IN4320 Machine Learning Assignment 2

Dilan Gecmen 4221168

March 14, 2018

Note: Worked together with Jasper Hemmes.

In this report we implemented the two following ways of semi-supervised learning for linear discriminant analysis (LDA):

1. **K-Means Clustering Semi Supervised for LDA**

   First we performed K-Means Clustering. K-Means clustering classifies a data set through k clusters. These k clusters are fixed a priori. For K-Means clustering the objective function is to minimize the total intra cluster variance:

   $$\text{Objective} = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

   where k is the number of clusters, n is the number of cases, x_i is case i, c_j is the centroid of cluster j, and $||x_i^{(j)} - c_j||^2$ is the distance function. The algorithm goes as follows:

   1. Initialize the means of the k clusters in the data.
   2. Attribute the closest cluster to each data point. This is done according to the Euclidean distance function.
   3. Set the position of each cluster to the mean of all data points belonging to that cluster.
   4. Repeat steps 2-3 until convergence. This means that in consecutive rounds the same points are assigned to each cluster.

   With K-Means Clustering unlabeled samples are labeled. This sample data is then used as trainings data to train the LDA model.

2. **K-Nearest Neighbors Semi Supervised for LDA** The K-Nearest Neighbors algorithm can be summarized as follows (k is is chosen a priori.):

   1. All euclidean distances between labeled and unlabeled points are calculated.
   2. Select the k smallest distances.
   3. Propagate the labels from the labeled points to the unlabeled points for those k distances.
   4. Repeat steps 1-3 until convergence. This means that all unlabeled points have been assigned a label.

   With K-Nearest Neighbors unlabeled samples are labeled. This sample data is then used as trainings data to train the LDA model.

Now we take the MAGIC Gamma Telescope Data Set from brightspace and normalized all 10 features on the full data set once before all other experiments. That is make all 10 feature standard deviations equal to 1. To standardize the data we use the following equation:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Now based on this normalized data set, we make learning curves against the number of unlabeled samples for a total of 25 labeled samples in the training set. We add 0, 10, 20, 40, 80, 160, 320, and 640 unlabeled samples to the 25 labeled samples. We plotted the expected error rates, see Figure 1.

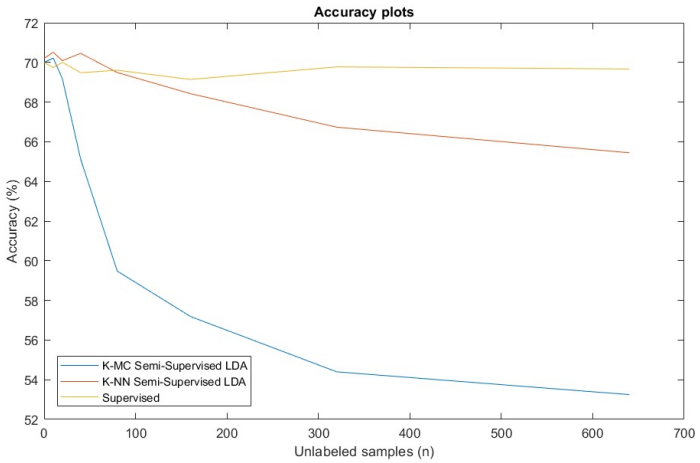

Figure 1: Error rates

As we can see in Figure 1 the results are as expected. We see that K-Nearest Neighbors Semi Supervised performs better than K-Means Clustering Semi Supervised. A reason for this could rely purely on the kind of dataset we used on which KNN performs better than KMC.
When the amount of unlabeled samples is equal to 0, we see that the accuracy of all three approaches stays the same. This is because all three approaches become the same, namely LDA.
The more unlabeled samples we use the higher the uncertainty becomes of a sample being labeled correct. This can also be seen in Figure 1 for both semi supervised approaches.

With the same preprocessed data set we plot the log-likelihood versus the number of unlabeled data, see Figure 2.

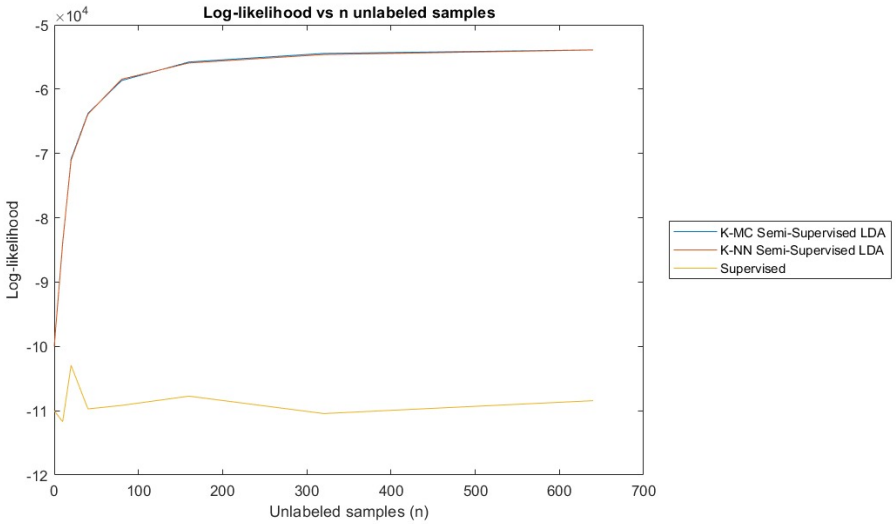

Figure 2: Log-likelihood versus the number of unlabeled data.

The more unlabeled samples we use the higher the log-likelihood becomes for the semi supervised methods. This means that it is less likely that a sample is labeled with the correct label. So, the accuracy of the methods becomes lower when the number of unlabeled samples is increased, as we already stated. For the supervised method we know that when the amount of unlabeled samples is increased the accuracy is quite high and constant. It stays around 70%. This means that for supervised learning it is more likely that an unlabeled sample is labeled correct in contrast to the semi supervised methods.

In the Figure 5 on we constructed two artificial data sets for two features as an example. K-Means Clustering performs better when a linear line can be drawn that divides the samples into 2 clusters (when there are 2 classes), see Figure 3. So, the linear line is drawn between the means of the 2 classes. On this data set K-Nearest Neighbor can fail.
K-Nearest Neighbors calculates all euclidean distances between labeled and unlabeled points. Unlabeled points are labeled according to the labeled points they are closest to. If we look at Figure 4 we see that KNN performs good as the distances between the classes are large, but the distance between the samples is small. Note that KMC will not perform well as the means are very close to each other.

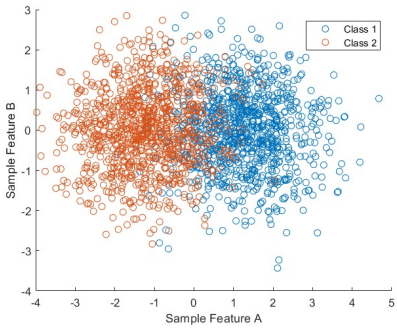

Figure 3: Set on which KMC performs better than KNN

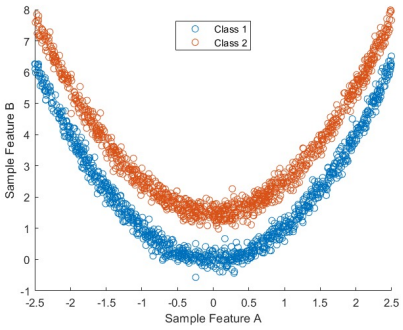

Figure 4: Set in which KNN performs better than KMC.

Figure 5: Artificial data sets for two features to illustrate the performance of KNN versus KMC