ISyE 6416 – Computational Statistics - Spring 2016
Bonus Project: "Big" Data Analytics
Initial Project Proposal

Team Member Names: Danielle Boccelli and William Henry Taslim

Project Title: Comparison of Classification Methods in Identifying Handwritten Digits

Problem Statement

Handwriting recognition technology plays a significant role in technological advancement in recognizing the words on scanned documents and potentially, in effectively verifying signatures. Our team is interested in comparing the efficiency and effectiveness of various classification methods that we learned in ISyE 6416 (expectation–maximization (EM) algorithm, hidden Markov model (HMM), support vector machines (SVM) and/or *k*-means clustering) in recognizing a set of handwritten digits. The result of our project will provide an insightful information about each method for tech companies or computer scientists who are looking for the best way to recognize handwritten words.

Data Source

The data source for this project is the Semeion Handwritten Digit Dataset (available at: archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit). The dataset was originally collected for the Semeion Research Center of Sciences and Communication in Rome, Italy for usage in machine learning research - specifically, classification.

The Semeion dataset consists of 1593 handwritten digits written by a total of 80 individuals (refer to Figure 1.). These digits were scanned, and transformed into a 16x16 "pixel" grayscale image (for a total of 256 data points for each handwritten digits). Each pixel is the average gray value from the digit, generated by stretching the image. The grayscale images were then converted to black and white images (boolean values), by using a threshold for what is considered black, and what is considered white. Each individual from which the data was collected, was asked to write each digit (one through nine) twice: the first in their normal way of writing, and the second quickly (or sloppily compared to their standard writing).

Figure 1. An example of the handwritten digits from the dataset

Methodology

The goal of the project is to compare different classification methods using the Semeion dataset. Although each individual is bound to have variations in their handwriting, the general structure of the individual digits should be enough to provide some distinction for the purposes of classification. Both k-means and the EM algorithm will be coded in R, as well as all measures of accuracy and statistical significance. HMM and SVM will be implemented using existing R packages, but accuracy will be coded in R as it was for k-means and EM.

K-means will be run multiple times, with random starts, in order to minimize the possible effect of local minima in distance measurements, and accuracy will be judged based on the average accuracy of the runs. Since the points in the dataset are binary, the absolute distance between the centroid and any member of the cluster will be 1 (if the point does not match the centroid), or 0 (if the point matches the centroid).

The EM algorithm will run until the expected log-likelihood reaches a level of convergence (which will be determined at a later point in time). EM gives a probability that a point is in a cluster, and so the cluster with the highest probability at convergence will be taken as the final cluster for a given digit.

Expected Results

There are a few potential problems that may arise with the project. Firstly, the data is not very granular, which means that a lot of information is lost when going from grayscale to black and white (a written digit cannot be fully described with 256 black or white boxes). Secondly, some digits may not be easy to classify due to their similarity (for example: 5 and 6, or 3 and 8), and so there may be higher misclassification for some digits over others.

With the knowledge we currently have, we expect that EM will provide improved accuracy than *k*-means clustering. SVM and HMM will likely out-perform EM and *k*-means since R packages are better optimized. Since there is a total of ten digits we are looking at, we can expect that random guess would be about 10% accurate - anything close to this would not be better than a random guess.