Semi – Supervised Learning: Summary

1. Based on the paper by Rob Fergus et al on Semi-supervised Clustering on Large image datasets, I started working on the work that was done so far.

2. So as per the method described in the paper, I could successfully complete the code for semi – supervised learning based on eigenfunctions.

3. Stepwise Procedure:

3a) Fit Function:

a. Take n dimensional input data X and rotate the data points using PCA making them independent.

For every dimension of the data:

b. Get the histograms by using b bins for the entire data. Get the edges of the bins of the histograms.

c. Get the k eigenvalues and eigenfunctions using these histograms by Equation 2 mentioned in the paper [Fergus et al](https://cs.nyu.edu/~fergus/papers/fwt_ssl.pdf)

d. Interpolate the eigenfunctions using the histogram bin edge means and create a 1-D interpolator.

e. Check if the original datapoints need to be scaled according to the interpolator range otherwise transform the original points to scale them as per the interpolator

f. Get the Approximated EigenValues by interpolating the transformed original data points using the interpolator built in (d).

g. Solve for Equation 1 mentioned in paper [Fergus et al](https://cs.nyu.edu/~fergus/papers/fwt_ssl.pdf) to get the eigenfunctions f.

h. Once you have these eigenfunctions, use Gaussian Mixture Model to hard-label these eigenfunctions.

3b) Predict Function:

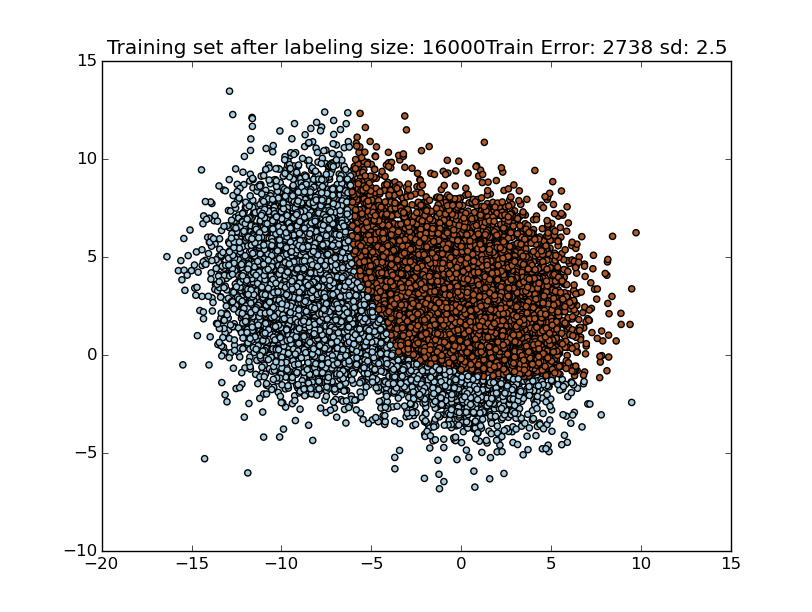
a. Interpolate the unseen data points and get the approximate eigen vectors. Solve for these eigen vectors using the alpha calculated before from Fit method and do hard labeling for these new points.

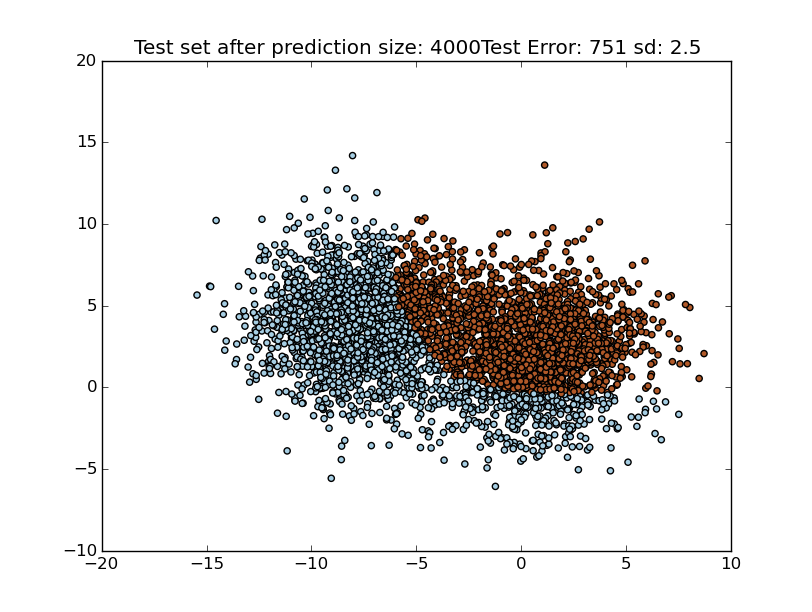
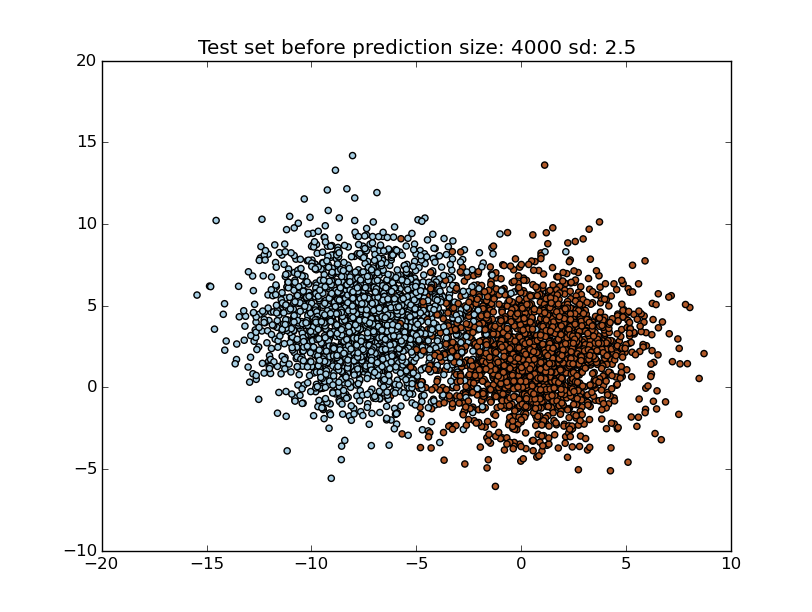
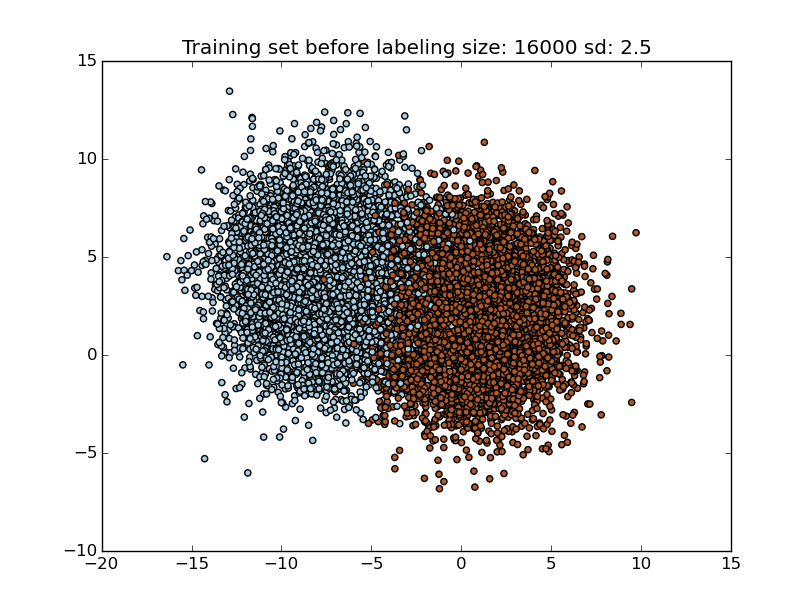
4. I tested for blobs dataset for 20000 datapoints with 5-fold cross validation, 2 clusters, 2 dimensions/features. Below are the results:

1) Variance = 2.4; Dataset = Blobs; Clusters = 2; Features = 2; bins = 10

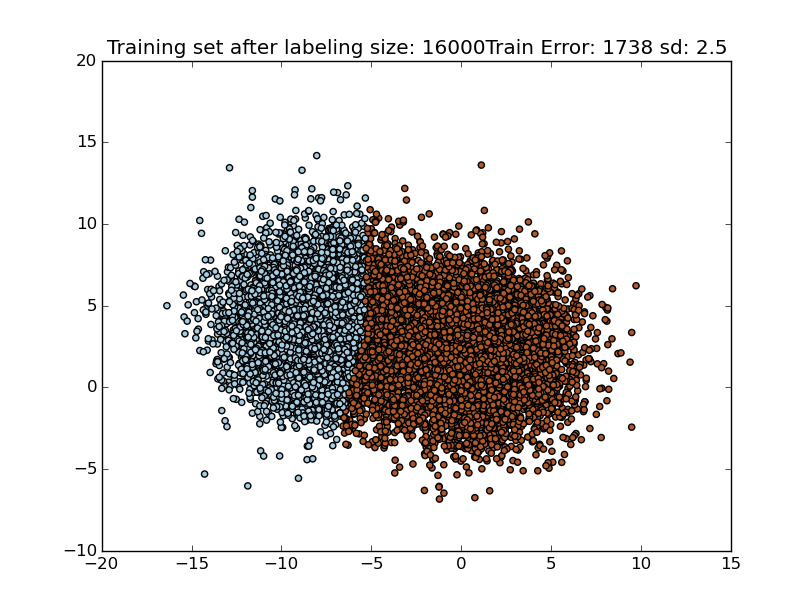
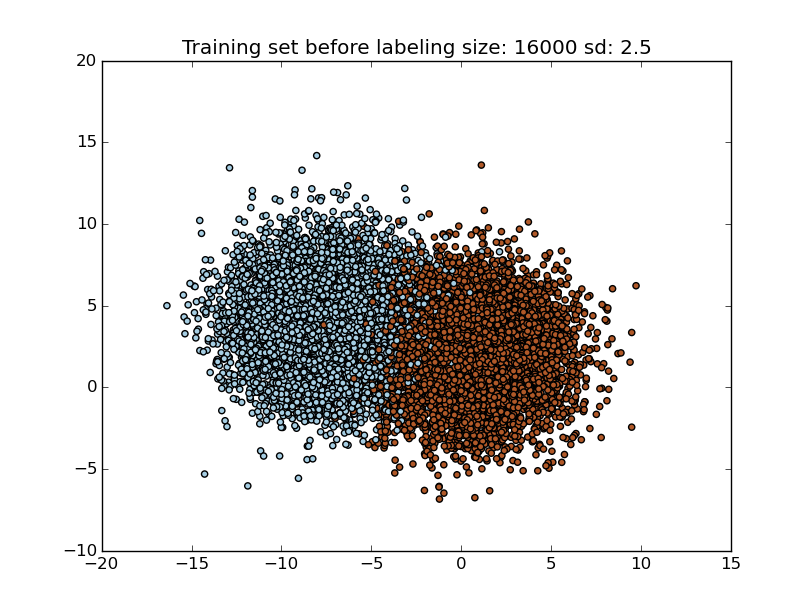
5 fold cross Validation Results (Training set size = 16000, Test set size = 4000)

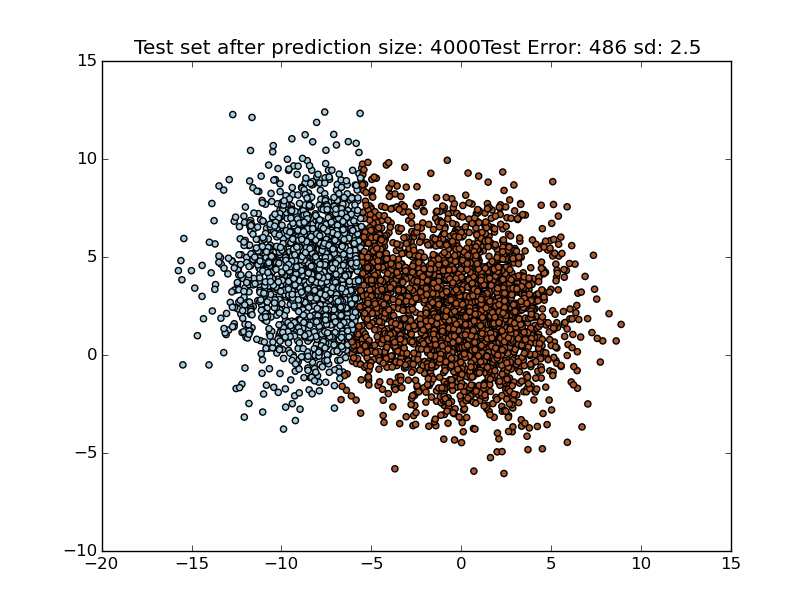
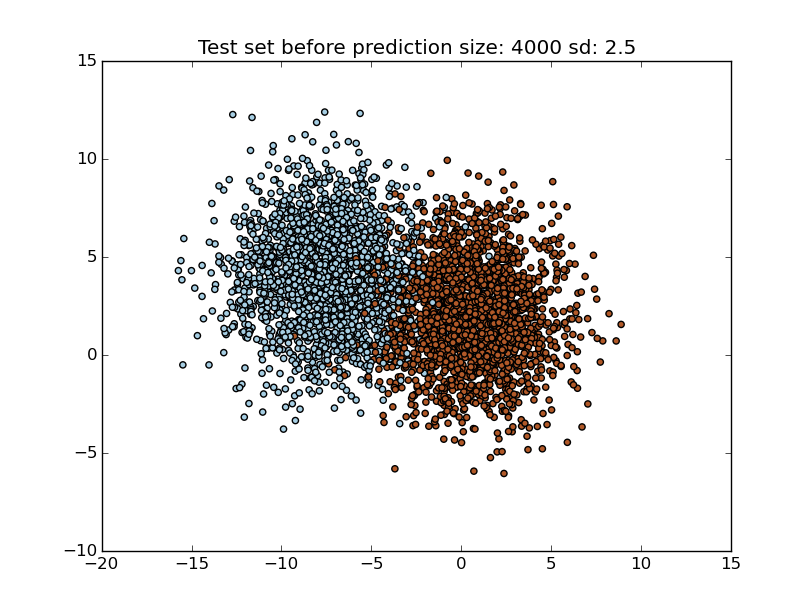
k = 1 Train Error: 0.1711 Test Error: 0.18775

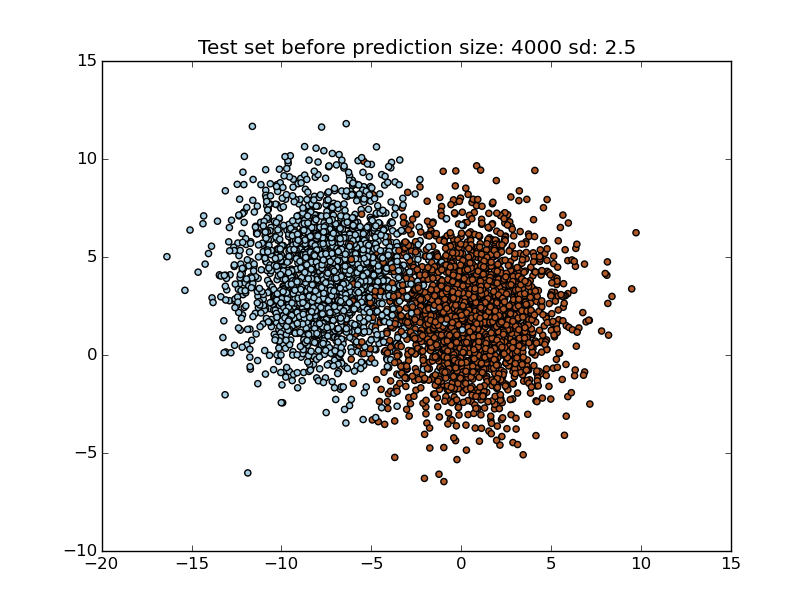
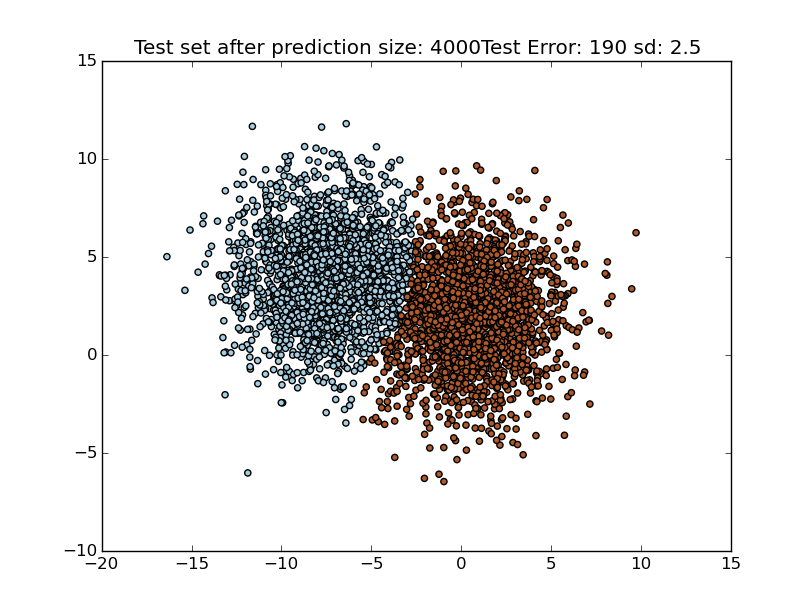
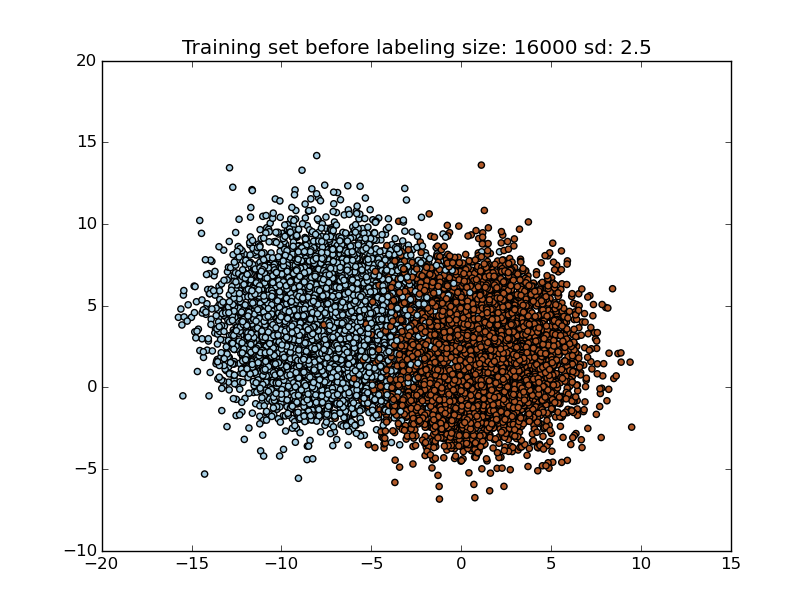
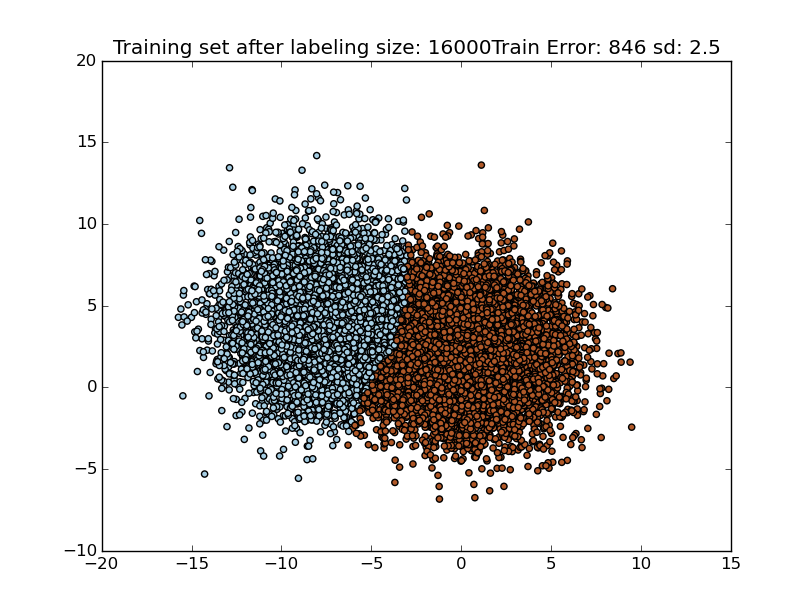




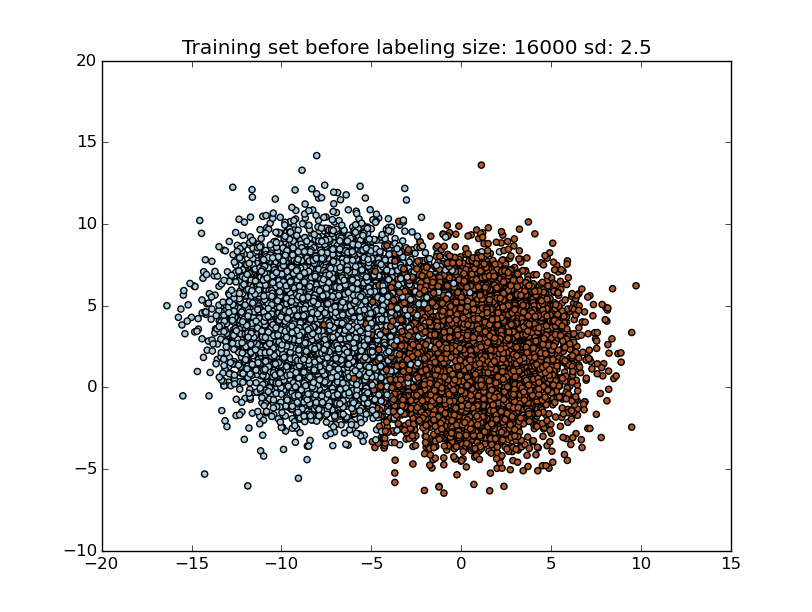
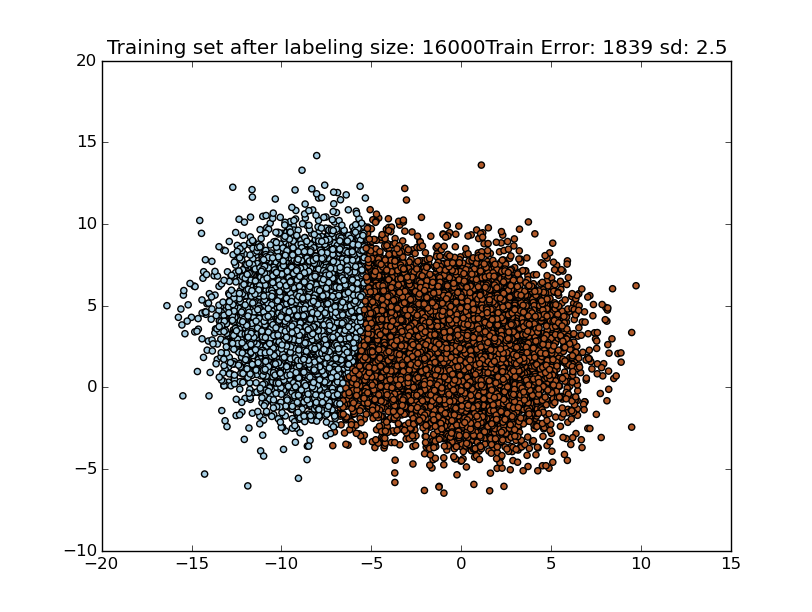
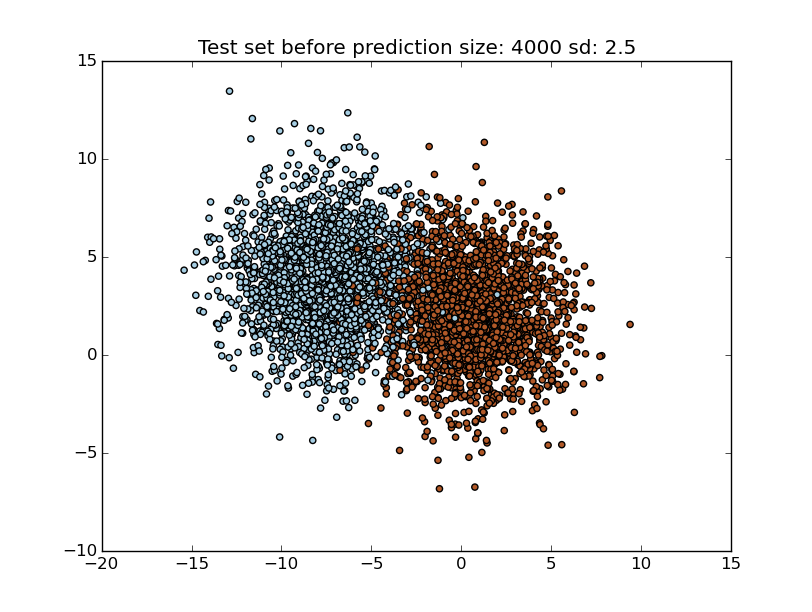
k = 2 Train Error: 0.1086 Test Error: 0.1215

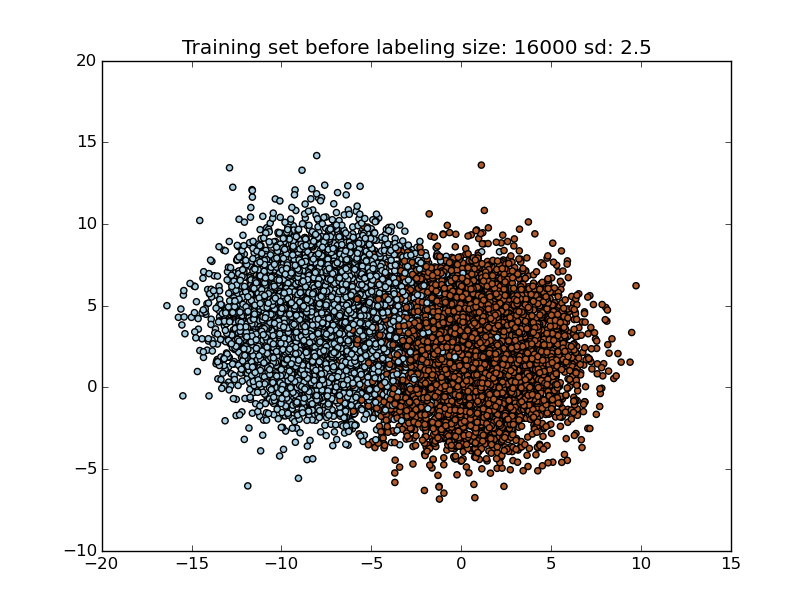


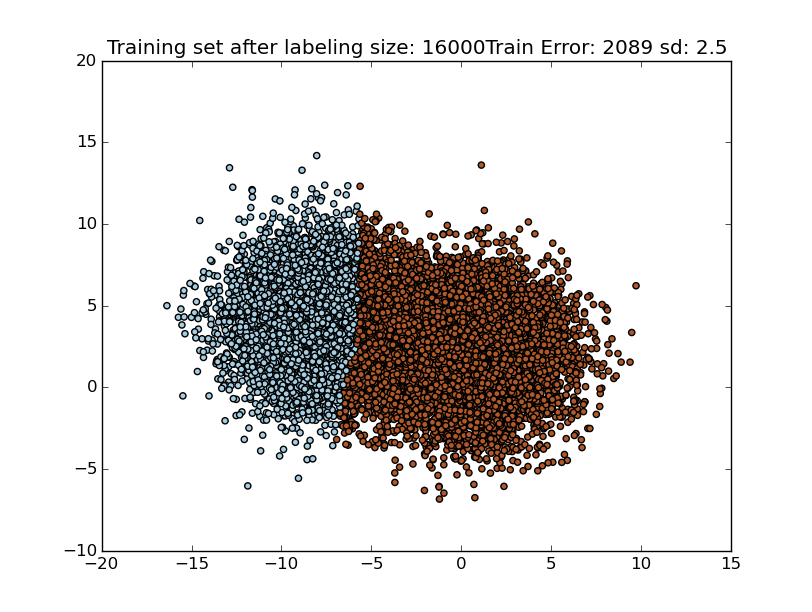
k = 3 Train Error: 0.052875 Test Error: 0.0475

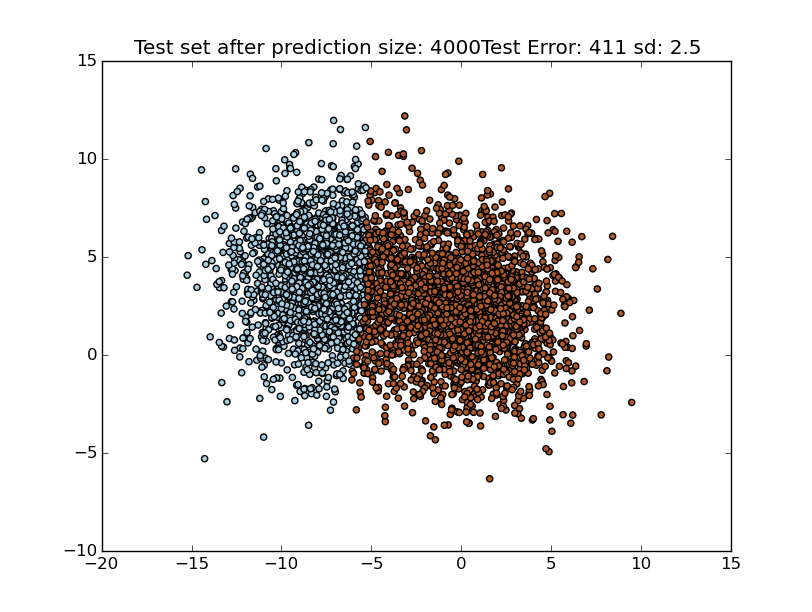


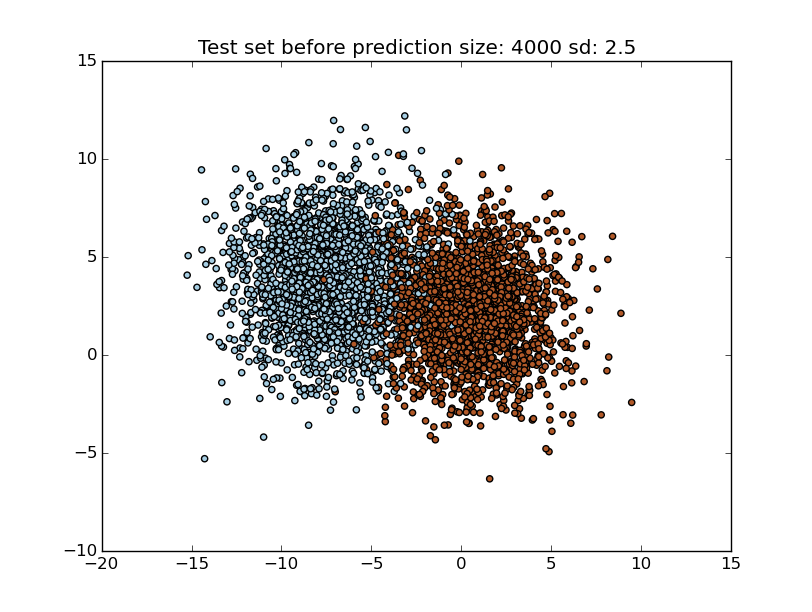
k = 4 Train Error: 0.1149 Test Error: 0.09925



k = 5 Train Error: 0.1306 Test Error: 0.102







ISSUES:

1. I tested for 10 features and 3 clusters but it is not clustering accurately.

2. I have those test results on git. I could only visualize 2 features at a time. I examined the eigenvalues and functions, they seem alright. I was unable to debug the whole issue. I wish to get it resolved once next semester starts or may be during the summer if time permits.