CEG-7570 Project

Part I – Feature Selection

Data sets to be used in the project: A data set of bank notes and a data set of wheat seeds.

Bank Notes: This data set contains 1372 data points (observations), each with 4 features. Some bank notes are genuine while others are forged. An image of each note is taken and the variance, skewness, kurtosis, and entropy of the image are computed and used as the features of the note.

Wheat Seeds: This data set contains 210 data points, each with 7 features. There are 3 classes of wheat seeds. An X-ray image of a seed is taken and area, perimeter, compactness, length, width, asymmetry, and groove length of the seed are measured and used as the features of the seed.

Training data set and Test data set: Knowing the class of each data point, split each data set into a Training data set and a Test data set. If there are m points in the smallest class in a data set, move the first m/2 points in each class to the Training data set and the next m/2 points to the Test data set. Therefore, if a data set contains n classes, there will be overall mn/2 points in the Training data set and mn/2 points in the Test data set.

Feature Selection: Knowing the class of each point in a Training data set, in Part I of the project we will select the feature that maximizes the class separability measure as computed by the Fisher's discriminant ratio. In Part II of the project we will use the selected feature to train a 1-D Bayesian classifier.

Part I of the project is worth **8 points**.