Efficient Implementation of Optimal Linear Discriminant Analysis

Jurjen Duintjer Tebbens

joint work with **Pavel Schlesinger***

Institute of Computer Science
Academy of Sciences of the Czech Republic
email: tebbens@cs.cas.cz

* Institute of Formal and Applied Linguistics, Charles University, Prague

Outline

1. Introduction: Classification Tasks
2. Fisher's Linear Discriminant Analysis
3. Fisher's criterion for the $p >> n$ problem
4. Efficient Implementation

1. Introduction: Classification Tasks

Classification task:

- Assign a given sample, based on its properties, to pre-defined classes

- Different classifiers utilize different decision rules, e.g. centroid based, nearest neighborhood and support vector machines methods

- Decision rules are derived from information gained from training samples, 'learning process'

- In applications like text document classification or protein fold class prediction, the samples may contain a very large number of variables

- To enhance efficiency one frequently performs a preprocessing step known as dimension reduction of the space of variables

- One of the simplest and most popular methods that incorporate dimension reduction is Fisher's Linear Discriminant Analysis (FLDA).

2. Fisher's Linear Discriminant Analysis

FLDA idea: project the variables on a space of small dimension such that class information is maximally preserved.

This is achieved with between-class- and within-class-variance. With

- n: Number of training samples
- p: Number of variables
- $x_i \in \mathbb{R}^p$: The ith sample
- $\bar{x} \in \mathbb{R}^p$: The grand mean of all samples, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

the *total variance matrix* $\mathbf{T} \in \mathbb{R}^{p \times p}$ is defined as

$$\mathbf{T} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T.$$

With

- g: The number of classes
- n_j: The number of samples in class j
- N_j: The set of indices i for which x_i is in class j
- $\bar{x}_j \in \mathbb{R}^p$: The mean vector in class j, $\bar{x}_j = \frac{1}{n_j} \sum_{i \in N_j} x_i$

the *between-class-variance matrix* $\mathbf{B} \in \mathbb{R}^{p \times p}$ is defined as

$$\mathbf{B} = \frac{1}{g-1} \sum_{j=1}^{g} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

and the *within-class-variance matrix* $\mathbf{W} \in \mathbb{R}^{p \times p}$ is defined as

$$\mathbf{W} = \frac{1}{n-g} \sum_{j=1}^{g} \sum_{i \in N_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T.$$

Clearly, $\mathbf{T}$, $\mathbf{B}$ and $\mathbf{W}$ are positive semi-definite.

One can easily prove that

$$(n-1)\mathbf{T} = (g-1)\mathbf{B} + (n-g)\mathbf{W}.$$

For a vector $v \in \mathbb{R}^p$:

- $v^T\mathbf{T}v$: Total variance
- $v^T\mathbf{B}v$: Between class variance
- $v^T\mathbf{W}v$: Within class variance

FLDA seeks projection vectors $c_i \in \mathbb{R}^p$, $i < g$, such that the projected samples $(c_1, \ldots, c_i)^T x_j$ have maximal between-class-variance and minimal within-class-variance. This leads to *Fisher's criterion*:

Fisher's criterion: A projection vector c_i must satisfy

$$\frac{c_i^T \mathbf{B} c_i}{c_i^T \mathbf{W} c_i} = \max_{c \in \mathbb{R}^p,\, c \neq 0} \frac{c^T \mathbf{B} c}{c^T \mathbf{W} c}.$$

The criterion is formulated for nonsingular $\mathbf{W}$. Then it is equivalent to finding the largest eigenpairs of the generalized eigenproblem

$$(\mathbf{B} - \lambda \mathbf{W})c = 0, \tag{1}$$

which can be transformed to a standard eigenproblem, e.g.

$$(\mathbf{W}^{-1}\mathbf{B} - \lambda \mathbf{I})c = 0. \tag{2}$$

The FLDA-projection space of dimension i, $i < g$, is spanned by the eigenvectors corresponding to the i largest eigenvalues.

3. Fisher's criterion for the $p >> n$ problem

In some modern applications (text document classification, protein fold prediction), the number of variables is so high that one cannot afford to work with the same number of samples (the '$p >> n$ problem'). As a sum of n rank one matrices,

$$\mathbf{W} = \frac{1}{n-g} \sum_{j=1}^{g} \sum_{i \in N_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$$

has rank($\mathbf{W}$) $\leq n$ and $\mathbf{W}$ must be singular.

This makes Fisher's maximization problem

$$\max_{c \in \mathbb{R}^p,\, c \neq 0} \quad \frac{c^T \mathbf{B} c}{c^T \mathbf{W} c}$$

challenging.

One group of strategies for the $p >> n$ problem solves a modified generalized eigenproblem

$$(\mathbf{B} - \lambda \tilde{\mathbf{W}})c = 0,$$

with a matrix

$$\tilde{\mathbf{W}} \approx \mathbf{W}$$

in some sense.

Often, $\tilde{\mathbf{W}}$ is chosen so that we can easily transform to a standard eigenproblem: Small diagonal perturbation, truncated SVD,..., see e.g. [Hong and Yang - 1991], [Cheng, Zhuang and Yang - 1992], [Krzanowski et al. - 1995].

Hence one solves a modified maximization problem

$$\max_{c \in \mathbb{R}^p,\, c \neq 0} \quad \frac{c^T \mathbf{B} c}{c^T \tilde{\mathbf{W}} c}.$$

Others address Fisher's criterion directly [Cheng, Liao, Ko, Lin and Yu - 2000], [Howland, Park et al. - 2003, 2004, 2005]. They argue that

$$\frac{c^T \mathbf{B} c}{c^T \mathbf{W} c} \tag{3}$$

is maximized for $c \in \text{null}(\mathbf{W})$. Hence the best projection vectors are to be chosen from null($\mathbf{W}$).

Indeed, a vector c from null($\mathbf{W}$) trivially has within-variance

$$c^T \mathbf{W} c = 0$$

and because $\mathbf{W}$ is positive semi-definite, the within-variance is minimal.

In addition, $c^T \mathbf{B} c = 0$ should be maximized.

Methods of this type choose their leading projection vectors from null($\mathbf{W}$). If necessary, further projection vectors are obtained by proceeding to the complement of null($\mathbf{W}$).

In the best case, Fisher's original idea to minimize within variance and maximize between variance is preserved, the criterion is modified as

$$\max_{c \in \mathbb{R}^p,\, \mathbf{W}c=0} \quad c^T \mathbf{B} c. \tag{4}$$

We refer to application of this criterion as to optimal linear discriminant analysis. Experiments seem to indicate that optimal linear discriminant analysis leads to the most powerful classification.

4. Efficient Implementation

A straightforward implementation (see e.g. [Krzanowski et al. - 1995] consists of

1. Computation of a spectral decomposition of $\mathbf{W}$
2. Defining the matrix $\mathbf{W}_N$ whose orthogonal columns span the null space of $\mathbf{W}$
3. Finding the leading eigenpairs of $\mathbf{W}_N^T \mathbf{B} \mathbf{W}_N$

$\mathbf{W}$ has rank at most n, hence null($\mathbf{W}$) has dimension at least $p-n$ and so has the eigenproblem in 3. Unfortunately, for $p >> n$ this is still very much.

How can we enhance efficiency?

Exploit the special structure of the variance matrices: For $\mathbf{T}$,

$$\mathbf{T} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{n-1}(\mathbf{X} - \bar{x}e)(\mathbf{X} - \bar{x}e)^T.$$

Here, $\mathbf{X} \in \mathbb{R}^{p\times n}$ is the sample matrix whose ith column contains the ith sample and $e = (1, 1, \ldots, 1)$.

Similarly, $\mathbf{B}$ and $\mathbf{W}$ can be written as products of rectangular matrices $\mathbf{R}_B, \mathbf{R}_W \in \mathbb{R}^{p\times n}$

$$\mathbf{B} = \mathbf{R}_B\mathbf{R}_B^T, \qquad \mathbf{W} = \mathbf{R}_W\mathbf{R}_W^T.$$

It suffices to compute the (economy size) singular value decomposition $\mathbf{U}\Sigma\mathbf{V}^T$ of $\mathbf{R}_W$ to obtain null($\mathbf{W}$). Indeed,

$$\mathbf{W} = \mathbf{U}\Sigma^2\mathbf{U}^T$$

and the singular vectors corresponding to zero singular values span the wanted null space.

Analogously, the eigenpairs of $\mathbf{W}_N^T \mathbf{B} \mathbf{W}_N = \mathbf{W}_N^T \mathbf{R}_B \mathbf{R}_B^T \mathbf{W}_N$ can be found by considering the SVD of $\mathbf{R}_B^T \mathbf{W}_N$.

This SVD technique is used in [Howland, Park et al. - 2003, 2004, 2005] and in the implementation of the statistical software called R-environment. It reduces computational costs significantly.

Storage costs, however, stay high. They are dominated by

$$\mathbf{W}_N \in \mathbb{R}^{p \times (p-n)}.$$

In many applications where p is large (say over 10.000), this is just too much to be able to store.

If a *sparse* matrix were available we could apply a sparse method and save storage costs.

The sample matrix $\mathbf{X}$ is sparse in a number of important applications. Unfortunately, in

$$\mathbf{B} = \frac{1}{g-1} \sum_{j=1}^{g} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

and

$$\mathbf{W} = \frac{1}{n-g} \sum_{j=1}^{g} \sum_{i \in N_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T.$$

sparsity is destroyed by the group means.

How can we combine exploitation of sparsity with optimal LDA and possibly low storage and computational costs ?

We propose an implementation based on eliminating the common null space of **B** and **W**. Vectors of the common null space are uninteresting because both within and between variance are minimal.

The common null space of **B** and **W** is the null space of **T**.
P r o o f : Any $v \in \mathbb{R}^p$ is in the null space of **B** iff $v^T \mathbf{B} v = v^T \mathbf{R}_B \mathbf{R}_B^T v = 0$ and the same holds for **W** and **T**. With

$$(n-1)\mathbf{T} = (g-1)\mathbf{B} + (n-g)\mathbf{W}$$

and the fact that **W** and **B** are positive semi-definite we have:

$$v^T \mathbf{T} v = 0 \quad \Leftrightarrow \quad \frac{v^T}{n-1}\left((g-1)\mathbf{B} + (n-g)\mathbf{W}\right) v = 0 \quad \Leftrightarrow$$

$$v^T \mathbf{B} v = 0 \quad \text{and} \quad v^T \mathbf{W} v = 0.$$

Hence the complement of the common null space is spanned by the eigenvectors corresponding to nonzero eigenvalues of $\mathbf{T}$. The eigendecomposition of

$$\mathbf{T} = \frac{1}{n-1}(\mathbf{X} - \bar{x}e)(\mathbf{X} - \bar{x}e)^T$$

can be obtained with the SVD of $\mathbf{X} - \bar{x}e$. Two crucial points:

- $\mathbf{X} - \bar{x}e$ is a rank one updated sparse matrix, hence the SVD can be computed with a sparse method.
- The complement of the common null space has dimension at most n. Hence after eliminating the common null space, we are left over with a small projected maximization problem

$$\max_{c \in \mathbb{R}^p,\, c \neq 0} \quad \frac{c^T \mathbf{P}^T \mathbf{B} \mathbf{P} c}{c^T \mathbf{P}^T \mathbf{W} \mathbf{P} c}$$

For the projected problem we can apply direct methods to find the largest eigenvectors of $\mathbf{P}^T\mathbf{B}\mathbf{P}$ in null($\mathbf{P}^T\mathbf{W}\mathbf{P}$).

Certainly, one could also compute the SVD of $\mathbf{X} - \bar{x}e$ with a direct method, which is backward stable.

On the other hand, a sparse method may compute singular vectors consecutively, without necessity to store the whole matrix of singular vectors.

The advantage of our elimination of the common null space is that we have the choice between direct and sparse methods.

Thank you for your attention.

This work is supported by the Program Information Society under project 1ET400300415.