موسسه آموزش عالی آزاد توسعه

برگزار کننده دوره‌های تخصصی علم داده

Homework 6: Winter 2020

Due date:۲۴ اسفند

Please email your HWs to y.zerehsaz@gmail.com
Please append all your codes to your response

This HW must be completed in **Python**.

Consider the "Spine" dataset to answer the following questions.

The Spine dataset contains information about patients belonging to one of three categories of lumbar spine malfunctions: 1) Normal, 2) Disk Hernia and 3) Spondylolisthesis with the last two categories being abnormal. Each patient is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence (PI), pelvic tilt (PT), lumbar lordosis angle (LL), sacral slope (SS), pelvic radius (PR) and grade of spondylolisthesis (GS).

Question 1:

a) Get the summary of data and check whether or not there are any missing values.

b) Get the structure of the data frame.

c) Compute the mean and standard deviation for the PI variable, and mean, median and standard deviation for GS variable. Use the agg function.

d) Group the whole dataframe based on the "Categories" (last column in Spine dataset). Compute the mean and standard deviation associated with each group for all variables. What do you think about the differences in means and standard deviations of variables among the levels of the variable "Categories"?

e) Compare the boxplots for the variable GS corresponding to the groups of the variable "Categories". You must locate all boxplots in one plot (see googleapps.py file).

Question 2:

Perform the PCA method on the ***scaled*** data (do not use the last column) and answer the following questions.

a) Scale the data and apply PCA on the scaled dataset.

b) Provide the matrix of directions (loadings). Interpret the first three directions. This is the "W" matrix in the slides.

c) Compute the explained variance ratio and decide on the number of components to choose (justify your answer).

d) Using the scree plot, how many components would you like to choose? (justify your answer).

e) Make a scatter plot of the first two PC scores (**use the output of Python i.e. fit_transform**) and interpret the observations (patients) based on the loading matrix and their position in the plot. Are there any unusual observations (possible outliers) in the data?

f) If you have found several outliers in part e, choose the worst one and answer the following questions.

I) Which observation does this outlier belong to?

II) On what variables (variable) do you think this outlier has the highest values?