Name: Matt Hall
Title: Exploring kNN, MLE and SVM Classification Techniques

Data Set – This project will be using the Letter Recognition Data Set (https://archive.ics.uci.edu/ml/datasets/Letter+Recognition) from the UCI Machine Learning Repository

Project Idea: PCA will first be used on the dataset to order the attributes from most important to least important. Classification performance will be measured using SVM, MLE and kNN classification methods. The classification will begin with the single most important attribute, as determined from PCA, then adding the 2nd most important, then the 3rd and so on. There are 16 attributes in this data set, which will result in 16 tests for each of the 3 classifiers. Each test will be run 10 times, measuring the average and variance at each stage. The objective will be to determine an optimal number of attributes for each classifier. Also, to observe how the efficiency for each classifier behaves against the others while classifying with 1 attribute up through the full 16 attributes.

Relevant Papers:

Software:
Code will be written in Python using the numpy library for randomization, calculating covariance matrices, linear algebra calculations, etc. The MLE and kNN classifcation code will be similar to what was performed in previous homeworks, however, it will be modified to handle varying numbers of attribute inclusion. Ideally, press the run button and a plot will be generated showing classifier efficiency versus number of attributes included. SVM and PCA code will be developed for this project.

Experiments:
PCA will be used to determine significance of attribute data. SVM, MLE and kNN classification methods will then be used beginning with the most significant attribute. The second most significant will then be added to the training dataset, followed by classification by the 3 classification methods. Then 3rd most significant will be added and classified and so on. In order to accurately measure efficiency, each classifier will be trained using the same training data and tested on the same test data for each of the 10 runs.