Alexandria University
Faculty of Engineering
Computer and Systems Engineering Dept.

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

Assignment 1
Data Exploration and Preprocessing

Objectives

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

2) Getting familiar with Python programming and Scikit-Learn visualization and preprocessing capabilities.

Download and Read Data

Download the image segmentation dataset http://archive.ics.uci.edu/ml/datasets/image+segmentation

Add both files: segmentation.data and segmentation.test together to get a large dataset.

Explore the dataset. State number of instances, attributes and classes.

Data Exploration

- **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

- **Boxplots**
  1- Plot Boxplots to compare the attributes of the dataset

- **Correlation Matrix**
  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.

Preprocessing

1. Normalization

Apply data normalization on the dataset using two different approaches;

- Min-max scaler
- Z-score normalization

Visualize the dataset 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**. Use an appropriate number of principal components.
Also use the attribute pca.explained_variance_ratio_ to know the variance percentage 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. Use an appropriate value of K

Visualize the correlation matrix after applying feature selection and state your conclusion.

.

Notes

1. You should deliver well documented code as well as a report showing all your work and analysis.
2. Extra visualization and/ or analysis is appreciated
3. You should be working in groups of 2.
4. Copied assignments will be penalized. So, not delivering the assignment would be much better.

Good Luck