IE 582 Statistical Learning for Data Mining
Homework 2, due November 16th, 2021

Instructions: Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

- A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
  - https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
  - https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps-917519326e41

  Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

- A Markdown html document. This can be created using RMarkdown for R and Python Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself. As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Task 1 – Dimensionality reduction

Consider a classification problem where you have only two features (the data is on Moodle). You decided to visualize the instances on a 2D feature space where the class information is color-coded. The R code to obtain the plot below is already given to you and after running to code you obtained Figure 3.

```
1  dat=read.csv('your_path_goes_here/IE582_Fall21_HW2_q1_data.csv,header=T)
2  lev=as.numeric(dat[,3])
3  plot(dat[,1],dat[,2],col=lev,pch=lev,xlab=names(dat)[1],ylab=names(dat)[2])
4  legend("topleft",paste("Class",levels(dat[,3])),col=unique(lev), pch= unique(lev))
```

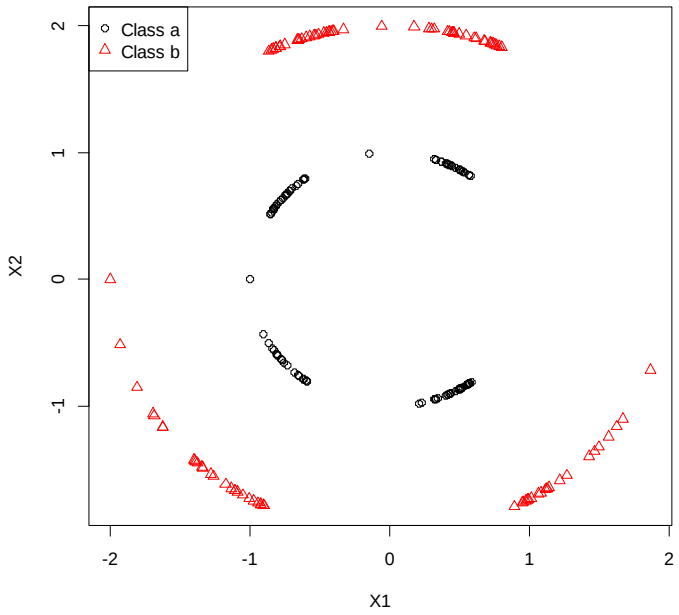

Figure 3. Scatter plot of the feature X1 versus X2 and the class information

a) Apply PCA to reduce the number of dimensions to one and visualize the instances on a scatter plot. Note that the scatter plot will show the observation number versus the observed value (as we have a single feature to represent the instance).

b) Apply MDS to reduce the number of dimensions to one and visualize the instances on a scatter plot as in part (a). Use at least two different similarity measure.

c) On a 2D scatter plot, one can observe how two observations from two classes are different.

1. What is your conclusion when you use PCA results in single dimension (your results from part a)?
2. What is your conclusion when you use MDS results in single dimension (your results from part b)? Compare the results from different similarity measures used in part (b)
3. Compare MDS results with PCA. What is your conclusion?

d) Suppose, you are not satisfied with your dimensionality reduction scheme in part (a). Add the following columns to your data, $X1^2, X2^2, X1 \times X2$ (three columns as functions of your original variables) and apply PCA. Comment on the PCA results (i.e. what are the eigenvalues? What do they refer to?).

Task 2 – Reconstructing Turkey Map

Please download the intercity distance information from Karayolları Genel Müdürlüğü's website from the link below:
https://www.kgm.gov.tr/SiteCollectionDocuments/KGMdocuments/Root/Uzakliklar/ilmesafe.xls

Suppose we apply MDS to this data to distance matrix to obtain latent variables in 2D dimensional space. Represent the new feature space on a 2D plot. You are expected to label each data point with the respective city's name. This plot should be similar to Turkey map. Are there any unusual observations? If yes, comment on your findings.

Task 3 – Dimensionality reduction for time series data

The conference paper by Liu et al. (2009) starts with the following statement: “Gestures have recently become attractive for spontaneous interaction with consumer electronics and mobile devices in the context of pervasive computing”. The aim is to provide efficient personalized gesture recognition on wide range of devices.

To achieve this, Liu et al. (2009) uses a single three-axis accelerometer to collect data from eight users to characterize eight gesture patterns. The library, uWaveGestureLibrary, consists over 4000 instances each of which has the accelerometer readings in three dimensions (i.e. x, y and z). Eight gestures are illustrated in Figure 1.

The dataset is provided in the following link:
https://drive.google.com/drive/u/1/folders/13553neknux7U8why55KM1WrjgkA9IJKm

Note that there are separate files for each axis and each row corresponds to one gesture in the files. First column has the class information. The information between second and last column is the time ordered observations in the corresponding axis (provided in the file name as X, Y or Z). Moreover, the data is split into training and test sets. For now, you are expected to work with only the training series. Hence, following files are to be used:

- uWaveGestureLibrary_X_TRAIN,
- uWaveGestureLibrary_Y_TRAIN,
- uWaveGestureLibrary_Z_TRAIN

a) Read the data and visualize one instance (all axes) from each class and try to relate the shape (time series) you see with the gestures shown in Figure 1 (this is just for fun, sometimes it is good to start with data visualization to understand what is going on). A 3D scatter plot would be interesting. Note that this is an acceleration information. You can transform this information to a velocity vector by computing the cumulative sum of acceleration over time. A cumulative sum operation on the velocity values will transform the series to a position information.

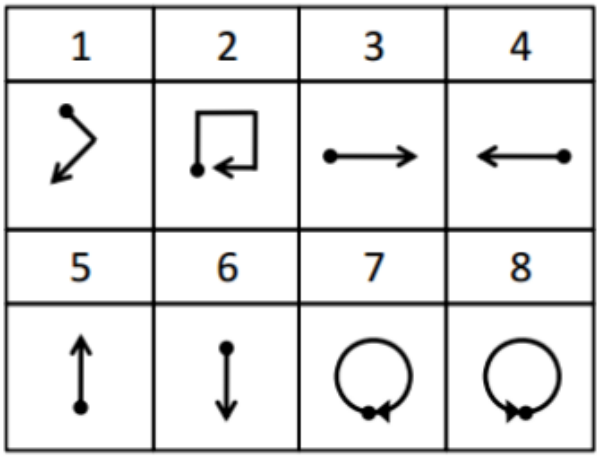

Figure 1: Gesture vocabulary considered by Liu et al. (2009). The dot denotes the start and the arrow the end

b) As you may have noticed, the data is provided as a regular data matrix (i.e. each row represents an instance and columns represent the time index of the observations). Assume that X, Y and Z variables are concatenated to obtain a single time series in an unusual way as follows:

$$\text{concatenated series} = [x_1\, x_2\, \ldots.\, x_T\, y_1\, y_2\, \ldots.\, y_T\, z_1\, z_2\, \ldots.\, z_T\,]$$

In other words, the information from X, Y and Z axis are concatenated to obtain a single series of length 3T (each axis is represented by a time series of length T).

Apply PCA to the time series from each class in the training data. You are expected to filter the data from each class and apply PCA to the representation. Work on the following questions based on your PCA application for the time series from each class.

- How much variability can be recovered by the first two components?
- Draw the each eigenvector (component) as a time series for each class. What do the eigenvectors imply in this setting? In total, you are expected to provide 8 plots. Each plot is expected to provide two time series (i.e. first eigenvector and the second eigenvector). Are there any interesting patterns/observations? If yes, provide your comments.