IN 4320 Machine Learning Assignment 2

Xiangwei Shi
4614909

Real

a. The first way of semi-supervised learning for LDA that I decide to use is self-learning from [1]. To be explicit, there are several steps.

Firstly, to start self-learning method, the initial LDA will be trained by the labeled data. Secondly, assign new unlabeled data with the existing labels by the LDA. Thirdly, using the original labeled data and newly-labeled data retrain the LDA. And then repeat the step two and three. Once the stop criterion, which will be explained later, is satisfied, the iterations will stop and the final LDA will be decided. The stop criterion applied is that the difference of the predicted labels for unlabeled data from two successive iterations is small enough or the times of iterations is big enough. For instance, when the average of L1 distance or L2 distance of the predicted labels from two successive iterations is smaller than 10^{-5} or the times of iteration is larger than 100, the stop criterion is satisfied. The reason of setting such stop criterion can be summarized as: 1) the small average difference of predicted labels means the converge; 2) the times of iterations guarantee the iteration will be stopped when the data is not linear separable.

The second way of semi-supervised learning for LDA that I will use is a method from [2]. The steps for this method are presented.

Firstly, pre-whiten all the data. Secondly, calculate the overall mean of the labeled data m and overall mean of data μ. Next, calculate the total covariance matrix over all labeled data T and the total covariance matrix of all data Θ. And then, transform the labeled data using the equation of $x \leftarrow \Theta^{\frac{1}{2}} T^{-\frac{1}{2}}(x-m)+\mu$. In this step, Moore-Penrose generalized matrix inverse replaces the inverse. Lastly, train the LDA with the transformed data.

b. Before drawing the learning curves, some steps for preprocessing the data have been implemented. Firstly, in the raw data, the two classes are labeled as 'g' and 'h'. I converted the two labels into 0 and 1, respectively. Secondly, for both semi-supervised learning methods explained above and the supervised learning, I repeated the experiment for 50 times. Lastly, for the first method, self-learning, the average distance between predicted labels of two iterations is set as 10^{-5} and the iteration limit is set as 1000.

The figure of the learning curves against the number of unlabeled samples for a total of 25 labeled samples in the training set is shown below, as Figure 1. From the figure, we can get that both error rates are decreasing as adding more unlabeled samples until the number of unlabeled samples is large enough, such as 160 for method 2 and 320 for method 1. Compared to the supervised error rates, the average error rates with small number of unlabeled samples of two semi-supervised learning methods are larger. Moreover, the first semi-supervised method could get similar error rates when large number of unlabeled samples are adding for training, which means there is a great decrease at the beginning of adding unlabeled samples. However, the error rates of the second semi-supervised learning method are always larger than the supervised one. For both semi-supervised learning methods, significant change in error rates happen on the begining of adding unlabeled samples for training, which means that both

semi-supervised learning methods improve the LDC performance after adding unlabeled samples.

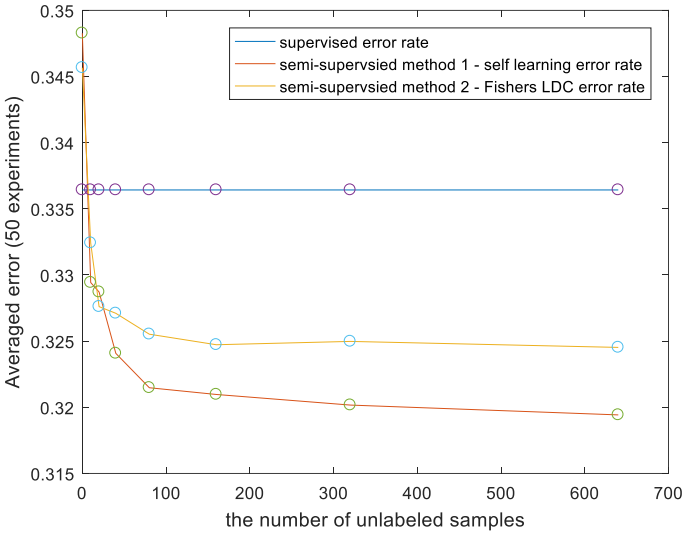

Figure 1 The learning curves against the number of unlabeled samples

c. The figure of log-likelihood of two semi-supervised learning methods is shown below as Figure 2. The steps of preprocessing data are the same as that of b. The equation for calculating the log-likelihood for semi-supervised learning can be summarized as

$$loglikelihood = \sum loglikelihood_{labeled\_0} + \sum loglikelihood_{labeled\_0} + \sum loglikelihood_{unlabeled}.$$

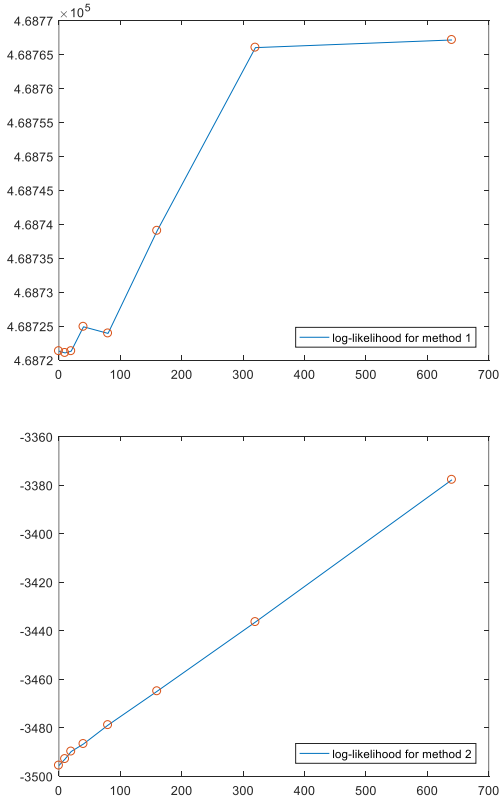

Figure 2 Log-likelihood of two semi-supervised learning methods

Unlike the error rates, both log-likelihood curves against the number of unlabeled samples

represent an increasing trend, which means the performance of both semi-supervised methods improve as increasing the number of unlabeled samples. However, comparing both log-likelihood curves, there are some differences, such as the values of log-likelihood and trend after adding large number of unlabeled samples for training. The possible reason for such quite difference of log-likelihood values is that the second semi-supervised method whitens the data and the distribution of the data will be changed. Another reason could be these methods are different from each other which generate the different LDCs. As for the non-increasing trend of the second log-likelihood curve, one possible explanation is adding more unlabeled samples than 640 could increase the log-likelihood value.

Imaginary

d. To generate the first data, which makes the first method outperforms the supervised method and the second method performs the opposite. The scatter plot and the classifiers of supervised method and two semi-supervised methods are shown below as Figure 3.

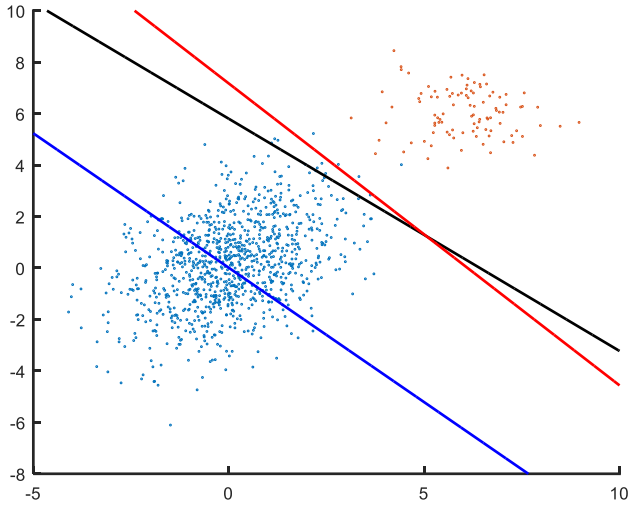

Figure 3 Scatter plot of the first dataset and three classifiers (black: supervised, red: method 1, blue: method 2)

As shown in Figure 3, the first dataset is made of two classes in 2-dimensional space. The data from both classes are Gaussian-distributed with the mean of (0,0) and (3,3), and with the covariance matrixes of [2 1; 1 3] and [1 0; 0 1]. And the numbers of the data from each class are 1000 and 100. To implement the semi-supervised methods, 25 labeled data and 25 unlabeled data are used for training. The constraints and preprocessing remain the same as b and c. The test errors of supervised method and both semi-supervised methods are 0.0056, 0.0020 and 0.2949. The first method outperforms the supervised learning in terms of error rate, however, the second method is doing the opposite.

The possible reason for this result could be the second semi-supervised method is not robust to the unbalanced data because of the transforming data. As for the first method, it is robust to the unbalanced data and adding unlabeled samples helps improve the performance. To verify this explanation, I changed the number of samples from two classes into 1000 and 1000. And the error rates become 0.0096, 0.0071 and 0.0035, where we can find that method 2 improves the performance with balanced data.

The second dataset is also made of two classes in 2-dimensional space. I generated 2 Gaussian-

distributed data with the mean of (0,0) and (0,4), and with the same covariance matrix, [1 0;0 1]. And the two classes are divided by whether their first feature is larger than 0 or not. The numbers of training data and unlabeled samples are the same as the first data and the data processing is the same as question b or c. The scatter plot of the data and three classifiers is shown below, as Figure 4.

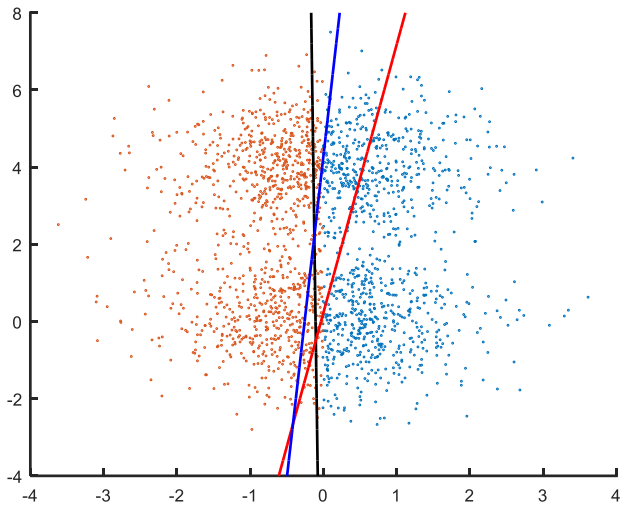

Figure 4 Scatter plot of the second dataset and three classifiers (black: supervised, red: method 1, blue: method 2)

As shown in Figure 4, the data is clearly linear separable. The numbers of data from two classes are almost the same. Therefore, supervised LDA can get a low error rate with 0.0481. In this scenario, method 2 performs slightly better than the supervised LDA with error rate as 0.0461. The first semi-supervised learning method performs the worst with error rate as 0.0577. The possible reason for this result could be the transformation in method 2 does not affect the distribution of labels and adding unlabeled samples improve the performance. However, the first method is initialized by supervised LDA and each iteration will lead the classifier into the direction with false labeled data, which could result in the error rate less than that of supervised LDA.

Reference:

[1] Yarowsky, D.: *Unsupervised word sense disambiguation rivaling supervised methods.* In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, pp. 189–196 (1995)

[2] Loog, Marco. *"Semi-supervised linear discriminant analysis using moment constraints."* IAPR International Workshop on Partially Supervised Learning. Springer, Berlin, Heidelberg, 2011.