Matt Hall

The dataset consists of 20,000 instances for training and testing. Each instance has 16 attributes and 1 class attribute. There are 26 potential classes, for each of the letters in the English alphabet. There was no attempt to normalize the data prior to running the PCA algorithm, however, this will need to be investigated b/c some attributes are counts while others are distances. MLE classified with 88% success rate without normalization, however, this may be improved by normalization.

Primary component analysis will be performed, followed by classification using MLE, kNN and SVM techniques.

The dataset will be randomized prior to each run. The first half of records will become part of the training data assigned to a new training matrix, while the second half will be part of the test data, maintained in a separate matrix from the training data.

At this point the experiments will be to run each classifier16 times after PCA is completed. Each of the 16 runs will correspond to dimension reduction of the raw data from the PCA results. There will be a total of 48 data points, 16 from each of the 3 classification methods4

PCA was performed on the dataset and the MLE classifier obtains 88% classification using all 16 dimensions from PCA. The 80% threshold was surpassed at 11 dimensions, while 1 dimension only provided 5% accuracy.