Alexandria University
Faculty of Engineering
Computer and Systems Engineering Dept.

Prof. Dr. Mohamed Abougabal
Prof Dr. Nagwa M. El-Makky
Dr. Marwan Torki
Special Topics in Information Systems

Assignment 2
Data Exploration and Preprocessing

Objectives

1) Getting familiar with commonly used operations in data exploration and preprocessing

2) Investigating scikit-learn preprocessing capabilities with more depth

Download and Read Data

Download the image segmentation dataset https://archive.ics.uci.edu/ml/machine-learning-databases/image/
Add both files: segmentation.data and segmentation.test together to get a large training dataset.

Explore the dataset, state how many readings and how many attributes are there. Also, state how many classes exist in the dataset.
Plot your data using plt.plot() to be able visualize it.
(Hint: sort your data according to the classes for a better visualization of the tasks given below).

Data Exploration

- **Pearson's correlation**
  1- Compute the Pearson's correlation coefficient between each 2 attributes (features). This should result in a dxd symmetric matrix where d is the number of features.
  2- Visualize your output matrix using imshow.

- **Covariance**
  1- Compute the Covariance matrix of the dataset.
  2- What is the relation between the covariance matrix of the dataset and the Pearson's correlation matrix of it?

- **Histograms**

1- Plot the data histogram for each class. Let each class have only one plot with a different color for each attribute.
2- Try different number of bins, bins= 5,10, 12

Preprocessing

1. Normalization

Apply data normalization on the dataset using two different approaches;

- Min-max scaler
- Z-score normalization

Visualize the data after normalization by each approach, using histograms or box-plots. What is the difference before and after each normalization?

2. Dimensionality reduction

- **Feature Projection**

Principal Component Analysis (PCA): PCA computes the principal components of a dataset and reduces its dimensionality. Apply PCA on the dataset **after being normalized by z-score normalization**.
Try more than one appropriate number of components.
Also use the attribute pca.explained_variance_ratio to know the variance captured by each component.

Visualize the correlation matrix of your dataset after applying PCA, what is your conclusion?

- **Feature selection**

In sklearn.feature_selection, use SelectKBest to reduce the number of data features. Try different values of K

Visualize the correlation matrix and plot your data after applying feature selection and state your conclusion about the resultant figures.

.

Notes

1. You should deliver well documented code as well as a report showing all your work and conclusions.
2. You should be working in groups of 2.
3. Copied assignments will be penalized; so not delivering the assignment would be much better.
4. You should write your code in python.

Good Luck