Robust Fisher Linear Discriminant Analysis

Jun Lu
EPFL
jun.lu@epfl.ch

Chen Liu
EPFL
chen.liu@epfl.ch

Abstract

This is the report for the course project in CS-454 Convex Optimization and its Applications. This report will focus on Fisher Linear Discriminant Analysis(Fisher LDA). We will introduce robust Fisher LDA, compare it with and demonstrate its superiority over classic version by both theoretical analysis and experiments. We will show that robust Fisher LDA can be modeled as a convex optimization problem and can be simplified by relevant theorem.

1 Introduction

Fisher Linear Discrimination Analysis(Fisher LDA)[1] is a simple but effective dimension reduction technique to do binary classification. It projects a dataset into a lower-dimensional space and preserve good class-separability. However, Fisher LDA's performance is too sensitive to few outliers. The robust version of Fisher LDA in [2] solves this problems by better estimating the statistical properties of the dataset like mean and covariance. It incorporates a convex optimization problem into Fisher LDA and improves its performance for unbalanced dataset.

In next two chapters, we will introduce Fisher LDA and robust Fisher LDA respectively. Chapter 4 will focus on our tricks to solve the convex optimization problems raised by robust Fisher LDA. Chapter 5 is about dataset and experiments. Finally, we analyze the result and draw a conclusion in the last chapter.

2 Fisher LDA

For a linear discriminant characterized by $w \in R^n$, we seek to obtain a scalar y by projecting the samples onto a line $w^T x$. Of all the possible lines we would like to select the one that maximizes the separability of the scalars. Fisher suggested maximizing the difference between the means of the projected samples, normalized by a measure of the within-class scatter, where the within-class scatter is defined as the sum of the cluster variance or projected samples. Thus the degree of discriminant is measured by the Fisher discriminant ratio:

$$f(w, \Sigma_x, \Sigma_y, \mu_x, \mu_y) = \frac{w^T(\mu_x - \mu_y)(\mu_x - \mu_y)^T w}{w^T(\Sigma_x + \Sigma_y)w} = \frac{(w^T(\mu_x - \mu_y))^2}{w^T(\Sigma_x + \Sigma_y)w} = \frac{w^T S_B w}{w^T S_W w} = J(w) \tag{1}$$

where μ_x and Σ_x denote the mean and covariance of positive instances, μ_y and Σ_y for negative instances, S_B is defined as between-class covariance matrix, S_W is defined as within-class covariance matrix.

Derive and equate this criteria to 0, we get $S_W^{-1} S_B w - J(w)w = 0$. By solving this generalized eigenvalue problem yields $w^* = S_W^{-1}(\mu_x - \mu_y)$ (i.e. project it to the eigenvector with largest eigenvalue). In chapter 5, we will use classic Fisher LDA to project original data into the directions of two eigenvectors with largest eigenvalues and see why we need to project it in such direction.

3 Robust Fisher LDA

In robust Fisher LDA, we take a probabilistic view to the statistical properties like Σ_x, Σ_y, μ_x and μ_y of the dataset. Instead of using 'definite' values, we use a convex set M as the feasible set of these parameters i.e $(\Sigma_x, \Sigma_y, \mu_x, \mu_y) \in M$. The optimization goal is to maximize the worst-case Fisher discriminant radio.

$$\begin{aligned} max\ min_{(\Sigma_x,\Sigma_y,\mu_x,\mu_y)\in M} f(w, \Sigma_x, \Sigma_y, \mu_x, \mu_y) \\ s.t\ w \neq 0 \end{aligned} \tag{2}$$

According to Kim's work in [2], we use $M = \{(\Sigma_x, \Sigma_y, \mu_x, \mu_y)|(\mu_x-\bar{\mu}_x)^T P_x(\mu_x-\bar{\mu}_x) \leq 1, (\mu_y-\bar{\mu}_y)^T P_y(\mu_y-\bar{\mu}_y) \leq 1, \|\Sigma_x - \bar{\Sigma}_x\|_F \leq \delta_x, \|\Sigma_y - \bar{\Sigma}_y\|_F \leq \delta_y\}$ where $\|M\|_F = (\sum_{i,j} M_{ij}^2)^{1/2}$. Parameters in this expression is obtained by resampling[3]. We first create 100 new sets which have the same size as training set and each of them are sampled uniformly from it. For each new set, we calculate its mean and covariance matrix. $\bar{\mu}_x, \bar{\mu}_y, \bar{\Sigma}_x, \bar{\Sigma}_y$ is set as the average value of corresponding mean or variance. P_x, P_y is calculated from covariance of mean in each new set Σ_{μ_x}, Σ_{μ_y} by $P_x = \Sigma_{\mu_x}/n$ and $P_y = \Sigma_{\mu_y}/n$. δ_x and δ_y is defined as the maximum deviation from the average covariance matrix in $\| - \|_F$ metric.

It is obvious that Σ_x, Σ_y and μ_x, μ_y's constraints are independent. So we can simplify equation(2) by $min_{(\Sigma_x,\Sigma_y,\mu_x,\mu_y)\in M}\frac{[w^T(\mu_x-\mu_y)]^2}{w^T(\Sigma_x+\Sigma_y)w} = \frac{min_{(\mu_x,\mu_y)\in M}[w^T(\mu_x-\mu_y)]^2}{max_{(\Sigma_x,\Sigma_y)\in M}w^T(\Sigma_x+\Sigma_y)w} = \frac{min_{(\mu_x,\mu_y)\in M}[w^T(\mu_x-\mu_y)]^2}{w^T(\Sigma_x+\Sigma_y+(\delta_x+\delta_y)I)w}$.

Lemma 1 Let $R(w, a, B) = \frac{(w^T a)^2}{w^T B w}$ and (a^*, B^*) be the optimal of following problem:

$$\begin{aligned} min\ a^T B^{-1} a \\ s.t\ (a, B) \in V \end{aligned} \tag{3}$$

Let $w^* = B^{*-1}a^*$, then (w^*, a^*, b^*) satisfy the minmax property

$$R(w^*, a^*, B^*) = a^{*T} B^{*-1} a^* = max_{w\neq 0} min_{(a,B)\in V} R(w, a, B) = min_{(a,B)\in V} max_{w\neq 0} R(w, a, b) \tag{4}$$

and the saddle point property

$$R(w, a^*, B^*) \leq R(w^*, a^*, B^*) \leq R(w^*, a, B), \forall w \in \mathbb{R}^n/\{0\}, \forall (a, B) \in V \tag{5}$$

The proof of lemma 1 is in Kim's paper[2]. Let $B = \bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I$ and $a = \mu_x - \mu_y$, we can solve the max-min optimization problem (2) using this lemma. The optimal value w^* is

$$w^* = (\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)^{-1}(\mu_x^* - \mu_y^*) \tag{6}$$

where μ_x^* and μ_y^* is the optimal value of the following optimization problem.

$$\begin{aligned} min\ (\mu_x - \mu_y)^T(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)^{-1}(\mu_x - \mu_y) \\ s.t\ (\mu_x - \bar{\mu}_x)^T P_x(\mu_x - \bar{\mu}_x) \leq 1,\ (\mu_y - \bar{\mu}_y)^T P_y(\mu_y - \bar{\mu}_y) \leq 1 \end{aligned} \tag{7}$$

Since covariance matrix is always semi-positive definite, $\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I$, P_x and P_y are all semi-positive definite. So (6) is a convex optimization problem and more specifically quadratically constrained quadratic programming (QCQP) problem. Such problem can be solved in $O(n^3)$ flops.[4]

4 Optimization Method

To solve the convex optimization problem(7), we can use general interior point method(IPM)[4]. There is a python library called cvxopt that we can rely on to implement IPM. However, IPM is too time and computing resource consuming to solve this problem. In this chapter, we design a parametric method which is not included in paper[2] but can estimate the optimal of problem(7) quickly and effectively.

We first write the Lagrangian of problem(7): $L(\mu_x, \mu_y, \lambda, \gamma) = (\mu_x - \mu_y)^T(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)^{-1}(\mu_x - \mu_y) + \lambda((\mu_x - \bar{\mu}_x)^T P_x(\mu_x - \bar{\mu}_x) - 1) + \gamma((\mu_y - \bar{\mu}_y)^T P_y(\mu_y - \bar{\mu}_y) - 1)$. Since problem(7) is convex, so the duality gap is zero and the optimal values should satisfy KarushKuhn-Tucker(KKT) conditions. We list some constraints of KKT condition below.

$$\lambda((\mu_x - \bar{\mu}_x)P_x(\mu_x - \bar{\mu}_x) - 1) = 0 \tag{8}$$

$$\gamma((\mu_y - \bar{\mu}_y)P_y(\mu_y - \bar{\mu}_y) - 1) = 0 \tag{9}$$

$$\frac{\partial L}{\partial \mu_x} = (\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)^{-1}(\mu_x - \mu_y) + \lambda P_x(\mu_x - \bar{\mu}_x) = 0 \tag{10}$$

$$\frac{\partial L}{\partial \mu_y} = (\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)^{-1}(\mu_y - \mu_x) + \gamma P_y(\mu_y - \bar{\mu}_y) = 0 \tag{11}$$

If $\lambda = 0$, according to equation(10), we have $(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)^{-1}(\mu_x - \mu_y) = 0$. Multiply both sides by $(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)$, we have $\mu_x - \mu_y = 0$ and this is contradict with the assumption that $\mu_x \neq \mu_y$. As a result, λ has to be non-zero. That is to say, the corresponding constraint is tight and according to equation (8), we have the following equations.(same for γ)

$$(\mu_x - \bar{\mu}_x)^T P_x(\mu_x - \bar{\mu}_x) = 1,\ (\mu_y - \bar{\mu}_y)^T P_y(\mu_y - \bar{\mu}_y) = 1 \tag{12}$$

With equality constraint above, we can parameterize μ_x and μ_y.

$$\mu_x(k_x) = \bar{\mu}_x + \frac{k_x}{\sqrt{k_x^T P_x k_x}},\ \mu_y(k_y) = \bar{\mu}_y + \frac{k_y}{\sqrt{k_y^T P_y k_y}},\ k_x, k_y \in \mathbb{R}^n \tag{13}$$

Under such parametric setting, we can see that $\mu_x(k_x)$ and $\mu_y(k_y)$ always satisfy equation(12). Combine equation(13)'s substitution and optimization object of problem(7) and we can obtain a unconstrained optimization problem.

$$min_{k_x,k_y \in \mathbb{R}^n} H(k_x, k_y) = (\mu_x(k_x) - \mu_y(k_y))^T(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)(\mu_x(k_x) - \mu_y(k_y)) \tag{14}$$

$H(k_x, k_y)$ is not convex with respect to k_x and k_y, but we can still use stochastic gradient descent(SGD)[5] to get a good estimate of the optimal value. The gradient of $H(k_x, k_y)$ is shown below. SGD process stops when the gradient's norm is lower than a threshold(usually 10^{-5}).

$$\frac{\partial H}{\partial k_x} = \frac{(k_x^T P_x k_x)I - P_x k_x k_x^T}{(k_x^T P_x k_x)^{3/2}}(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)(\mu_x(k_x) - \mu_y(k_y)) \tag{15}$$

$$\frac{\partial H}{\partial k_y} = \frac{(k_y^T P_y k_y)I - P_y k_y k_y^T}{(k_y^T P_y k_y)^{3/2}}(\bar{\Sigma}_x + \bar{\Sigma}_y + (\delta_x + \delta_y)I)(\mu_y(k_y) - \mu_x(k_x)) \tag{16}$$

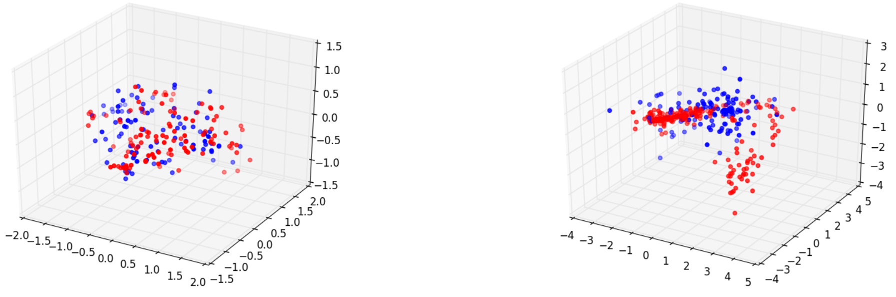
5 Experiment

5.1 Dataset Description

To demonstrate and compare the performance of fisher LDA and robust fisher LDA, we used the sonar and ionosphere benchmark problems from the UCI repository (LINK). The two benchmark problems have 208 and 351 points, and the dimension of each data point is 60 and 34, respectively.

For the ionosphere dataset, there are 225 instances labeled class 1 (positive), 126 instances labeled class 2 (negative); for the sonar dataset, there are 111 entries labeled class 1 (positive), 97 entries labeled class 2 (negative). Both dataset is unbalanced, especially for ionosphere dataset.

To have a more intuitive view of both datasets, we use principle component analysis(PCA)[6] to project the whole dataset into 3-d space. We plot the spatial distribution of whole dataset, with positive and negative points marked as blue and red respectively shown in Fig 1(a) and 1(b). We can find the ionosphere dataset is more clustered than sonar. Later we will show robust Fisher LDA on the ionosphere has higher test accuracy than sonar.



(a) sonar

(b) ionosphere

Figure 1: PCA on two dataset into 3-d dimensions.

5.2 Solving Unbalanced Data Set

Concerning a linear discriminant problem, an unbalanced data set leads to biased decisions towards the majority class and therefore an increase in the generalization error. As referred in the section of Data Description, the number of samples per class in our problem is not equally distributed. To address this problem, we considered the following approaches.

Random Under Sampling - Randomly select a subset the majority classes' data points so that the number of data points of each class is equal to the number of points of the minority class - Class 1.

Random Over Sampling - Randomly add redundancy to the data set by duplicating data points of the minority classes so that the number of data points of each class is equal to the number of points of the majority class - Class 2.

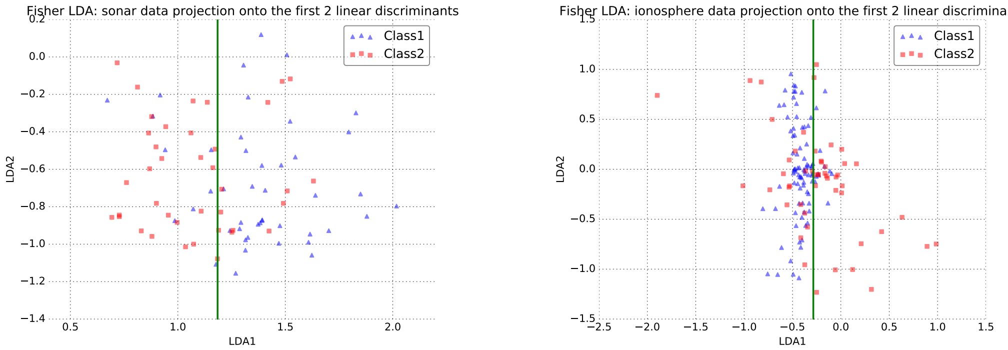
If we compare these two methods, the *Random Under Sampling* would not be useful in our problem, because the dataset is small and *Random Under Sampling* will lose many information. As a result, We use *Random Over Sampling* to deal with the unbalanced dataset to make the two classes equally distributed.

5.3 Fisher LDA Numerical Results

We sampled randomly from the dataset sonar and ionosphere separately with 60% as the training set, and 40% as the test set. After getting related two eigenvectors from the training set, we project the test set into these two directions. In Fig 2(a) and 2(b), the LDA1 is the direction with largest eigenvalue, the LDA2 is the direction with the second largest eigenvalue. For the sonar dataset, if we project it to LDA1, the accuracy is 75.00%; if we project it to LDA2, the accuracy is 53.57%. For the ionosphere dataset, if we project it to LDA1, the accuracy is 78.01%; if we project it to LDA2, the accuracy is 43.26%. The vertical line in the figure is the threshold used for the direction of LDA1. From the simple test, we show the rightness of Fisher linear discriminant criterion.

5.4 Fisher LDA and Robust Fisher LDA Comparison

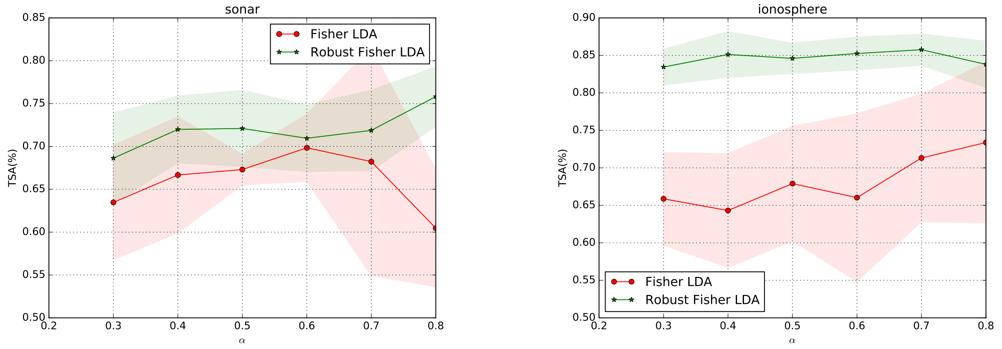
To compare classic Fisher LDA and robust Fisher LDA, We test with an increasing number of training set. Fig 3(a) 3(b) show the comparison of Test-set accuracy(TSA) between Fisher LDA and Robust Fisher LDA on the two datasets. As described in section Dataset Description, we illustrate that the Robust Fisher LDA on dataset ionosphere would have higher accuracy. Which is very clear shown in following figures.



(a) sonar.

(b) ionosphere.

Figure 2: Project sample datas into two directions.



(a) sonar

(b) ionosphere

Figure 3: Test-set accuracy(TSA) for sonar and ionosphere for different training set size. Shadow area is the standard deviation.

6 Conclusion

In this project, we look into the Fisher Linear Discrimination Analysis(Fisher LDA) and use convex optimization techniques to derive and test its robust version. We have tested both models on two datasets and prove that for noisy data, robust Fisher LDA outperforms classic Fisher LDA.

7 Reference

[1]. Fukunaga, Keinosuke. *Introduction to statistical pattern recognition.* Academic press, 2013.

[2]. MLA Kim, Seung-Jean, Alessandro Magnani, and Stephen Boyd. *Robust fisher discriminant analysis.* Advances in Neural Information Processing Systems 18 (2006): 659.

[3]. B.Efron and R.Tibshirani. *An Introduction to Bootstrap.* Chapman and Hall, London UK,1993.

[4]. S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[5]. Bottou, Lon. *Large-scale machine learning with stochastic gradient descent.* Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010. 177-186.

[6]. Jolliffe, Ian. *Principal component analysis.* John Wiley & Sons, Ltd, 2002. APA