Discriminant Analysis: A Unified Approach

Peng Zhang & Jing Peng
Tulane University
Electrical Engineering & Computer Science Department
New Orleans, LA 70118
{zhangp,jp}@eecs.tulane.edu

Norbert Riedel
Tulane University
Mathematics Department
New Orleans, LA 70118
riedel@math.tulane.edu

Abstract

Linear discriminant analysis (LDA) as a dimension reduction method is widely used in data mining and machine learning. It however suffers from the small sample size (SSS) problem when data dimensionality is greater than the sample size. Many modified methods have been proposed to address some aspect of this difficulty from a particular viewpoint. A comprehensive framework that provides a complete solution to the SSS problem is still missing. In this paper, we provide a unified approach to LDA, and investigate the SSS problem in the framework of statistical learning theory. In such a unified approach, our analysis results in a deeper understanding of LDA. We demonstrate that LDA (and its nonlinear extension) belongs to the same framework where powerful classifiers such as support vector machines (SVMs) are formulated. In addition, this approach allows us to establish an error bound for LDA. Finally our experiments validate our theoretical analysis results.

1 Introduction

The purpose of this paper is to present a unified theoretical framework for discriminant analysis. Discriminant analysis [6, 10] has been successfully used as a dimensionality reduction technique for many classification problems. According to Fisher’s criterion, one has to find a projection matrix \( W \) that maximizes:

\[
J(W) = \frac{|W^T S_b W|}{|W^T S_w W|}. \tag{1}
\]

where \( S_b \) and \( S_w \) are so-called between-class and within-class matrices, respectively. In practice, the “small sample size” (SSS) problem is often encountered, where \( S_w \) is singular. Therefore, the maximization problem can be difficult to solve.

To address this issue, the term \( \varepsilon I \) is added, where \( \varepsilon \) is a small positive number and \( I \) the identity matrix of proper size. This results in maximizing

\[
J(W) = \frac{|W^T S_b W|}{|W^T (S_w + \varepsilon I) W|}. \tag{2}
\]

It can then be solved without any numerical problems.

In [9], Friedman discusses regularized discriminant analysis with regard to the small sample size problem. Equation (2) is a special case of his regularized discriminant analysis in practice. However, Friedman leaves open the question of whether an optimal choice for the parameter \( \varepsilon \) exists, and if so, whether this optimal choice is a unique one? Further, Friedman’s analysis is based on traditional parametric methods in statistics, and Gaussian distributions for underlying random variables are often assumed. That is, his regularization technique is one of improving the estimation of the covariance matrices that are possibly biased due to the small data size. The validity of this technique can be very limited. For example, the nonlinear case of discriminant analysis, called generalized discriminant analysis (GDA), becomes very popular in practice and shows promising potential in many applications. The idea is to map the data into a kernel induced feature space and then perform discriminant analysis there. However, in a high and possible infinite dimensional feature space, GDA is not well justified from Friedman’s point of view. Thus, a rigorous justification for the regularization term \( \varepsilon I \) is still missing, especially when GDA is considered.

In [7], Evgeniou, Pontil and Poggio present a theoretical framework for the regularization functional (the so called regularization network)

\[
\min_f H[f] = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \lambda ||f||_K^2, \tag{3}
\]

where \( ||f||_K^2 \) is a norm in a Reproducing Kernel Hilbert Space \( H \) defined by the positive definite function \( K \), \( l \) is the number of data points or examples and \( \lambda \) is a regularization parameter. In their paper they justify the above functional
using statistical learning theory that is mostly developed by Vapnik [22].

In this paper, we will show that discriminant analysis can be cast in the framework of statistical learning theory. More specifically, Fisher’s criterion (1) is equivalent to the regularization network (3) without regularization, which is,

\[ \min_{\lambda} \sum_{i=1}^{l} (y_i - f(x_i))^2. \]

And Equation (2) is equivalent to the regularization network (3) with the parameter \( \epsilon \) corresponding to the regularization term \( \lambda \). Thus the role of \( \epsilon \) is well justified by statistical learning theory. Further we show that the optimal choice of \( \lambda \) exists, and that it is unique. Additionally we establish an error bound for the function that discriminant analysis tries to find.

1.1 Related Work

The relationship between least squares regression and Fisher’s linear discriminant has been well known for a long time. There is a good review in [6]. Actually in the seminal paper [8], Fisher already pointed out its connection to the regression solution. In [17], Mika also provides an extensive review of LDA. Another widely used technique in statistics, canonical correlation analysis, is also closely related to least squares regression [1].

The regularized least squares (RLS) methods have been studied for a long time, under different names. In statistics, ridge regression [12] has been very popular for solving badly conditioned linear regression problems. After Tikhonov published his book [21], it was realized that ridge regression uses the regularization term in Tikhonov’s sense. In the 1980’s, weight decay was proposed to help prune unimportant neural network connections, and was soon recognized [13] that weight decay is equivalent to ridge regression. Recently this old method was found essential in the framework of statistical learning theory, labeled by different names (RLS [18], Regularized Network [7], least squares SVMs [19], and proximal SVMs [11]). It is demonstrated that this regression method is fully comparable to SVMs when used as a classifier [18, 24]. Most recently, an error bound for RLS given a finite sample data set was developed in [5].

Many elements discussed in this paper are scattered throughout the literature and may look familiar to some audience but not to others. The fact is that most are not comprehensive enough for this topic and unclear issues pop up for many. For example, many researchers such as Mika [17] apply the regularization technique to address the SSS problem but they still follow Friedman’s principle that is quite limited. It is our goal to adopt a simple yet unified approach to making this topic more comprehensive.

2 Discriminant Learning Analysis

In this section, we first review LDA using Fisher’s criterion, and then go on to study it from a learning theory point of view, which we call discriminant learning analysis (DLA).

2.1 Linear Discriminant Analysis

In LDA, within-class, between-class, and mixture scatter matrices are used to formulate the criteria of class separability. Consider a \( J \)-class problem, where \( m_0 \) is the mean vector of all data, and \( m_j \) is the mean vector of \( j \)-th class data. A within-class scatter matrix characterizes the scatter of samples around their respective class mean vectors, and it is expressed by \( S_w = \sum_{j}^{J} \sum_{l}^{l_j} (x_l^j - m_j)(x_l^j - m_j)^T \), where \( l_j \) is the size of the data in the \( j \)-th class. A between-class scatter matrix characterizes the scatter of the class means around the mixture mean \( m_0 \). It is expressed by \( S_b = \sum_{j=1}^{J} l_j (m_j - m_0)(m_j - m_0)^T \). The mixture scatter matrix is the covariance matrix of all samples, regardless of their class assignments, and it is given by \( S_m = \sum_{j=1}^{J} \sum_{l=1}^{l_j} (x_l^j - m_0)(x_l^j - m_0)^T = S_w + S_b \). Fisher’s criterion is used to find the projection matrix that maximizes (1). In order to determine the matrix \( W \) that maximizes \( J(W) \), one can solve the generalized eigenvalue problem:

\[ S_w w = \lambda_i S_b w_i. \]

The eigenvectors corresponding to the largest eigenvalues form the columns of \( W \). For a two class problem, it can be written in a simpler form:

\[ S_w w = m_1 - m_2, \]

where \( m_1 \) and \( m_2 \) are the means of the two classes.

2.2 LDA as Least Squares Function Estimation

LDA is a supervised subspace learning problem where we are given a set \( z = \{ (x_l, y_l) \}_{l=1}^l \) of \( l \) training examples drawn i.i.d. from the probability space \( Z = X \times Y \). Here probability measure \( \rho \) is defined, and \( x_i \) are the \( n \)-dimensional inputs. We first consider two class problems where we have \( y_l \in \{-1, 1\} \) as class labels. Without loss of generality, we assume \( x \) has zero mean, i.e., \( E(x) = 0 \). Let us consider:

\[ f_{MSE} = \arg \min_{f} \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2. \]

This is a mean squared function estimation problem. Here we first consider the hypothesis space \( H_L \), the function class of linear projections. That is, \( H_L = \{ f | f(x) = w^T x, x \in X \} \). Thus we will solve

\[ w_{opt} = \arg \min_{w} \frac{1}{l} \sum_{i=1}^{l} (y_i - w^T x_i)^2. \]
It turns out that the solution of (6) is the same as the solution obtained by Fisher’s criterion.

**Lemma 1** The linear system derived by the least squares criterion (6) is equivalent to the one derived by Fisher’s criterion (1), up to a constant, in two class problems.

Proof: We rewrite (6) as:

\[
\sum_{i=1}^{l} (y_i - w^T x_i)^2 = (y - X^T w)^T (y - X^T w) = (y^T - w^T X)(y - X^T w) = l - 2(l_1 m_1 - l_2 m_2)^T w + w^T S_m w.
\]

We use the fact that \( X_y = l_1 m_1 - l_2 m_2 \). Taking the derivative with respect to \( w \) and setting the result to 0 we obtain

\[
S_m w = (l_1 m_1 - l_2 m_2).
\]

(7)

This is equivalent to Equation (4). When \( S_w \) is full rank, three equations \( S_m w = (l_1 m_1 - l_2 m_2) \), \( S_m w = m_1 - m_2 \), and \( S_w w = m_1 - m_2 \) have the same solution \( w \) up to a constant factor, given that the overall mean is 0. For details see Appendix A.

### 2.3 Interpretation of LDA within Learning Theory

When both classes are gaussian distributed, the LDA can be shown optimal. However more often in reality they are not, then it is not clear how good the LDA is. The conventional view of LDA is limited here. The learning framework, on the other hand, is more general. There is no need to assume the distribution of the data to discuss the ‘goodness’. Instead, the learning approach directly aims to find a good function mapping and the goodness is well defined (see Appendix C and ??).

In a general sense, LDA tries to find a transformation \( f \) such that the transformed data have a large class mean difference and small within-class scatters. The equivalence between (6) and (1) can be understood in another way: one minimizes the mean squared error with respect to each class mean while keeping the mean difference constant. We define \( f_o \) as the best function that minimizes the mean squared error. That is

\[
f_o = \arg\min f \int_Z (y - f(x))^2 dp.
\]

(8)

Given a set of data \( z \), one tries to learn the function \( f_z \) that is as close as possible to the optimal mapping \( f_o \). That is, \( f_{opt} = \arg\min_{f_z} \int_Z (f_z - f_o)^2 dp \). However, the probability measure \( p \) is unknown and so is \( f_o \). Instead, we may consider (5) that is called the empirical error (risk) minimization. Because (6) is equivalent to Fisher’s criterion, it is clear that Fisher’s criterion leads to an empirical error minimization problem. From the statistical learning theory point of view, however, without controlling the function norm, solving equation (5) often leads to overfitting data and the solution is not stable. And when the sample size is smaller than the dimensionality, it is an ill-posed problem and the solution is not unique. That is where diverse LDA techniques have been developed to solve the SSS problem.

### 2.4 Discriminant Learning Analysis as Regularized Least Squares Regression

Following [7], we instead minimize over a hypothesis space \( H \) the following regularized functional for a fixed positive parameter \( \lambda \):

\[
f_{DLA} = \arg\min_{f \in H} \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \lambda \|f\|_H^2.
\]

(9)

Again, we consider the linear projection function space \( H_L \). In this case, \( \|f\|^2 = w^T w \).

\[
w_{opt} = \arg\min_{w \in H} \frac{1}{l} \sum_{i=1}^{l} (y_i - w^T x_i)^2 + \lambda w^T w.
\]

(10)

Thus, \( \sum_{i=1}^{l} (y_i - w^T x_i)^2 + \lambda w^T w = l - 2(l_1 m_1 - l_2 m_2)^T w + w^T (S_m + \lambda I) w \). Taking the derivative with respect to \( w \) and setting the result to 0, we have

\[
(S_m + \lambda I) w = (l_1 m_1 - l_2 m_2).
\]

(11)

In Appendix A, it is shown that (11) is equivalent to

\[
(S_w + \lambda I) w = m,
\]

(12)

and therefore solving it is equivalent to maximizing (2). We call this approach discriminant learning analysis. Since for any given positive regularization term \( \lambda \), (9) always has a unique solution, we have

**Proposition 2** Discriminant learning analysis is a well-posed problem.

### 2.5 Nonlinear DLA

In the new framework of discriminant learning analysis, the nonlinear case is not an “extension” of the linear case by the “kernel” trick. In the above analysis we have adopted the linear hypothesis space \( H_L \), whereby we have been looking for an optimum solution in that space. If we choose a general Reproducing Kernel Hilbert Space (RKHS) as our hypothesis space, we will obtain nonlinear function mappings as the subspace. The derivation of DLA in a general function space is simply carried out in terms of function
analysis, in a similar way as in linear space. Here we skip the details of the derivations. Briefly speaking, for a two class problem, the nonlinear DLA solution is found by applying PCA to remove the null space of $W$. If $S_w$ is full rank then solve regular LDA; else in the null space of $S_w$, find the eigenvectors of $S_b$ with largest eigenvalues.

**DMLDA:** Apply PCA to remove the null space of $S_b$ first, then find the eigenvectors of $S_w$ corresponding to the smallest eigenvalues.

1 Some researchers also maximize $J(W) = \frac{W^T S_w W}{W^T S_w W}$, it is equivalent to maximizing $J(W) = \frac{W^T S_w W}{W^T S_w W}$.
It should be noted that PCA+LDA and Scatter-LDA can be equivalent when $S_w$ and $S_m$ span the same subspace. However, they are different when $S_b$ totally or partially spans the null space of $S_w$, thus $S_w$ and $S_m$ span different subspaces. For facial images the latter case turns out to be more common. In [3], Chen et al. prove that the null space of $S_w$ contains discriminant information. They also show that Scatter-LDA is not ‘optimal’ in that it fails to distinguish the most discriminant information in the null space of $S_w$. Thus they proposed the newLDA method. However, newLDA fell short of making use of any information outside of that null space. DLDA, on the other hand, discards the null space of $S_b$. This is very problematic. First, it fails to distinguish the most discriminant information in the null space of $S_w$, as in Scatter-LDA. Second, the null space of $S_b$ is not completely useless. This can be seen in a two class case. DLDA simply chooses $m$ as the subspace. However this can be biased because it ignores $S_w$.

It should be noted that in [16] there is a misleading claim, where the authors mention that there is no significant difference between newLDA and DLDA. This incorrect claim is also pointed by others [14]. In a word, all these techniques make “hard” decisions, either discarding a null space, or only working in a null space. In the following, we can see that DLA always works in the whole space, not making any “hard” decisions. Instead, it ‘fuses’ information from both subspaces. DLDA can work in the whole space because it is well-posed.

We first consider how LDA is solved. For the moment let us assume that $S_w$ is full rank, and thus $S_w^{-1}$ exists. For symmetric matrix $S_w$ there is a matrix $U$ such that $U^T U = I$, $S_w = U S U^T$, and $S_w' = U S^{-1} U^T$, where $S$ is a diagonal matrix. $U^T$ is a transform matrix. Consider the following in a two class LDA

$$w = S_w^{-1} m = U S^{-1} U^T m.$$  

The between vector $m$ is transformed to a new basis by $U^T$, then multiplied by a constant $\frac{1}{\lambda_l}$ along each dimension, and finally transformed again by $U^T$ back to the origin basis.

To illustrate this process. We consider a simple case where the data are in two dimensions. $x_1$ and $x_2$ are the original axis bases, as shown in Figure 1. $x_1'$ and $x_2'$ are the orthogonal axis bases in which $S_w$ is diagonalized. We choose $S_w$ such that $U^T S_w U = \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}$. Thus $\Sigma_{11} = \frac{1}{2}$ and $\Sigma_{22} = 2$. Let $\begin{bmatrix} m_1' \\ m_2' \end{bmatrix} = U^T m$. Then $m_1'$ and $m_2'$ are the projection on $x_1'$ and $x_2'$, respectively. By multiplying $U^{-1}$, $m_1' = \frac{1}{\sqrt{2}} m_1$ and $m_2' = 2 m_2$. The LDA solution $S_w^{-1} m$ is shown in Figure 1-(a) and denoted by “LDA”.

Now consider DLA: $(S_w + \lambda I)w = m$. Notice that if matrix $U$ can diagonalize $S_w$ then it can diagonalize $(S_w + \lambda I)$ at the same time: $U^T (S_w + \lambda I) U = U^T S_w U + \lambda U^T U = \Sigma + \lambda I$. Thus, $(S_w + \lambda I)^{-1} = U (\Sigma + \lambda I)^{-1} U^T$. To illustrate the effect of the regularization term, we choose $\lambda = 1$ and in Fig. 1, $\Lambda_1$ are the eigenvalues of $S_w + \lambda I$. Now $\Lambda_1^{-1} = \frac{1}{\frac{1}{2}}$, and $\Lambda_2^{-1} = \frac{2}{3}$. The vector $(S_w + \lambda I)^{-1} m$ is shown in Fig. 1-(a) and denoted by “DLA”.

We now consider the case where the SSS problem occurs. We keep $\Lambda_2 = 1.5$ the same but $\Lambda_1 = 1$ (that is, the original eigenvalue of $S_w$ is 0 along this dimension). This case is plotted in Fig. 1-(b). DLA takes advantage of information from both worlds. However, PCA+LDA will throw away the dimension $x_1$ and only keep the direction $x_2$. Thus it ignores critical discriminant information. newLDA, on the other hand, will only keep the dimension $x_1$ and throw away any discriminant information along $x_2$. DLA simply chooses $m$.

Thus, when the SSS problem does not occur, DLA has a smoothing effect on the eigenvalues of $S_w$. If the SSS problem does occur, DLA fuses and keeps a balance between the information both in and out of the null space of $S_w$. This balance is achieved by choosing proper $\lambda$.

5 Experiments

5.1 Feret Facial Images

We tested DLA algorithm against PCA+LDA, newLDA, DLDA and Scatter-LDA on FERET facial image data (http://www.itl.nist.gov/iad/humanid/feret/). We extracted 400 images for the experiment, where there are 50 individuals with 8 images from each. The images have been preprocessed and normalized using standard methods from the FERET database. The size of each image is 150 x 130 pixels, with 256 grey levels per pixel.
We randomly choose five images per person for training, and the remaining three for testing. Thus the training set has 250 images while the testing set has 150. Subspaces are calculated from the training data, and the 1-NN classifier (3-NN failed for all the methods due to five images per person in the training data) is used to obtain the accuracy rates after projecting the data onto the subspace. To obtain average performance, each method is repeated 10 times. Thus there are total 1,500 testing images for each method. The regularization term \( \lambda l \) is chosen by 10-fold cross-validation. The average accuracy rates are shown in Figure 2. The X-axis represents the dimensionality of the subspace. For each technique, the higher the dimension, the less discriminant the dimension. For most techniques, the accuracy rates increase quickly in the first 15 dimensions, and then increase slowly with additional dimensions.

DLA is uniformly better than any other algorithms, demonstrating its efficacy. It achieves the highest accuracy rate of 0.977. newLDA performs quite well in these experiments, again demonstrating that the most discriminant information is in the null space of \( S_w \), for the facial recognition task. On the other hand, Scatter-LDA does not perform well at lower dimensional subspaces. But it eventually performs better than PCA+LDA, when dimensions are more than 42. All methods achieve their highest accuracy rate with a 49 dimensional subspace, which is not surprising, for this is a 50-class problem. It is noticed that the performance of newLDA and Scatter-LDA (its tail is not shown in the plot) drops quickly with unnecessary dimensions.

### Table 1. Classification error rates in subspaces computed by GDA and NDLA, using 3-NN classifier, on 11 UCI data sets.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>GDA</th>
<th>NDLA</th>
<th>Better?</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast cancer</td>
<td>0.449</td>
<td>0.30</td>
<td>✓</td>
</tr>
<tr>
<td>cancer Wisconsin</td>
<td>0.033</td>
<td>0.038</td>
<td>×</td>
</tr>
<tr>
<td>credit</td>
<td>0.20</td>
<td>0.174</td>
<td>✓</td>
</tr>
<tr>
<td>heart Cleve</td>
<td>0.214</td>
<td>0.185</td>
<td>✓</td>
</tr>
<tr>
<td>heart Hungary</td>
<td>0.236</td>
<td>0.216</td>
<td>✓</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.088</td>
<td>0.077</td>
<td>✓</td>
</tr>
<tr>
<td>new thyroid</td>
<td>0.030</td>
<td>0.029</td>
<td>✓</td>
</tr>
<tr>
<td>pima</td>
<td>0.310</td>
<td>0.294</td>
<td>✓</td>
</tr>
<tr>
<td>glass</td>
<td>0.088</td>
<td>0.068</td>
<td>✓</td>
</tr>
<tr>
<td>sonar</td>
<td>0.162</td>
<td>0.170</td>
<td>×</td>
</tr>
<tr>
<td>iris</td>
<td>0.073</td>
<td>0.065</td>
<td>✓</td>
</tr>
<tr>
<td>average</td>
<td>0.171</td>
<td>0.147</td>
<td>✓</td>
</tr>
</tbody>
</table>

5.2 UCI Data Sets

In these experiments, we compare nonlinear LDA (GDA) and nonlinear DLA (NDLA) in two class classification problems. We did not test other nonlinear techniques such as KDDA [16] due to their complex implementation. The implementation of kernel DLDA is very involved. We use 11 data sets from the UCI machine learning database. They are all two classification problems. For each data set, we randomly choose 60% as training and the remaining 40% as testing. We train GDA and NDLA on the training data and obtain projections. We then project both training and test data on the chosen subspace and use the 3-NN classifier to obtain error rates. Note that for the two class case, one dimensional subspace is sufficient. We use the Gaussian kernel \( k(x_1, x_2) = \exp\left(-\|x_1-x_2\|^2/\sigma^2\right) \) to induce the Reproducing Kernel Hilbert Space for both GDA and NDLA. The kernel parameter \( \sigma \) and regularization term \( \lambda \) were determined through 10-fold cross-validation. We repeat the experiments 10 times on each data set to obtain the average error rates.

The results are shown in Table 1. On 9 data sets out of 11, NDLA performs better than GDA. Especially on the breast cancer credit, heart cleve, heart Hungary and glass data, it is significantly better than GDA, with 95% confidence.

6 Summary

This paper presents a learning approach, Discriminant Learning Analysis for effective dimension reduction. While discriminant analysis is formulated as a least square approx-
Appendix A

In this appendix, we say $Aw = c_1 v$ and $Aw = c_2 m$ are equivalent in the sense that the solution $w$ in the same direction, where $A$ is a matrix, $c_1, c_2$ are scalars and $m$ is a vector. First we show that $S_m w = m$ and $S_w w = m$ have the same solution (same set of eigenvectors), where $m = m_1 - m_2$ is the mean difference of the two class problem.

We know that solving $S_w w = m$ is equivalent to solving

$$S_w^{-1} S_w \Phi = \Phi \Lambda$$

(17) where $\Phi$ and $\Lambda$ are the eigenvector and eigenvalue matrices of $S_w^{-1} S_w$. Since we have $S_w = S_m = S_b$, following [10] (pp. 454), (17) can be converted to

$$(S_m - S_b) \Phi \Lambda = S_b \Phi$$

$$S_m \Phi \Lambda = S_b \Phi (I + \Lambda)$$

$$S_m^{-1} S_b \Phi = \Phi \Lambda (I + \Lambda)^{-1}.$$  

(18)

This shows that $\Phi$ is also the eigenvector matrix of $S_m^{-1} S_b$, and its eigenvalue matrix is $\Lambda (I + \Lambda)^{-1}$. When the components of $\Lambda$ and $\alpha_1$ are arranged from the smallest to the largest as $\alpha_1 \geq ... \geq \alpha_n$, the corresponding components of $\Lambda (I + \Lambda)^{-1}$ are also arranged as $\frac{\alpha_1}{1 + \alpha_1} \geq ... \geq \frac{\alpha_n}{1 + \alpha_n}$.

That is, the $t$ eigenvectors of $S_m^{-1} S_b$ corresponding to the $t$ largest eigenvalues are the same as the $t$ eigenvectors of $S_w^{-1} S_b$ corresponding to the $t$ largest eigenvalues. As a special case (two class problems), $S_m w = m$ and $S_w w = m$ share the same “eigenvector” solution.

Now we show that $S_m w = (l_1 m_1 - l_2 m_2)$ and $S_m w = m_1 - m_2$ share the same solution also. Consider that the overall mean $m_0 = 0$. $l m_0 = l_1 m_1 + l_2 m_2 = 0$, we have $m_1 = \frac{1}{l} m_2 = -\frac{1}{l} m$, and $l_1 m_1 - l_2 m_2 = 2m_1 m_2$.

Thus $S_m w = (l_1 m_1 - l_2 m_2)$ becomes $S_m w = 2m_1 m_2$.

With a constant $c$, the solution of $S_m w = cm$ is still in the same direction along the mean difference $m$, and thus is equivalent to solving $S_m^{-1} S_b \Phi = \Phi \Lambda$.

It is easy to show that $(S_m + \lambda I) w = (l_1 m_1 - l_2 m_2)$ is equivalent to $(S_w + \lambda I) w = m_1 - m_2$, simply replacing $S_w$ by $S_w$ and $S_m$ by $S_m + \lambda I$ and $S_m + \lambda I$ in the above analysis.

Appendix B: Multi-Class LDA

In Appendix A, we showed that for a two class problem, LDA and DLA correspond to solving (4) and (12), respectively. $S_m$ can be written as $S_m = U_m \Lambda_m^{1/2} \Lambda_m^{1/2} T_m$. (4) can be written as $\Lambda_m^{1/2} U_m^{T} w = \Lambda_m^{1/2} U_m^{T} m \equiv v$. Note that $v$ can be treated as a “data point.”

For a multi-class problem, Assuming $S_m$ is full rank, LDA corresponds to the system $S_b W = S_m W \Lambda$. It can be write as

$$N_b V = V \Lambda,$$

(19)

where $V = \Lambda_m^{1/2} U_m^{T} S_b U_m \Lambda_m^{-1/2}$. This is a simple eigenvalue problem. By solving $V$, we can compute $W$ by $W = U_m \Lambda_m^{-1/2} V$. The eigenvectors in $V$ with the largest eigenvalues have a one-to-one correspondence to the eigenvectors in $W$ with the corresponding largest eigenvalues in the original equation. Note that $S_b = \sum_{j=1}^{J} (\sqrt{\lambda_j} m_j)(\sqrt{\lambda_j} m_j)^T$ and $N_b$ is

$$N_b = \Lambda_m^{-1/2} U_m^{T} S_b U_m \Lambda_m^{-1/2} = \sum_{j=1}^{J} (\sqrt{\lambda_j} m_j)(\sqrt{\lambda_j} m_j)^T.$$  

(20)

If we treat every vector $\sqrt{\lambda_j} m_j$ as a point, then $N_b$ is the covariance of these $J$ points. Equation (19) can also be viewed as a PCA problem on these $J$ points. In the above analysis if we replace $S_m$ by $S_m + \lambda I$, we obtain that DLA corresponds to solving $S_b W = (S_m + \lambda I) W \Lambda_w$. Further

$$S_b W = (S_w + \lambda I) W \Lambda_w.$$  

(21)

The eigenvectors corresponding to the largest eigenvalues are chosen as the DLA subspace basis.

Appendix C: Overview of Statistical Learning Theory

The first step in the development of learning theory is the assumption of existence of a probability measure $\rho$ on the product space $Z = X \times Y$, from which the data are drawn. One way to define the expected error of $f$ is the least squares error: $\varepsilon(f) = \int_Y f(x) - y)^2 d\rho$ for $f : X \rightarrow Y$.

We try to minimize this error by “learning.” Define $\rho(f) = \int_Y \rho(y|x, f) d\rho$ where $\rho(y|x)$ is the conditional probability measure on $Y$ w.r.t. $x$. It has been shown [4] that the expected error $\varepsilon_{\rho}$ of $f_\rho$ represents a lower bound on the error $\varepsilon(f)$ for any $f$. Hence at least in theory, the function $f_\rho$ is the ideal one and so is often called the target function. However, since measure $\rho$ is unknown, $f_\rho$ is unknown as well.
Starting from the training data \( z = \{(x_i, y_i)\}_{i=1}^l \), statistical learning theory as developed by Vapnik [22] builds on the so-called empirical risk minimization (ERM) induction principle. Given a hypothesis space \( H \), one attempts at minimizing
\[
\frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2, \quad \text{where } f \in H, \quad (22)
\]
However, in general, without controlling the norm of approximation functions, this problem is ill-posed. Generally speaking, a problem is called well-posed if its solution exists, and is unique and stable (depends continuously on the data). A problem is ill-posed if it is not well-posed.

Regularization theory [21] is a device that was developed to turn an ill-posed problem into a well-posed one. The crucial step is to replace the functional in (22) by the following regularized functional
\[
\frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2, \quad f \in H, \quad (23)
\]
where \( \lambda \) is a regularization term, and \( \|f\|_K^2 \) the norm of \( f \) in \( L^2_0(X) \). Now minimizing (23) makes it well-posed and can be solved by elementary linear algebra. Let \( f_{\lambda,z} \) be the minimizer of (23), the question then is: How good an approximation is \( f_{\lambda,z} \) to \( f_{\mu} \), or how small is \( \varepsilon(f_{\lambda,z}) = \int_X (f_{\lambda,z}(x) - f_{\mu}(x))^2 d\rho_X \)? Further, what is the best choice for \( \lambda \) to minimize this error? In [5], the answers to these questions are given. They state: For each \( m \in \mathbb{N} \) and \( \delta \in [0,1) \), there is a function \( E_{m,\delta} = E : \mathbb{R}^+ \rightarrow \mathbb{R} \), such that, for all \( \lambda > 0 \),
\[
\int_X (f_{\lambda,z}(x) - f_{\mu}(x))^2 d\rho_X \leq E(\lambda)
\]
with confidence \( 1 - \delta \). And there is a unique minimizer of \( E(\lambda) \) that is found by a simple algorithm to yield the “best” regularization parameter \( \lambda = \lambda^* \).

References