German University in Cairo
Faculty of Media Engineering and Technology
Spring 2019

CSEN 1083 – Data Mining

Assignment #1

(Due on: February 28th at mid-night)

Problem 1

Implement the Principal Component Analysis (PCA) algorithm given in lectures. Your function should take the data as input and outputs the principal components in addition to the projection of the data samples on each principal component. Apply your function to the data given in the file "Data.txt". Refer to the notes at the end of this document to know what you are allowed to use and what you are not allowed to use from ready-made functions in Python.

Deliverables:

- Your code.
- A plot of the data showing the two identified principal components on the same plot. Save your plot as "Data_PCA.png".

Problem 2

Consider the data given in the file "EPL.xlsx". The data represents the outcome of all games in the English Premier League in the season 2017/2018 that ended in the win of one of the two competing teams. Apply the function you implemented in Problem 1 to this dataset. The attributes given in the file for each game are as follows:

Home Team
Away Team
HS: Home Team Shots
AS: Away Team Shots
HST: Home Team Shots on Target
AST: Away Team Shots on Target
HF: Home Team Fouls Committed
AF: Away Team Fouls Committed
HC: Home Team Corners
AC: Away Team Corners
FTR: Full-time Result

The home team and away team names are given for reference only. They should be removed from the analysis. The FTR field represents who won the game (H if the home team won and A if the away team won). It should only be used to identify which team won the game only. Therefore, this data has 8 dimensions excluding the teams names and the FTR field.

CSEN 1083 – Data Mining

Assignment #1

(Due on: February 28th at mid-night)

Deliverables:

- Your code.
- A10-bin histogram of the projection of the data on each of the identified principal components. Bars in the histogram corresponding to the games in which the home team won should be colored in red, while bars corresponding to games in which the away team won should be colored in blue. Save the histograms as PNG files. Name your files (Proj_PC1.png, Proj_PC2.png, …, Proj_PC8.png).
- In order to evaluate how each principal component provides better representation of the data, for each principal component, compute the difference between the mean of the projections of the data points representing when the home team won and the mean of the projections of the data points representing when the away team won. Plot the distance computed versus the principal component index. Name your plot “Distance.png”.

Notes:

- This is an individual assignment.
- You are not allowed to use the function that computes the covariance matrix in Python. You need to implement it.
- You are allowed to use the function that computes the eigenvalues and eigenvectors in Python and any other function you might need.