AI 221: Machine Exercise 6

Instructions:

- Read and answer each problem using computer code. This MEX should be done *individually*.
- Each item should be answered as either a Jupyter Notebook or a MATLAB Live Script, exported as a single PDF file for the entire MEX. Make sure to HIGHLIGHT your final answers.
- When done, submit the PDF file through UVLE.

Trip Advisor Travel Reviews

Download the Travel Reviews data set from the UCI Repository:

http://archive.ics.uci.edu/ml/datasets/travel+reviews

The data set contains 980 unique users of Trip Advisor with 11 features each: Their average feedback scores on destinations in East Asia. These include art galleries, dance clubs, juice bars, restaurants, museums, resorts, parks, beaches, theaters, and religious institutions.

The feedback score is a rating with 4-Excellent, 3-Very Good, 2-Average, 1-Poor, and 0-Terrible.

The goal of this problem is to find users with similar or dissimilar rating patterns:

a. **[10 pts]** Perform K-means clustering directly on the 11-feature data set. Based on the Silhouette Score, what is the best K? Report the silhouette scores of each user as a bar plot.

b. **[20 pts]** Redo item (a), but this time, reduce the 11-feature data down to 2-features using PCA first, before feeding the 2-feature data into K-means clustering. In addition to the silhouette score bar plot, display the 2-D data colored by cluster, for the best K that you found.

c. **[20 pts]** Using Kernel Density Estimation, analyze the 2-D data set if there are users that can be considered outliers or unusual raters. Use a confidence level of 95%. Plot the KDE results in 2-D.

Anomaly Detection in a Wastewater Treatment Plant

Download the Wastewater Treatment Plant data set from the UCI Repository:

https://archive.ics.uci.edu/ml/datasets/water+treatment+plant

The data set contains 527 measurements of 38 process variables in a wastewater treatment plant in Spain: each row of values represent one day's worth of measurements. These measurements are related to both the process equipment and the water quality, all of which are numeric and continuous.

The goal of this problem is to find days when we suspect anomalies in operation.

a. **[20 pts]** Directly perform DBSCAN on the 38-feature data set. Decide on a suitable values of minPts and epsilon. List the outliers that you found.

b. **[30 pts]** Reduce the data set into 2 dimensions using a suitable dimensionality reduction method. In the 2-dimensional space, perform KDE, One-Class SVM, and Local Outlier Factor with a suitable set of hyper-parameters. Compare the performance of the three methods.

END OF EXERCISE