Machine Learning [NETW1013]
Spring 2019
Assignment 6

Deadline: Sunday 14th of April

For all tasks listed below, use the data set of house prices used previously in Assignment 1

Principal component analysis

1- Assume that the input data set of the house data (column 4 to column 21) is equal to x Calculate the Correlation matrix of the x
Matlab code: Corr_x = corr(x)

2- Use the correlation matrix to identify the relation between different parameters.

3- Calculate the covariance matrix using "cov" function
Matlab code : x_cov=cov(x) ;

4- Used the Matlab SVD function to identify the principal components of the House prices data set using the cov of the house data set
Matlab code : [U S V] = svd(x_cov)

5- Use the EigenValue produced from the SVD function to find K where K is the minimum number of dimensions that can be used to describe a house. This will reduce the number of dimensions from m to K

$$\text{EigenValues } = [\lambda_1\ \lambda_2\ \ \lambda_3 \ldots \ldots .\lambda_m]$$

Where m is the number of dimensions?
The Eigen values are the diagonal of the matrix S

Hint (calculate $\alpha = 1 - \frac{\sum_{i=1}^{i=K} \lambda_i}{\sum_{i=1}^{i=m} \lambda_i}$ and find K that would make $\alpha \leq 0.001$)

6- Use the Eigen vectors to transform the data set to the reduced dimension data set
Reduced_Data=R= $U(:,1{:}K)^T\ x^T$

7- Use the Eigen vector to produce an approximate data out of the reduced data by multiplying by the Eigen vectors matrix.

8- Estimate the error in the data produced by the dimension reduction
Error = $\frac{1}{m}\sum_{1}$ (approximate data − Reduced_Data)

9- Use linear regression to estimate house prices based on the data set produced using principal component analysis.

K means clustering

1- Use K means clustering to find the clusters involved in the House data set and find the optimal number of clusters and their respective center points

2- Use K means on the reduced data set and compare the produced clusters on the real data in both cases

Anomaly detection

Apply anomaly detection to the house data set and use to build an anomaly detection system.

Submission & Grading

For submission upload your files to a github folder and submit a link to it using the following form:

https://docs.google.com/forms/d/e/1FAIpQLSf3koMKnUqn1JmabYtw3fV5mAWRrz2IDkXe6VM9pkOEq6x3DA/viewform?usp=sf_link

Part	Points
Principal Component Analysis	40
K Means Clustering	30
Anomaly Detection	30
Total	100

Submissions by email will NOT be considered.