Semi-supervised Sparse LDA

Gudmundur Einarsson
DTU Compute – Technical University of Denmark
guei@dtu.dk

Overview

- A semi-supervised extension of the supervised classification method sparse-LDA [3] using a regularizer similar to [1], motivated from spectral dimensionality reduction.
- Experimental results on the mnist data set [4] to evaluate the gain of adding unlabelled samples.
- Experimental results on the ECG part of the UCR time-series data set [2].
- Brief analysis on the added runtime.

Introduction

LDA is a method used for supervised classification and a sparse version of it, using an l_1-norm on the parameters was published in 2011 [3]. Sparse-LDA is a very flexible method, since it can handle the $p >> n$ case and multiple classes. It also provides a low dimensional embedding of the data, optimized for class separation, that due to the nature of the l_1-norm provides model-selection on the parameters, that can further give interpretational insights on the data. Given a $n \times p$ data-matrix $\boldsymbol{X}$ and an $n \times K$ indicator matrix $\boldsymbol{Y}$ for class membership, the sparse-LDA algorithm works by solving the *sparse optimal scoring problem*:

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbf{R}^K,\boldsymbol{\beta}\in\mathbf{R}^p}{\arg\min} & \underbrace{\underbrace{\|\boldsymbol{Y\theta}-\boldsymbol{X\beta}\|^2}_{\text{Optimal Scoring}}+\lambda\|\boldsymbol{\beta}\|_1}_{\text{Sparse OS}} \\ \text{s.t. } & \tfrac{1}{n}\boldsymbol{\theta}^T\boldsymbol{Y}^T\boldsymbol{Y\theta}=1,\ \boldsymbol{\theta}^T\boldsymbol{Y}^T\boldsymbol{Y\theta}_\ell=0\ \forall \ell<k, \end{aligned} \tag{1}$$

Here, n is the number of data-samples, p is the number of features and K is the number of different classes. We seek to find the sparse $\boldsymbol{\beta}_i, i \in \{1,...,K-1\}$, discriminant vectors. These vectors are used to project the data into a lower dimensional space, at most $K-1$-dimension, where data-points are classified according to the nearest centroid.

The optimization problem 1 was initially solved in [3] using a block-update scheme for the $(\boldsymbol{\theta}, \boldsymbol{\beta})$ pairs. The $\boldsymbol{\theta}$ scoring vectors can be found with a closed form solution, but the $\boldsymbol{\beta}$ discriminant vectors were found using least angle regression. The optimization has been improved with other methods that are yet to be published.

Here we explore adding *yet another regularization term* that can aid when we have unlabelled samples. Results on a couple of data examples are shown with some empirical results. We also explore the overhead in computation.

Method

The method is an extension of [1]. To leverage unlabelled samples we have to make some assumptions. The main assumption for semi-supervised learning is that:

> *Nearby samples are similar, and more likely to belong to the same class.*

The keypoint for this semi-supervised extension is to add a regularization term that encodes the similarity of the additional samples, thus leveraging the similarity to get a better classification.

Constructing the regularization term

We begin by defining the p-nearest neighbour graph G of the data $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, ..., \boldsymbol{x}_m$, where the first n points have labels, and the rest have none. p is an integer parameter that controls the connectivity of the graph. Points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are said to be neighbours, and connected with an edge in G, if either is one of the p closest points to the other one. We denote $N_p(\boldsymbol{x}_i)$ as the set containing the p closest points to $\boldsymbol{x}_i$ in G. The adjacency matrix A for G can then be defined as:

$$A_{ij} = \begin{cases} 1, & \text{if} \quad \boldsymbol{x}_i \in N_p(\boldsymbol{x}_j) \quad \text{or} \quad \boldsymbol{x}_j \in N_p(\boldsymbol{x}_i) \\ 0, & \text{otherwise} \end{cases}$$

Now we have encoded the similarity of the points. A natural request for the regularization term is to make points that are considered neighbours remain as close as possible after the projection with the discriminant vectors. Thus we define the additional regularization term as:

$$J(\boldsymbol{\beta}) = \sum_{ij}(\boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{x}_j\boldsymbol{\beta})^2 A_{ij}$$

This term can be represented in matrix notation as

$$J(\boldsymbol{\beta}) = 2\boldsymbol{\beta}^T\boldsymbol{X}^T L\boldsymbol{X\beta},$$

where L is the graph *Laplacian*, $L = D - A$, where D is the *degree* matrix. We refer to $J(\boldsymbol{\beta})$ as the *semi-supervised-regularizer*.

This idea for this regularization term comes from spectral and graph-based clustering [5]. The optimization is similar to having the elastic net penalty instead of only the l_1-norm in the optimal scoring formulation. The elastic net peanlty adds a ridge regularization term, but it can also be defined with an arbitrary coefficient matrix. In our case, that matrix is $\boldsymbol{W} = \boldsymbol{X}^T L \boldsymbol{X}$. So given that we can solve the sparse optimal scoring problem with an elastic net regularizer, the only added computation is the creation of the elastic net coefficient matrix $\boldsymbol{W}$.

Added time complexity

The time-complexity of the computation of L is the same as applying k-nearest neighbour to the data set, which is $O(nk + np)$. Afterwards, it is a matter of storing the matrix, which is of size $p \times p$, which can lead to a large memory footprint if p is very large.

Experimental setup

To evaluate the method we examine the performance on two data sets, the mnist data set of handwritten digits [4] and ECG data from the UCR time series classification archive [2].

Mnist

For evaluation we use fixed parameters for the number of neighbours ($k = 5$) and regularization parameter for the l_1-norm ($\lambda = 10^{-3}$), this is to achieve approximately 75 percent zeroes in each discriminant vector. We vary the number of labelled samples from 2 to 50 and the number of unlabelled samples as 10, 100 and 200. We run each experiment 50 times and report mean accuracy on the entire test data set with 30000 examples. The samples used for the training are randomly sampled from the training set. The data is normalized prior to training by subtracting the mean and applying unit scaling. The mnist images are of size 28×28 so $p = 784$. Examples of the mnist digits can been seen in figure 1.

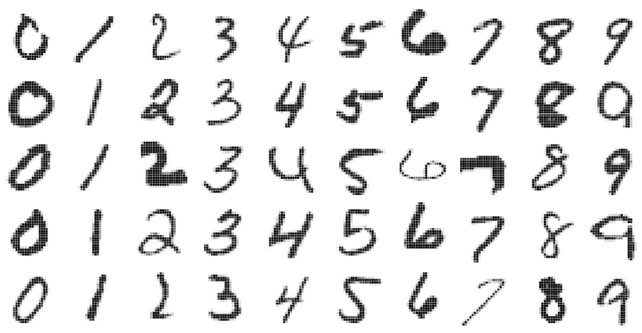

Fig. 1. *Examples of digits from the mnist data set.*

ECG data

The ECG data consists of ECG time series with $p = 136$ measurements per series. There are two classes and in the training data there are $n = 23$ samples. We train the semi-supervised method with 2 to 9 samples for each class and use the rest as unlabelled samples. The evaluation is done on a test set of 861 samples. The l_1-norm regularizer parameter is set to $\lambda = 5 \cdot 10^{-4}$, yielding approximately 100 zeroes in the discriminant vector. The norm parameter for the neighbourhood similarity norm is 0.005 with $k = 3$. An example of a ECG time seried can be seen in figure 2..

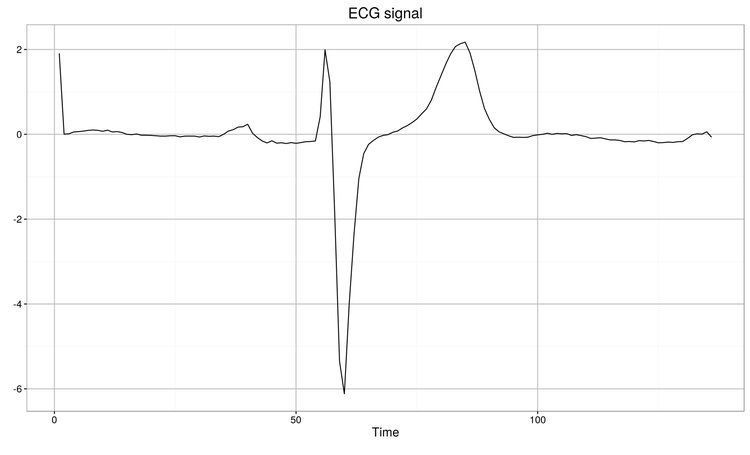

Fig. 2. *Examples of an ECG signal.*

Results

The results are summarized in figures 3.,4. and 5. There is no-crossvalidation done on the other regularization parameters. That is needed to give this subject a full treatment. There are clear differences between the 2-class and 10-class cases. There is a very slight improvemnt for the mnist data-set and the improvement is more dramatic for the ECG-data. We can conclude that it helps to have the unlabelled samples.

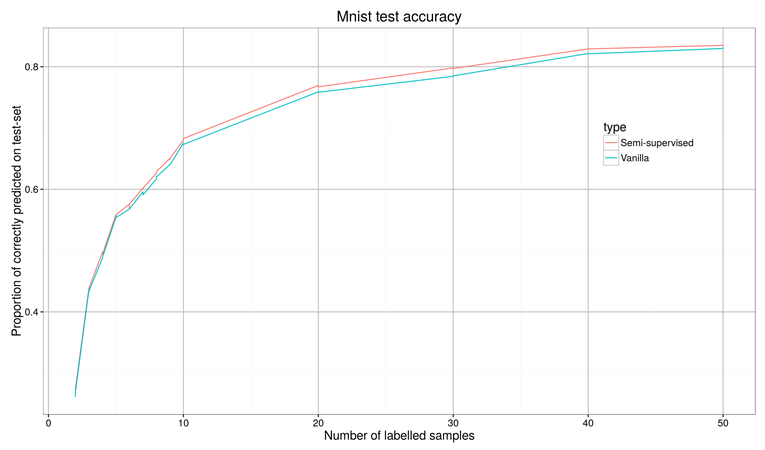

Fig. 3. *Comparison of using the graph similarity regularizer and not on the mnist data set, with varying number of labelled and unlabelled samples.*

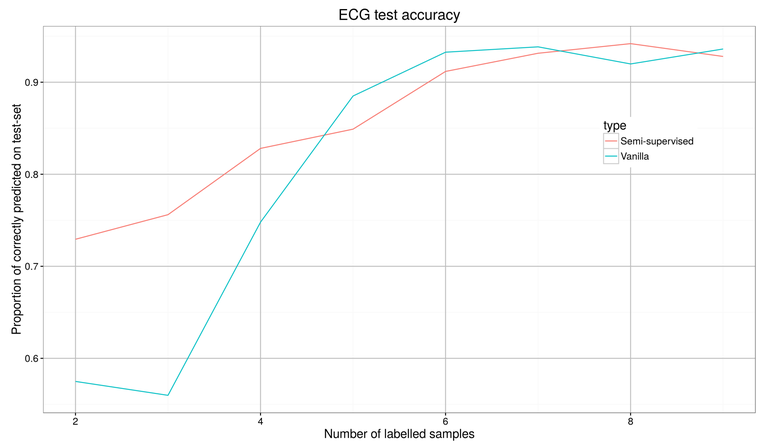

Fig. 4. *Comparison of using the graph similarity regularizer and not on the ECG data set, with varying number of labelled and unlabelled samples.*

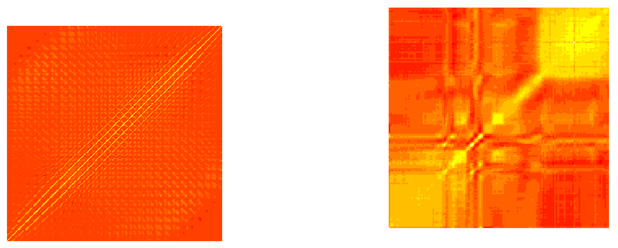

Fig. 5. *The semi-supervised-regularization matrix. On the left we have the matrix for the mnist data-set, which seems to encode spatial similarity. On the right we see the matrix for the ECG data set, which encodes temporal similarity along with features that should be similar further a part in the time-domain.*

References

[1] Deng Cai, Xiaofei He, and Jiawei Han. Semi-supervised discriminant analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE, 2007.

[2] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive. *URL www. cs. ucr. edu/~ eamonn/time_series_data*, 2015.

[3] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 2012.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.