th} individual has a column j's mode nucleobase (the most frequently occurring nucleobase in that position across the 995 data points) for his/her j^{th} nucleobase, and $Y_{i,j} = 1$ otherwise. Note that all mutations appear as a 1, even if they are different mutations, so if the mode for column j is 'G', then if individual i has an 'A', 'T', or 'C', then $Y_{i,j}$ would be 1. The first three columns of the data provide meta-data and therefore should be ignored when creating the binary matrix Y.

a. What would be the dimensions of the returned vectors after performing PCA on the binary matrix Y?
b. Project the data onto their first two principal components and display the associated scatter plot. Use different colors to specify each population.
c. Explain briefly one or two basic facts about the plot you obtained in the previous part. What can you interpret from the first two principal components? What aspects of the data are the first two principal components able to capture?

Hint: Think about history and geography.

d. Now create another scatter plot with each individual projected onto the subspace spanned by the first and third principal components. Then play with different labeling schemes (with labels derived from the meta-data) to explain the clusters that you see.

e. Briefly explain what information does the third principal component capture?

f. Plot the nucleobase index versus the absolute value of the third principal component. What do you observe? Give a possible explanation.
Hint: Think about chromosomes.

Note 1: At first, the data should be centred but not normalized. Also the output must be the normalized principal components (i.e. unit-length eigenvectors).

Note 2: All plots must include a legend.

Note 3: Using library implementation of PCA is allowed and recommended.

Recommended MATLAB function: `pca()`

Recommended Python function: `sklearn.decomposition.PCA()`

5. How Twitter Reacts to a Crisis? (12 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 6 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. A Tribute to Stephen Hawking (14 Pts.)

Keywords: *Clustering Problem, K-Means Method, K-Means++ Method, Hierarchical Clustering, Silhouette Plot*

Legendary physicist Stephen Hawking was 21 when he was diagnosed with *Amyotrophic Lateral Sclerosis* (*ALS*) in 1963. ALS is a rare disease which leads to gradual decline of the brain's ability to control muscles and usually kills the patient within about four years. However, Hawking managed to survive for over 55 years until he passed away at the age of 76 in 2018.

In this problem, we aim to investigate the functional change in a patient diagnosed with ALS over time. To do so, we rely on the *Amyotrophic Lateral Sclerosis Functional Rating Scale* (*ALSFR*) which is a measurement used for evaluating the functional status of ALS patients.

Figure 7 When Hawking was diagnosed with Amyotrophic Lateral Sclerosis in 1963, few thought he would live more than a couple of years

a. Select 10 features with highest covariance with ALSFRS slope. You must support your choice by appropriate visualization.

b. Train a K-Means model on the reduced dataset. Using a clear strategy, find the best possible value for the parameter K and use it afterwards.

c. Evaluate the model performance by reporting the centre of clusters and silhouette, and explain the details. Also use bar plot to show the centres.

d. Tune parameters and plot with K-Means++.

e. Return the model with optimal parameters and interpret the clustering results.

f. Apply hierarchical clustering on three different linkages and compare the corresponding silhouette plots along with dendrograms.

g. Compare the results of the above methods.

Note 1: Using built-in implementations is permissible.

Note 2: Please search for terms and definitions the may sound unfamiliar to you.

Recommended MATLAB functions: `cluster(), kmeans(), linkage(), silhouette(), dendrogram(),bar(),histogram()`

7. Are You Into Fashion? (20 Pts.)

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

Zalando is a Berlin-based company which offers fashion and lifestyle products to customers in European countries. Founded in 2008, the company has recently published a dataset which contains 70,000 images, each belonging to one of the 10 groups of different fashion articles. This dataset is known as fashion-MNIST, since it is often considered to be an alternative for the well-known MNIST dataset.

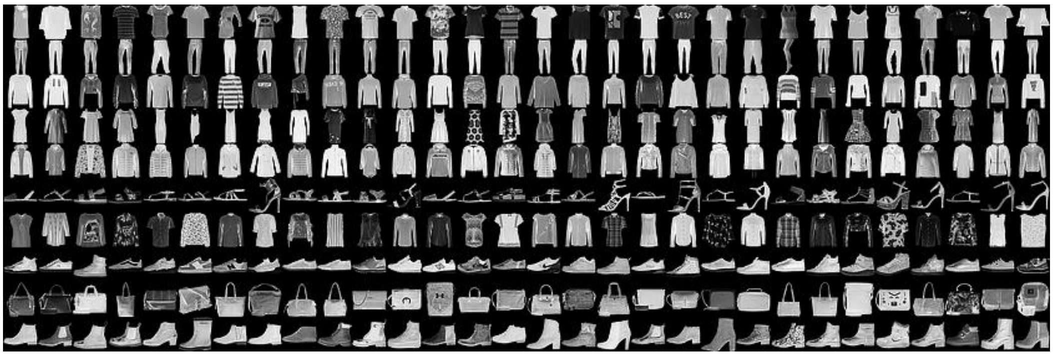

Figure 8 Examples of the images available in the dataset. Each row represents a different class.

In this problem we aim to use both **PCA** and **K-Means** in order to cluster the images in a reduced version of this dataset so that we obtain clusters which enable us to understand and interpret the data.

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. Apply LDA on the dataset and project the samples onto their first and second linear discriminants and display the results.

c. Now perform K-Means on the data using their first two principal components. Select the initial centroids randomly and set $K = 4, 7, 10$. Display the result of clustering with final centres highlighted. Also compare the results together and with the plot you obtained in part (a).

d. Repeat part (c) considering the initial centroids as the mean of the samples such that for K=4, initial centroids are the mean of samples of classes {1,3,5,7}, {2,4}, {6,8,10} and {9}, for K=7, initial centroids are the mean of samples of classes {1,3,5}, {2}, {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.

Note 1: In all scatter plots, samples of different clusters must be highlighted with different colors.

Note 2: You are allowed to use built-in functions and libraries for this part.

Recommended MATLAB functions: `pca(), kmeans(), bar(), histogram()`

8. Some Explanatory Questions (5 Pts.)

Please answer the following questions as clear as possible:

a. Why doesn't it sound reasonable if the number of PCs is greater than dimensions?
b. Explain how PCA can be applied to the problem of image compression.
c. Is it possible to make a kernel version of K-Means clustering? If yes, how? And what would be an advantage of kernel-based K-Means? If no, why?
d. 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 (K) is increased? Justify your answer.
e. What value of the parameter K in K-Means can always lead to a variance of zero, and why?

Good Luck!
Ali Abbasi