Dylan Plecki
dplecki2@illinois.edu
December 13, 2015
CS 498 Probability in CS

Homework 11: Take Home Final

Included in my submission is all of my outputs from 12 test runs, including plots, histograms, confusion matrices, and results of all relevant data. This has been attached in the *output.zip* file. My source code is written in Anaconda Python utilizing Numpy, Scikit Learn, Pandas, and MatPlotLib, and is in the *source_code.zip* file.

For my clustering algorithm, I used non-hierarchal K-means clustering with a various number of clusters. For my classifier, I used SkLearn's Random Forest library with a 75% training / 25% testing split. In an attempt to improve the accuracy of my classifier, I varied the sample block size between 16 samples, 32 samples, and 64 samples. I also varied the vocabulary size between 10-50 clusters in increments of 10 clusters. The results of the improvements are described below.

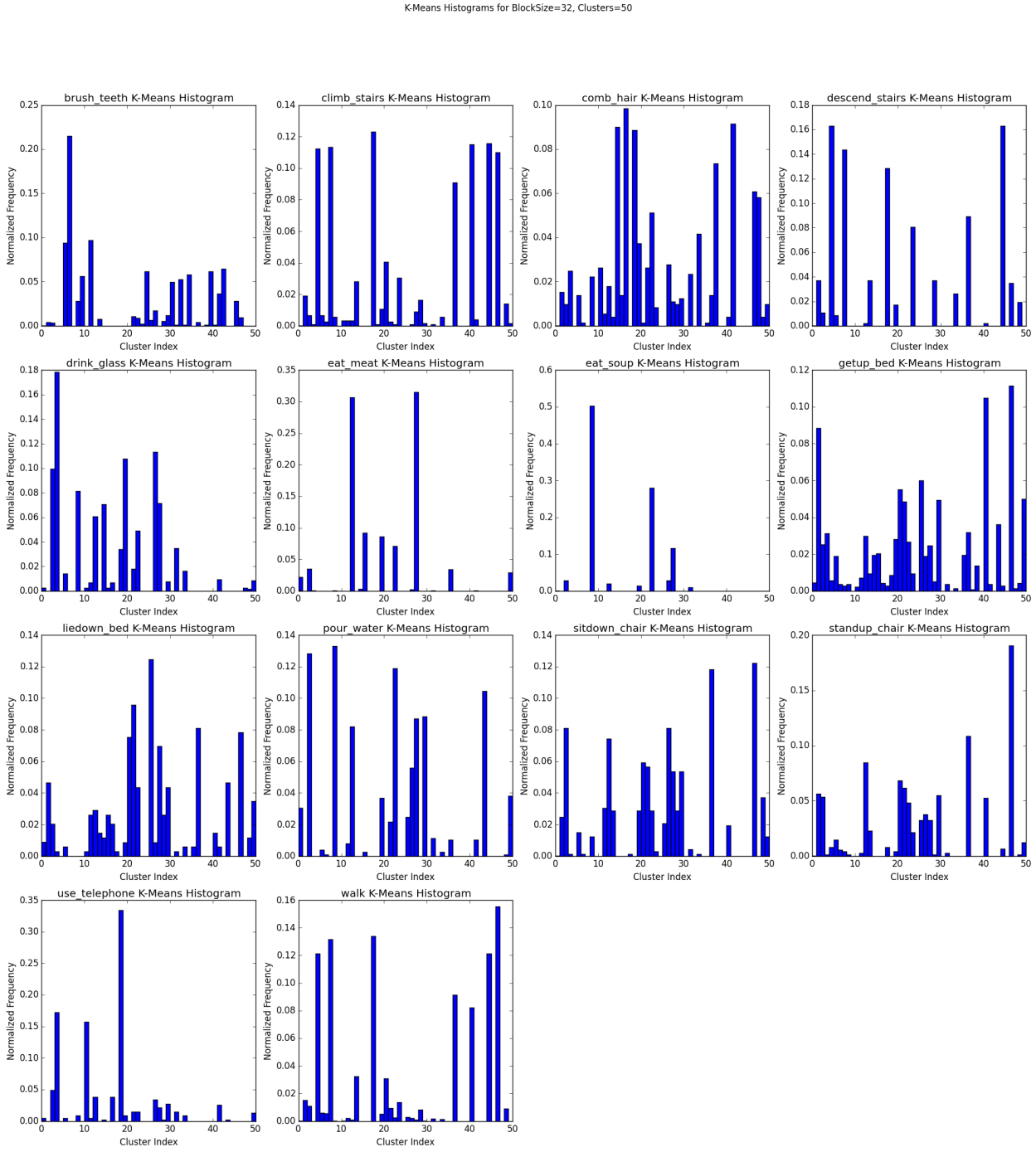
All other data (and there's a lot of it) that isn't described here can be found in the *output.zip* file. The below are general observations with a few acute examples of the dataset and my results.

Total error rate and accuracies for all runs of classifier:

Classifier Accuracies	10 Cluster Vocabulary Size	20 Cluster Vocabulary Size	30 Cluster Vocabulary Size	40 Cluster Vocabulary Size	50 Cluster Vocabulary Size	Average Accuracy
16 Sample Block Size	69.0%	67.1%	71.9%	66.2%	74.3%	69.7%
32 Sample Block Size	59.5%	65.7%	74.3%	66.2%	78.1%	68.8%
64 Sample Block Size	69.0%	71.9%	69.0%	68.1%	62.4%	68.1%
Average Accuracy	65.8%	68.2%	71.7%	66.8%	71.6%	68.8%

From the above table, it is fairly obvious that changing the sample block size and vocabulary size doesn't have a whole lot to do with the resulting classifier accuracy, but do offer some improvements towards the higher vocabulary sizes in most circumstances. Most of the varying accuracies most likely come from random errors, and the majority of the accuracies are centered around the average of 69%.

K-means histograms for most accurate classifier (32 sample block size / 50 cluster vocabulary size):



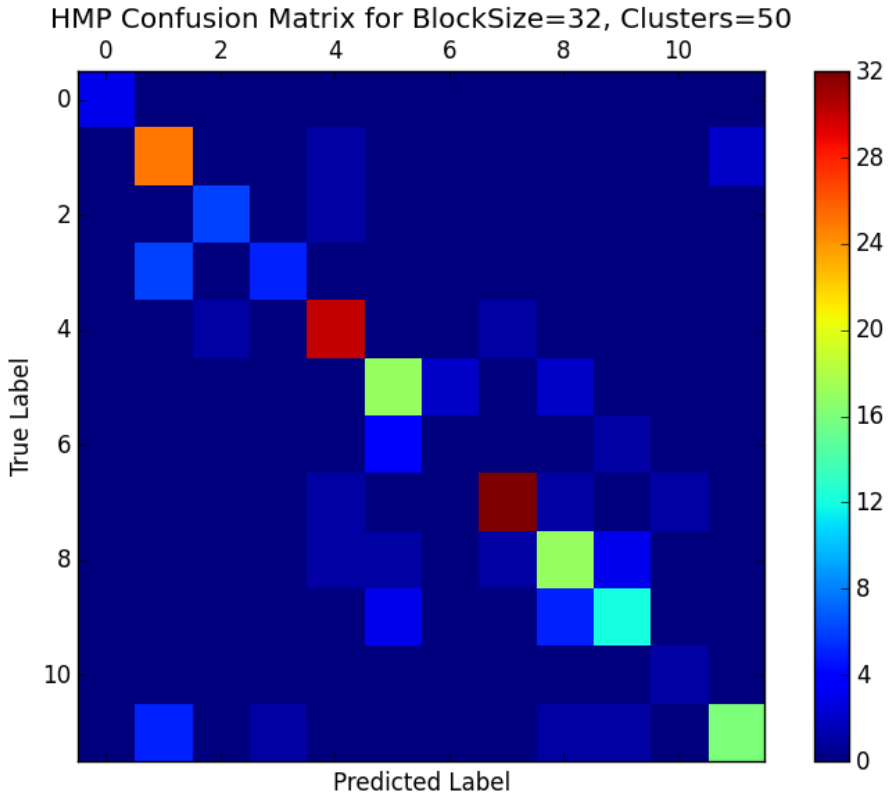
The histograms above represent distinguishing factors between the different activities. We can see from above that, for example, using a telephone or eating soup are relatively simple movements compared to the more complex such as getting out of bed or combing one's hair. We can also from these histograms learn to distinguish between the different activities based on shared (or distinct) cluster frequencies. This is where the learning classifier will come in.

Results from most accurate classifier (32 sample block size / 50 cluster vocabulary size):

```
SVM Results for BlockSize=32, Clusters=50

Accuracy score of the SVM: 0.780952
Confusion matrix (text-form):
[[ 3  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 25  0  0  1  0  0  0  0  0  0  2]
 [ 0  0  6  0  1  0  0  0  0  0  0  0]
 [ 0  6  0  5  0  0  0  0  0  0  0  0]
 [ 0  0  1  0 30  0  0  1  0  0  0  0]
 [ 0  0  0  0  0 17  2  0  2  0  0  0]
 [ 0  0  0  0  0  4  0  0  0  1  0  0]
 [ 0  0  0  0  1  0  0 32  1  0  1  0]
 [ 0  0  0  0  1  1  0  1 17  3  0  0]
 [ 0  0  0  0  0  3  0  0  5 12  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0]
 [ 0  5  0  1  0  0  0  0  1  1  0 16]]
```

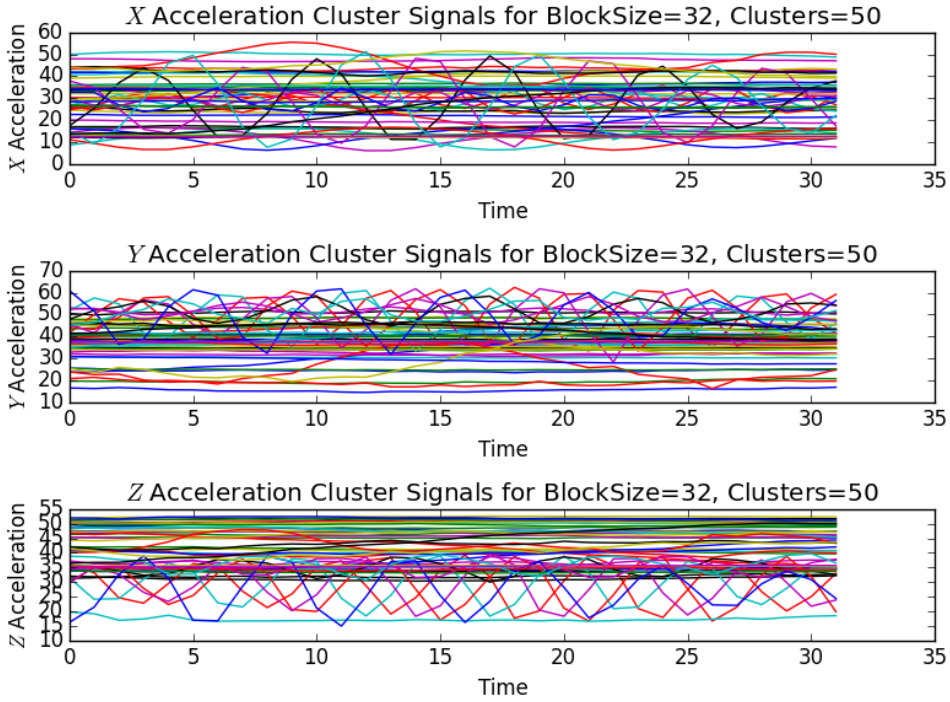
Class confusion matrix for most accurate classifier (32 sample block size / 50 cluster vocabulary size):



The above confusion matrix shows very good accuracy among many of the labels, as denoted by negatively-sloped the diagonal line of color which signifies correct predictions for the tested labels.

Signal plot for most accurate classifier (32 sample block size / 50 cluster vocabulary size):

Note: This looks different than David Forsyth's signal plot due to the non-hierarchal clustering and the larger vocabulary size. Thus, it is much more cluttered and has less defined curves.



Less cluttered signal plot from less accurate classifier (32 sample block size / 20 cluster vocabulary size):