Assignment 2 - Semi Supervised Learning

Nirmal Roy, 4724429

March 13, 2018

Question1

Semi supervised learning is the method of labeling unlabeled data and using them to better estimate our model. This in turn may help in increasing testing accuracy of unknown data samples.

For this exercise, we implement two algorithms for the task of semi supervised learning. The ideas were inspired from the slides of Xiaojin Zhu, Department of Computer Sciences University of Wisconsin, Madison, USA[1]. They are described in the following two sub sections:

Expectation Maximization for Mixture of Gaussians

Assumption: Each class has a Gaussian distribution.
Algorithm:

1. We pick up 25 random points with labels from the dataset. And find mean, covariance matrix and prior of each class.

2. We draw the unlabeled points randomly now and initialize the EM algorithm with the means, covariances and priors computed above.

3. For each point x_i, we calculate the posterior probabilities b_i using the Gaussian class conditional density. We update the mean as $\mu_{new} = \frac{\sum b_i * x_i}{\sum b_i}$, covariance as $C_{new} = \frac{\sum b_i * (x_i - \mu_{new})^2}{\sum b_i}$ and priors as $\frac{\sum b_i}{N}$

4. We keep on iterating till we the means and covariance matrices do not change much.

5. Finally, the unlabeled points are given the class to which they have the higher posterior probability.

The Graph Laplacian

Assumption: A graph is given on the labeled and unlabeled data. Instances connected by heavy edge tend to have the same label.
Algorithm:

1. Graph is composed of all labeled and unlabeled data points ($X_l \cup X_u$)

[1] http://pages.cs.wisc.edu/ jerryzhu/pub/sslicml07.pdf

2. We find the squared Euclidean distance D_{ij} of all the points from each other and calculate weights as $W_{ij} = \exp(-D_{ij})$

3. We define a Threshold $T =$ mean($\sum W_{ij}$) - sd($\sum W_{ij}$)

4. Connect nodes i and j if $W_{ij} > T$. Form the weighted adjacency matrix A_w, weights being W_{ij} between two points and the unweighted adjacency matrix.

5. Calculate the Laplacian matrix $L = diag - A_w$ where $diag$ is the diagonal matrix formed by summing each row of A. In other words, degree of each node.

6. Let y be the label vector for $X_l \cup X_u$. We optimize the equation $\frac{1}{2} * y^T * L * y$ where we fix the labels of X_l and subject to the constraint that the labels of unlabeled data $y_u \in [-1, 1]$. We change label 2 to -1 for ease of computation. Also relaxing the labels to in the interval $[-1, 1]$ makes the constraint and cost function convex. The final label vector is such that the overall smoothness of the graph is optimized.

7. Set $y_u = 1$ if $y_u >= 0$. Else, set $y_u = -1$

Question2

As can be seen from figure 1, the graph based algorithm gives better results than the EM algorithm for this dataset. The supervised error with 25 training points is 31%. The graph based method is better than that even after 1000 unlabeled samples. Although for large number of unlabeled data, they will dominate the initial 25 training set and hence both the learners saturate in the 640-1000 range. The EM algorithm although goes below the 31% mark initially, does not do well eventually. I believe the dataset is not suited for EM algorithm and the Gaussian assumption is hurting the algorithm. With larger and larger unlabeled points it is getting harder for the algorithm to model the dataset. We can see that the best case scenario for our learner is an error of about 28% which quite close to the 31% mark. I think this because we are getting the unlabeled data randomly from the dataset itself along with the training set. Hence, a considerable improvement from the supervised learner cannot be expected.

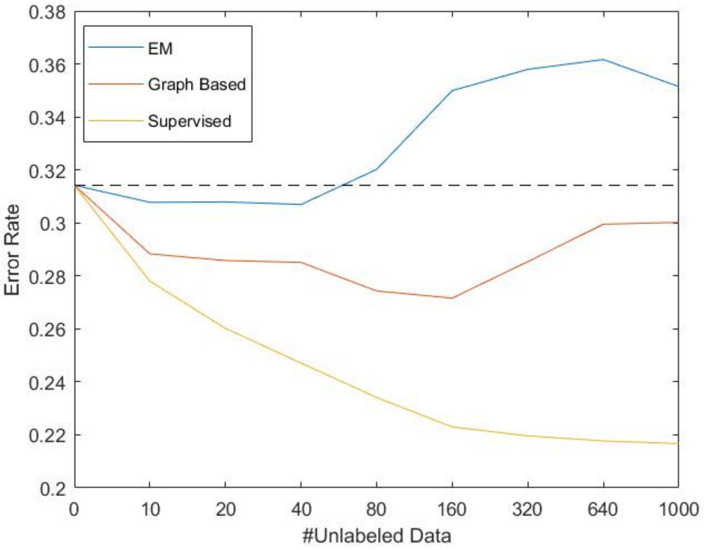

Figure 1: Plot of the error rate vs number of unlabeled data for our algorithms. The yellow line is the plot of the error rate for supervised learner considering $X_u + X_l$ training points. The dashed line at 0.3142 is a reference line for the error on 25 training samples(without any unlabeled data) randomly chosen from the dataset. All experiments are repeated 250 times.

Question3

Plotting log-likelihood on the test set gives an estimate of the confidence of the classifier in its performance. The plots saturate after addition of about 200 unlabeled data. Which means the algorithms found the best set of parameters around that mark to model the data. It is important to note that the plot gives an indication of the classifier's confidence, which doesn't increase much after 200 unlabeled points, and might not be reflective of the actual error. But we can intuitively say from the error plot that graph based method should have higher likelihood than EM algorithm and as seen from figure 2, that indeed is the case.

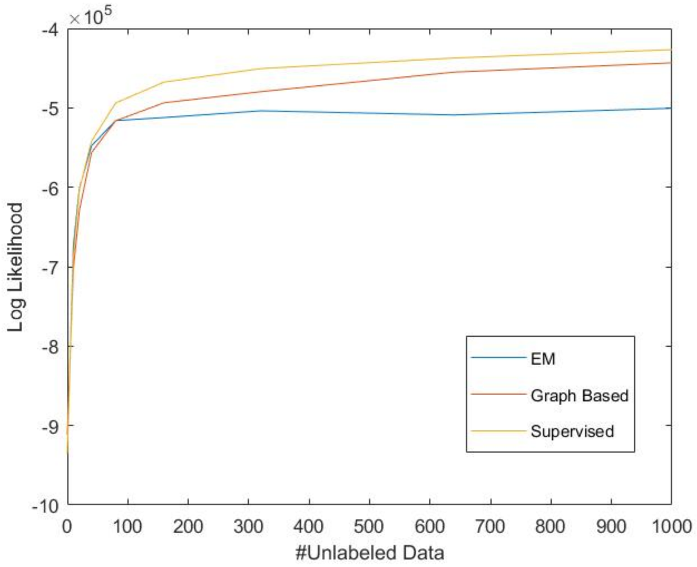

Figure 2: Plot of the log-likelihood on the test set vs number of unlabeled data for our algorithms. The yellow line is the plot of the log-likelihood for supervised learner considering $X_u + X_l$ training points. All experiments are repeated 250 times.

Question4

Intuitively, EM algorithm should work better when we have proper distributions for the two classes that our algorithm can model under the Gaussian assumption. Whereas, the graph based algorithm will have better performance when the two classes are separated from each other and points from each class are near to each other. This is a technique of label propagation and the algorithm will be bad if the two classes overlap in any way. Because in that case, labels of one class might propagate to the other class, causing a decrease in accuracy. Keeping these intuitions in mind, we generated two datasets: **banana dataset** and **Higleyman dataset** [figure 3].

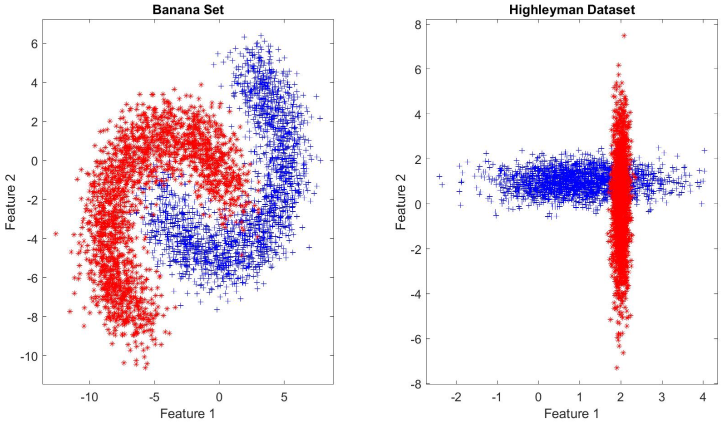

Figure 3

The results are tabulated below. All experiments have been averaged over 250 runs and we have taken 1000 unlabeled samples to compare the results.

Table 1: Error rates for the learners on the two datasets

	Banana Classes	**Highleyman Classes**
EM algorithm	18.2 %	16.8%
Graph based algorithm	15.16%	22.71%
Supervised Error(25 training points)	16.41 %	20.58 %

We were right with our intuitions. The highleyman dataset with its high overlap affects the distance based graph algorithm, picking up wrong links and reducing performance accuracy. Where as the EM algorithm can model the two Gaussian well and give good results. On the other hand, EM fails with the banana dataset given its shape. Whereas the graph based model picks up the local distance between the points really well and gives better results than normal LDA.