CSE 494: Artificial Intelligence for Cyber Security: Spring 2019

Lab 1: Unsupervised Machine Learning demo

Objectives of the lab:

- Employ *k-means* clustering algorithm on a real world dataset using the *scikit-learn* package
- Experiment with different parameters of the algorithm and *evaluate* the output of the clusters
- Extend this clustering algorithm to successfully implement agglomerative clustering using *scikit-learn*

OPTIONAL:

1. To refresh your knowledge on k-means clustering algorithm, you may choose to visit this brief tutorial on the same: http://www.cs.cmu.edu/~cga/ai-course/kmeans.pdf (taken from cmu course on ML)
2. Also, we will be heavily relying on 2 python packages: numpy for numerical processing and pandas for data storage frames and queries. Please look up some tutorials online if you are not familiar with these.

The lab will be divided into two sessions:

1. *K means Python Demo with Scikit Learn (30 min.) – We will be implementing the following on the dataset Colleges.csv uploaded on blackboard – the data is publicly available on Kaggle https://www.kaggle.com/flyingwombat/us-news-and-world-reports-college-data*
    a. Analysis of the attributes and pre-processing the data to convert the attributes into a feature matrix in numpy array format.
    b. Run K-means on the dataset with k=2, 3 and 4 clusters. Initialize clusters randomly using the scikit-learn *k-means* method. Extract the labels for the samples after n iterations.
    c. Initialize clusters using our own selected starting points. Check for the differences in labels.
    d. Dropping some variable prior to clustering – evaluate the change in the clusters with respect to random initialization.
    e. Dealing with missing data in the feature matrix.
    f. CLUSTER EVALUATION:
        i. Evaluate the clusters when labels are unknown - https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient
        ii. Evaluate the clusters when labels are known – Precision, Recall, F1.

2. Agglomerative clustering Demo with scikit-learn (30 min.)

a. Implement the agglomerative clustering algorithm on the dataset with single linkage, and euclidean distance as the “affinity” .
b. OPTIONAL - Plot the dendogram
c. Check the cluster labels with respect to the *k-means clustering* with 4 labels in the previous case.
d. Selecting the clusters at different levels instead of the complete tree
e. Evaluate the clusters this time using the same metrics before for evaluation.