Math 444 – Data Mining Spring 2017

Assignment 4

In this exercise, the goal is to test different classification algorithms, and in particular, their success rate. You need a *training set* and an independent test set. Load the already familiar data `HandwrittenDigits.mat`, which is your training set $\mathscr{T}$. Then load the file `HandwrittenDigitsTestset.mat`, constituting your test set $\mathscr{S}$.

1. Write a PCA classifier using the training set. The performance of the classifier depends on the number m of principal components that you include in the reduced model. To test the performance, run your classifier with several values of m, e.g., $m = 5, 10, \ldots, 80$, and plot the success frequencies $S_f(\mathscr{T})$ and $S_f(\mathscr{S})$ as functions of m.

2. Implement the k-nearest neighbor algorithm for the same data. Notice that for this method, the success rate with the training set is always $S_f(\mathscr{T}) = 1$ (why?). Test and plot the success rate $S_f(\mathscr{S})$ as a function of k, letting k vary from $k = 1$ to $k = 30$.

3. The file `ForestSpectra.mat` contains helicopter-borne scatterometer data of forest spectra. The data comes supposedly from four types of forest: Birch, Fir, Pine or Shrub.

   (a) Download the annotated data, and find the first few LDA directions. Investigate how well clustered the data is by plotting the LDA projections.

   (b) Download `ForestSpectraTest.mat`, and using an LDA classifier, investigate the performance.