IN4320 Machine Learning: Assignment 2
Semi-Supervised Learning

Dimitropoulos Georgios: 4727657

March 13, 2018

1 Real

1.1 A)

1st way of semi-supervised learning for LDA:

Step 1) Setting: We are given a dataset which contains a lot of labeled data. Also it contains a lot of unlabeled data. We have as a classifier the two-class linear discriminant analysis (LDA) which is based on the assumption that the class-conditional distributions are Gaussian with the same covariance matrix and its parameters are estimated through maximum likelihood (ML).
Step 2) We split the labeled data into training and test set and we use the training data in order to train our classifier and we test its performance with the test set. Hence, we measure the performance of our initial supervised classifier, namely the LDA in our case.
Step 3) After that, we want to investigate if the addition of extra unlabeled data is going to increase the performance of our classifier (supervised learner). The first way that we propose in order to perform the semi-supervised learning for the LDA is to perform a clustering algorithm in the unlabeled data (and particularly the K-Means algorithm) which finds clusters (majority of labels). After that, the output of the K-Means algorithm is going to give labels to our unlabeled data. Our reasoning behind clustering (majority) based on the global consistency assumption which gives information about the conditional probabilities, is that that we are going to have a better and bigger labeled data set and this is going to improve the performance of our classifier.
Step 4) Now, that we trained our classifier with more data, we are going to test it with the test set in order to investigate if the extra amount of data which we used to train it, increases its performance.
Step 5) Finally, we are going to check, by adding more and more unlabeled samples, which take their labels through the K-Means algorithm that we described before, how the expected error rate changes and to compare it with the initial supervised error rate in order to see if our first proposed way of semi-supervised learning for LDA improves its performance and for which number of unlabeled samples this is the case.

2nd way of semi-supervised learning for LDA:

Step 1) Setting: Again, we are given a dataset which contains a lot of labeled data. Also it contains a lot of unlabeled data. We have as a classifier the two-class linear discriminant analysis (LDA) which is based on the assumption that the class-conditional distributions are Gaussian with the same covariance matrix and its parameters are estimated through maximum likelihood (ML).
Step 2) We split the labeled data into training and test set and we use the training data in order to train our classifier and we test its performance with the test set. Hence, we measure the performance of our initial supervised classifier, namely the LDA in our case.
Step 3) After that, we want to investigate if the addition of extra unlabeled data is going to increase the performance of our classifier (supervised learner).The second way that we propose in order to perform the semi-supervised learning for the LDA is based on a semi-supervised setting which is called self learning (Yarowsky algorithm) [1],[2]. In this approach, firstly, the most "confident" sample (the one that has the largest distance from the decision boundary, that we have after the first time that we applied our supervised classifier, which is trained in the labeled training set)

takes the same label as the label that the nearest labeled sample point to this has. So we started with our classifier which is trained on the labeled data only, and after that, labels are predicted for the unlabeled objects with the above self learning procedure. These objects, with the labels which they took through the above procedure are used in the sequel to retrain our classifier. This new classifier is now used to relabel the unlabeled objects.
Step 4) Now, that we trained our classifier with more data, we are going to test it with the test set in order to investigate if the extra amount of data which we used to train it, increases its performance.
Step 5) Finally, we are going to check, by adding more and more unlabeled samples, which take their labels through the above self learning procedure that we described before, how the expected error rate changes and to compare it with the initial supervised error rate in order to see if our second proposed way of semi-supervised learning for LDA improves its performance and for which number of unlabeled samples this is the case.

1.2 B)

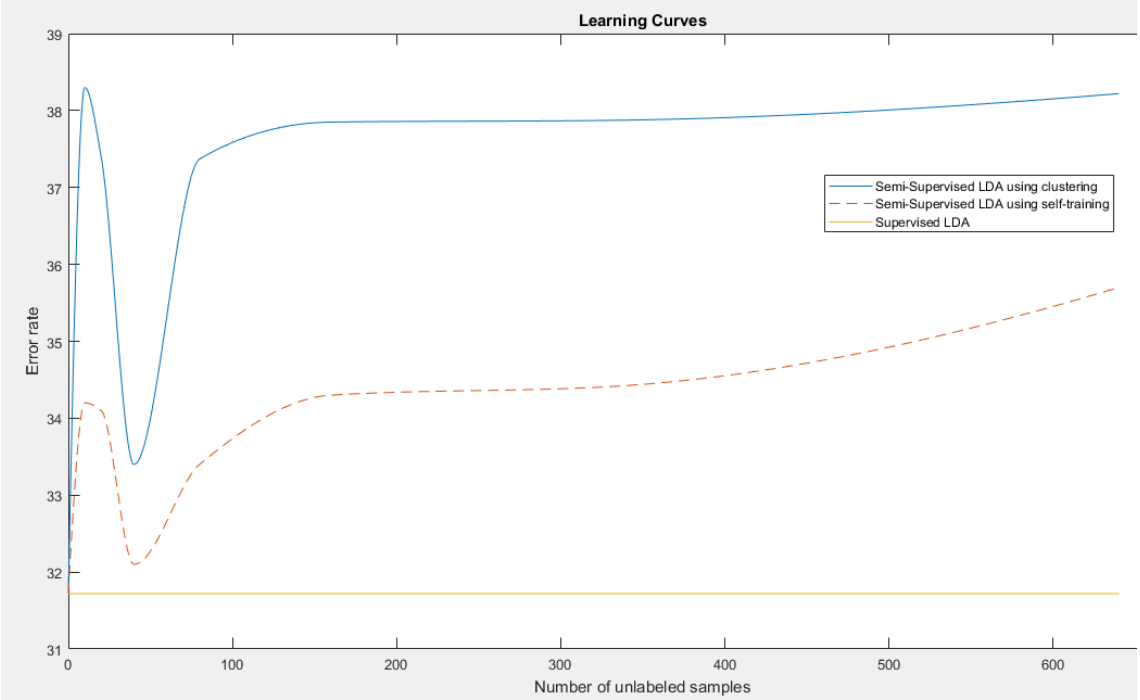

Figure 1: Error rate against the number of unlabeled samples for the two proposed ways of semi-supervised learning and the supervised learning on the normalized MAGIC Gamma Telescope Data Set.

In this question, we worked on the MAGIC Gamma Telescope Data Set. At first, we normalized all 10 features in this full data set. In the sequel we split the dataset set in training and testing set. After that, based on this normalized data we trained our classifier (LDA) on 25 randomly selected labeled data samples and we tested its performance with the testing set. We repeated this procedure several times and we took the average error which was 31.72%. Sequentially, we performed the semi-supervised learning for the LDA with the two approaches which we proposed above in a) by adding 0, 10, 20, 40, 80, 160, 320, and 640 unlabeled samples respectively. Each time we trained our LDA with the extra amount of unlabeled data and tested its performance with the testing set.Based on these experiments we made the learning curves against the number of unlabeled samples for a total of 25 labeled samples in the training set, which can be depicted in the above figure 1. In this figure we can depict the error rate against the number of unlabeled samples for the two proposed ways of semi-supervised learning (using clustering and using self-learning) and the supervised learning. As we can see from the figure, by adding a few unlabeled data (10) the error of both ways of semi-supervised learning is increased (38.3% for the first semi-supervised

approach(clustering) and 34.2% for the second semi-supervised approach(self-training)). After that by adding a more few unlabeled data (20) the error of both ways of semi-supervised learning is slightly decreased(37.4% for the first semi-supervised approach(clustering) and 34.1% for the second semi-supervised approach(self-training)).In the sequel adding more few unlabeled data (40) the error of both ways of semi-supervised learning is decreased significantly(33.4% for the first semi-supervised approach(clustering) and 32.1% for the second semi-supervised approach(self-training)). However, in case that we add even more unlabeled data (80) the error is increased significantly again (37.37% for the first semi-supervised approach(clustering) and 33.4% for the second semi-supervised approach(self-training)). Finally, by adding new unlabeled data (160,320,640) the error rate seems to remain stable with a minor increase(37.85%, 37.87%, 38.22% for the first semi-supervised approach(clustering) and 34.3%, 34.3%, 35.7% for the second semi-supervised approach(self-training) respectively). Hence, we conclude that by keep adding unlabeled data has a negative impact and the error starts to be increased again in this specific dataset. Also we notice that the supervised learning outperforms the two proposed ways for semi-supervised learning in terms of error rate in this specific dataset.

1.3 C)

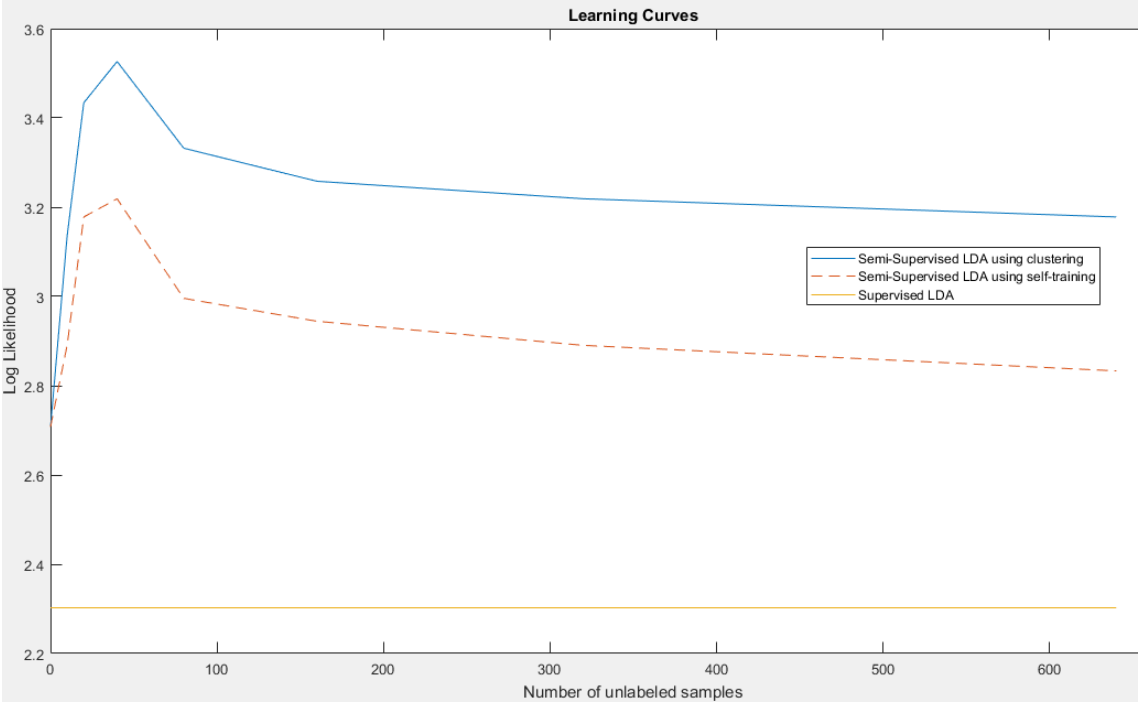

Figure 2: Log Likelihood against the number of unlabeled samples for the two proposed ways of semi-supervised learning and the supervised learning on the normalized MAGIC Gamma Telescope Data Set.

In this question, we made the same type of plots. However, in this case we plotted the log-likelihood versus the number of unlabeled data with the same preprocessed data set as in the previous question b). As it can be depicted in the above figure 2, the log likelihood for both the two proposed ways of semi-supervised learning is increased in case we add few unlabeled samples (10,20,40). After that in case we add more unlabeled samples (80) the log likelihood for both the two proposed ways of semi-supervised learning is slightly decreased. Finally by adding even more unlabeled samples (160,320,640) the log likelihood for both the two proposed ways of semi-supervised learning seems to be stable.Hence, based on the figures 1 and 2 we can conclude that the error rate and the log-likelihood are not monotonically related in this particular dataset.

2 Imaginary

2.1 D)

1st Artificial Dataset
Our first imaginary dataset consists of the points (-4,0),(-3,0),(-2,0), (-1,0),(1,0),(2,0),(3,0),(4,0) and (1,6).So this dataset contains 9 samples in the 2-d space, thus each point has two features. This dataset can be depicted in the figure 3, where the true labels of the samples are shown. Namely we have a 2-class problem, where the class 1 consists of the points (-4,0),(-3,0),(-2,0),(-1,0) and the class 2 of the points (1,0),(2,0),(3,0),(4,0),(1,6) respectively. However, these are the true labels and we do not know them in this imaginary scenario.
Let say that we have only 2 random labeled sample points (let suppose the (-4,0) and (1,6)). So these 2 points consist our training set. After that, we train our LDA classifier with these 2 points and we test it with the remaining 7 points. The result of this classifier can be depicted in figure 4. We can see that we achieve an error rate of 14.29% (point (1,0) is misclassified).
In the sequel, we want to investigate if the addition of extra unlabeled data is going to increase the performance of our classifier (supervised learner). Hence, we implemented the first way that we proposed in order to perform the semi-supervised learning for the LDA. Namely we want to add one unlabeled data. Let say that we want to add the sample point (-3,0). So we perform the K-Means algorithm which is going to give label to this unlabeled data. We run the algorithm and the point (-3,0) took the label 1.In the sequel we retrain our classifier with these 3 points (-4,0),(1,6) and (-3,0) and we test it with the remaining 6 points. The result of this classifier can be depicted in figure 5. We can see that we achieved an error rate of 66.6% (points (1,0),(2,0),(3,0),(4,0) are misclassified).
After that we repeat this procedure. Let say that we want to add now the sample point (-2,0). So we perform the K-Means algorithm which is going to give label to this unlabeled data. We run the algorithm and the point (-2,0) took the label 1.In the sequel we retrain our classifier with these 4 points (-4,0),(1,6),(-3,0) and (-2,0) and we test it with the remaining 5 points. The result of this classifier can be depicted in figure 6. We can see that we achieved an error rate of 80% (points (1,0),(2,0),(3,0),(4,0) are misclassified).

Now we want to investigate if the implementation of our second way that we proposed in order to perform the semi-supervised learning for the LDA is going to increase the performance of our classifier (supervised learner). Again, let say that we have only 2 random labeled sample points (let suppose the (-4,0) and (1,6)). So these 2 points consist our training set. After that, we train our LDA classifier with these 2 points and we test it with the remaining 7 points. The result of this classifier can be depicted in figure 4. We can see that we achieved an error rate of 14.29% (point (1,0) is misclassified).
Now, we want to add again one unlabeled data.So we perform the self-learning algorithm now which is going to label firstly, the most "confident" sample (the one that has the largest distance from the decision boundary, that we have after the first time that we applied our supervised classifier, which is trained in the labeled training set). As it can be depicted form the figure 4 the most "confident" sample is the point (-3,0) which takes the same label as the label that the nearest labeled sample point to this has. So it going to take the label 1.In the sequel we retrain our classifier with these 3 points (-4,0),(1,6) and (-3,0) and we test it with the remaining 6 points. The result of this classifier can be depicted in figure 7. We can see that we achieved an error rate of 66.6% (points (1,0),(2,0),(3,0),(4,0) are misclassified).
After that we repeat this procedure. As it can be depicted form the figure 7 the most "confident" sample now is the point (4,0) which takes the same label as the label that the nearest labeled sample point to this has. So it going to take the label 2. In the sequel we retrain our classifier with these 4 points (-4,0),(1,6),(-3,0) and (4,0) and we test it with the remaining 5 points. The result of this classifier can be depicted in figure 8. We can see that we achieved an error rate of 0% as all points are classified correctly.
To conclude, in this 1st Artificial dataset we were able to prove the following inequality(two unlabeled points were added to the two training samples and were tested with the remaining five points) :

Error rate of LDA after Self-Learning=0%<error rate of supervised LDA=14.29%<error rate of LDA after clustering=80%

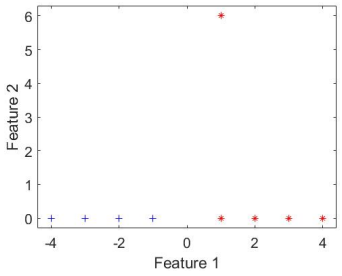

Figure 3: 1st Artificial Dataset

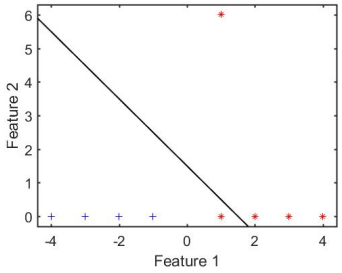

Figure 4: LDA on the 1st Artificial Dataset

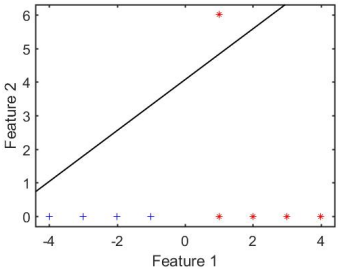

Figure 5: LDA after Clustering with one unlabeled data

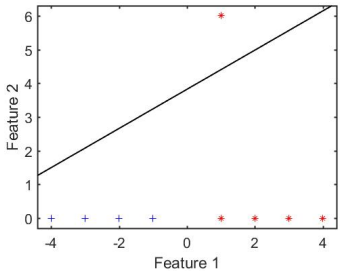

Figure 6: LDA after Clustering with two unlabeled data

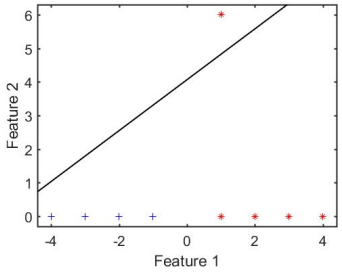

Figure 7: LDA after Self-Learning with one unlabeled data

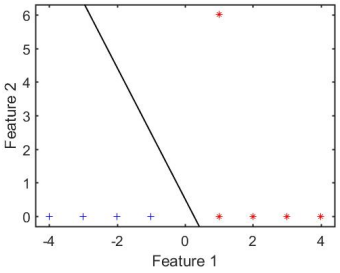

Figure 8: LDA after Self-Learning with two unlabeled data

In the second artificial dataset we are going to prove the other way around.

2nd Artificial Dataset

Our second imaginary dataset consists of the points (0,-4),(0,-3),(0,-2),(0,-1),(0,1),(0,2),(0,3),(0,4) and (6,1). So this dataset contains 9 samples in the 2-d space, thus each point has two features. This dataset can be depicted in the figure 9, where the true labels of the samples are shown. Namely we have a 2-class problem, where the class 1 consists of the points (0,-4),(0,-3),(0,-2), (0,-1) and the class 2 of the points (0,1),(0,2),(0,3),(0,4),(6,1) respectively. However, these are the true labels and we do not know them in this imaginary scenario.

Let say that we have only 2 random labeled sample points (let suppose the (0,-4) and (0,1)). So these 2 points consist our training set. After that, we train our LDA classifier with these 2 points and we test it with the remaining 7 points. The result of this classifier can be depicted in figure 10. We can see that we achieve an error rate of 14.29% (point (0,-1) is misclassified).

In the sequel, we want to investigate if the addition of extra unlabeled data is going to increase the performance of our classifier (supervised learner). Hence, we implemented the first way that we proposed in order to perform the semi-supervised learning for the LDA. Namely we want to add one unlabeled data. Let say that we want to add the sample point (0,2). So we perform the K-Means algorithm which is going to give label to this unlabeled data. We run the algorithm and the point (0,2) took the label 2.In the sequel we retrain our classifier with these 3 points (0,-4),(0,1) and (0,2) and we test it with the remaining 6 points. The result of this classifier can be depicted in figure 11. We can see that we achieved an error rate of 33.3% (points (0,-1) and (6,1) are misclassified).

After that we repeat this procedure. Let say that we want to add now the sample point (-2,0). So we perform the K-Means algorithm which is going to give label to this unlabeled data. We run the algorithm and the point (-2,0) took the label 1.In the sequel we retrain our classifier with these 4 points (0,-4),(0,1),(0,2) and (-2,0) and we test it with the remaining 5 points. The result of this classifier can be depicted in figure 12. We can see that we achieved an error rate of 0% as all points are classified correctly.

Now we want to investigate with what way the implementation of our second way that we proposed in order to perform the semi-supervised learning for the LDA is going to affect the performance

of our classifier (supervised learner). Again, let say that we have only 2 random labeled sample points (let suppose the (0,-4) and (0,1))). So these 2 points consist our training set. After that, we train our LDA classifier with these 2 points and we test it with the remaining 7 points. The result of this classifier can be depicted in figure 10. We can see that we achieved an error rate of 14.29% (point (0,-1) is misclassified).
Now, we want to add again one unlabeled data.So we perform the self-learning algorithm now which is going to label firstly, the most "confident" sample (the one that has the largest distance from the decision boundary, that we have after the first time that we applied our supervised classifier, which is trained in the labeled training set). As it can be depicted form the figure 10 the most "confident" sample is the point (6,1) which takes the same label as the label that the nearest labeled sample point to this has. So it going to take the label 2.In the sequel we retrain our classifier with these 3 points (0,-4),(0,1) and (6,1) and we test it with the remaining 6 points. The result of this classifier can be depicted in figure 13. We can see that we achieved an error rate of 16.6% (point (0,-1) is misclassified).
After that we repeat this procedure. As it can be depicted form the figure 13 the most "confident" sample now is the point (0,4) which takes the same label as the label that the nearest labeled sample point to this has. So it going to take the label 2. In the sequel we retrain our classifier with these 4 points (0,-4),(0,1), (6,1) and (0,4) and we test it with the remaining 5 points. The result of this classifier can be depicted in figure 14. We can see that we achieved an error rate of 20% (point (0,-1) is misclassified).
To conclude, in this 2nd Artificial dataset we were able to prove the following inequality(two unlabeled points were added to the two training samples and were tested with the remaining five points) :

Error rate of LDA after clustering=0%<Error rate of supervised LDA=14.29%<Error rate of LDA after Self-Learning=20%

As a conclusion we could state that in the second imaginary dataset, the global consistency assumption leads to a better and bigger labeled training set and so our first way of semi-supervised learning leads to improvements. On the hand, the nature of the 1st imaginary dataset and the rational of our second way of semi-supervised learning leads to improvements in the first case.

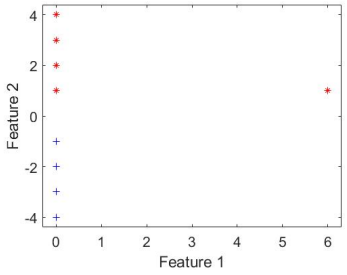

Figure 9: 2nd Artificial Dataset

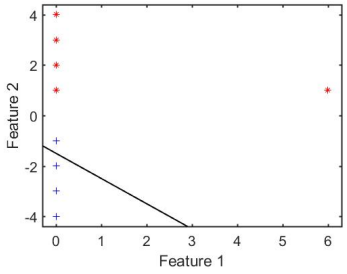

Figure 10: LDA on the 2nd Artificial Dataset

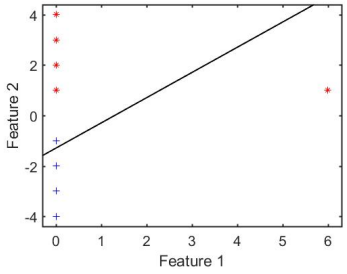

Figure 11: LDA after Clustering with one unlabeled data

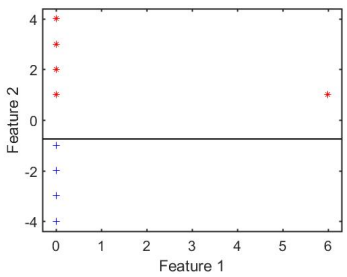

Figure 12: LDA after Clustering with two unlabeled data

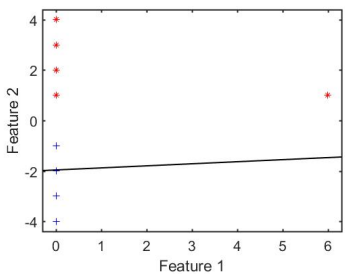

Figure 13: LDA after Self-Learning with one unlabeled data

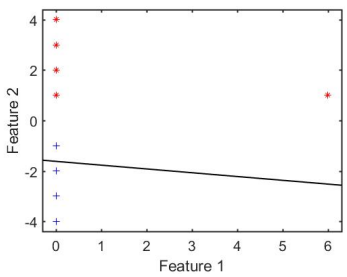

Figure 14: LDA after Self-Learning with two unlabeled data

References

[1] *G. J. McLachlan, "Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis," Journal of the American Statistical Association, vol. 70, no. 350, pp. 365–369, 1975*

[2] *D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pp. 189–196, 1995.*