The Many Flavors of Penalized Linear Discriminant Analysis

Daniela M. Witten
Assistant Professor of Biostatistics
University of Washington

May 9, 2011
Fourth Erich L. Lehmann Symposium
Rice University
Overview

- There has been a great deal of interest in the past 15+ years in penalized regression,

\[
\min_{\beta} \{||y - X\beta||^2 + P(\beta)\},
\]

especially in the setting where the number of features \(p\) exceeds the number of observations \(n\).

- \(P\) is a penalty function. Could be chosen to promote
  - sparsity: e.g. the lasso, \(P(\beta) = ||\beta||_1\)
  - smoothness
  - piecewise constancy...

- How can we extend the concepts developed for regression when \(p > n\) to other problems?

- A Case Study: Penalized linear discriminant analysis.
The classification problem

- The Set-up:
  - We are given \( n \) training observations \( x_1, \ldots, x_n \in \mathbb{R}^p \), each of which falls into one of \( K \) classes.
  - Let \( y \in \{1, \ldots, K\}^n \) contain class memberships for the training observations.
  - Let \( X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \).
  - Each column of \( X \) (feature) is centered to have mean zero.
The classification problem

- **The Set-up:**
  - We are given $n$ training observations $x_1, \ldots, x_n \in \mathbb{R}^p$, each of which falls into one of $K$ classes.
  - Let $y \in \{1, \ldots, K\}^n$ contain class memberships for the training observations.
  - Let $\mathbf{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$.
  - Each column of $\mathbf{X}$ (feature) is centered to have mean zero.

- **The Goal:**
  - We wish to develop a classifier based on the training observations $x_1, \ldots, x_n \in \mathbb{R}^p$, that we can use to classify a test observation $x^* \in \mathbb{R}^p$.
  - A classical approach: linear discriminant analysis.
Linear discriminant analysis

**FISHER’S DISCRIMINANT PROBLEM**

\[
\maximize_{\beta} \left\{ \beta^T \Sigma_b \beta \right\} \text{ subject to } \beta^T \Sigma_w \beta \leq 1.
\]

**NORMAL MODEL**

\[
x_i | y_i = k \sim N(\mu_k, \Sigma_w)
\]

**OPTIMAL SCORING PROBLEM**

\[
\minimize_{\beta, \theta} \left\{ ||Y\theta - X\beta||^2 \right\} \text{ subject to } \theta^T Y^T Y\theta = 1.
\]
LDA via the normal model

- Fit a simple normal model to the data:

  \[ x_i | y_i = k \sim N(\mu_k, \Sigma_w) \]

- Apply Bayes’ Theorem to obtain a classifier: assign \( x^* \) to the class for which \( \delta_k(x^*) \) is largest:

  \[
  \delta_k(x^*) = x^* \Sigma_w^{-1} \mu_k - \frac{1}{2} \mu_k \Sigma_w^{-1} \mu_k + \log \pi_k
  \]
Fisher’s discriminant

A geometric perspective: project the data to achieve good classification.
Fisher’s discriminant

A geometric perspective: project the data to achieve good classification.
Fisher’s discriminant

A geometric perspective: project the data to achieve good classification.
Fisher’s discriminant

A geometric perspective: project the data to achieve good classification.
Fisher’s discriminant and the associated criterion

Look for the discriminant vector $\beta \in \mathbb{R}^p$ that maximizes

$$\beta^T \hat{\Sigma}_b \beta \text{ subject to } \beta^T \hat{\Sigma}_w \beta \leq 1.$$ 

- $\hat{\Sigma}_b$ is an estimate for the between-class covariance matrix.
- $\hat{\Sigma}_w$ is an estimate for the within-class covariance matrix.
- This is a generalized eigen problem; can obtain multiple discriminant vectors.
- To classify, multiply data by discriminant vectors and perform nearest centroid classification in this reduced space.
- If we use $K - 1$ discriminant vectors then we get the LDA classification rule. If we use fewer than $K - 1$, we get reduced-rank LDA.
LDA via optimal scoring

▶ Classification is such a bother. Isn’t regression so much nicer?
▶ It wouldn’t make sense to solve

$$\min_{\beta} \{ ||y - X\beta||^2 \}.$$ 
▶ But can we formulate classification as a regression problem in some other way?
LDA via optimal scoring

- Let $Y$ be a $n \times K$ matrix of dummy variables; $Y_{ik} = 1_{y_i = k}$.

$$
\begin{align*}
\text{minimize} \{ & ||Y\theta - X\beta||^2 \} \quad \text{subject to} \quad \theta^T Y^T Y \theta = 1.
\end{align*}
$$

- We are choosing the optimal scoring of the class labels in order to recast the classification problem as a regression problem.

- The resulting $\beta$ is proportional to the discriminant vector in Fisher’s discriminant problem.

- Can obtain the LDA classification rule, or reduced-rank LDA.
Linear discriminant analysis

**FISHER’S DISCRIMINANT PROBLEM**
maximize $\{\beta^T \hat{\Sigma}_b \beta\}$ subject to $\beta^T \hat{\Sigma}_w \beta \leq 1$.

**NORMAL MODEL**
$x_i | y_i = k \sim N(\mu_k, \Sigma_w)$

**OPTIMAL SCORING PROBLEM**
minimize $\{||Y\theta - X\beta||^2\}$ subject to $\theta^T Y^T Y \theta = 1$. 
LDA when $p \gg n$

When $p \gg n$, we cannot apply LDA directly, because the within-class covariance matrix is singular.

There is also an interpretability issue:

- All $p$ features are involved in the classification rule.
- We want an **interpretable** classifier. For instance, a classification rule that is a
  
  - sparse,
  - smooth, or
  - piecewise constant

  linear combination of the features.
Penalized LDA

- We could extend LDA to the high-dimensional setting by applying (convex) penalties, in order to obtain an interpretable classifier.
- For concreteness, in this talk: we will use $\ell_1$ penalties in order to obtain a sparse classifier.
- Which version of LDA should we penalize, and does it matter?
Penalized LDA via the normal model

- The classification rule for LDA is
  \[ x^* T \hat{\Sigma}_w^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k T \hat{\Sigma}_w^{-1} \hat{\mu}_k, \]

where \( \hat{\Sigma}_w \) and \( \hat{\mu}_k \) denote MLEs for \( \Sigma_w \) and \( \mu_k \).

- When \( p \gg n \), we cannot invert \( \hat{\Sigma}_w \).

- Can use a regularized estimate of \( \Sigma_w \), such as
  \[
  \Sigma_w^D = \begin{pmatrix}
  \hat{\sigma}^2_1 & 0 & \ldots & 0 \\
  0 & \hat{\sigma}^2_2 & \ddots & \vdots \\
  \vdots & \ddots & \ddots & 0 \\
  0 & \ldots & 0 & \hat{\sigma}^2_p
  \end{pmatrix}.
  \]
Interpretable class centroids in the normal model

- For a sparse classifier, we need zeros in estimate of $\Sigma_w^{-1}\mu_k$.
- An interpretable classifier:
  - Use $\Sigma^D$, and estimate $\mu_k$ according to
    \[
    \text{minimize } \left\{ \sum_{j=1}^{p} \sum_{i:y_i=k} \frac{(X_{ij} - \mu_{kj})^2}{\sigma_j^2} + \lambda \|\mu_k\|_1 \right\}.
    \]
  - Apply Bayes’ Theorem to obtain a classification rule.
- This is the nearest shrunken centroids proposal, which yields a sparse classifier because we are using a diagonal estimate of the within-class covariance matrix and a sparse estimate of the class mean vectors.

Citation: Tibshirani et al. 2003, Stat Sinica
Penalized LDA via optimal scoring

▶ We can easily extend the optimal scoring criterion:

$$\min_{\beta, \theta} \left\{ \frac{1}{n} \| \mathbf{Y} \theta - \mathbf{X} \beta \|^2 + \lambda \| \beta \|_1 \right\} \text{ subject to } \theta^T \mathbf{Y}^T \mathbf{Y} \theta = 1.$$ 

▶ An efficient iterative algorithm will find a local optimum.

▶ We get sparse discriminant vectors, and hence classification using a subset of the features.

Citation: Clemmensen Hastie Witten and Ersboll 2011, Submitted
Penalized LDA via Fisher’s discriminant problem

- A simple formulation:

\[
\max_{\beta} \{ \beta^T \hat{\Sigma} b \beta - \lambda \| \beta \|_1 \} \text{ subject to } \beta^T \tilde{\Sigma}_w \beta \leq 1
\]

where \( \tilde{\Sigma}_w \) is some full rank estimate of \( \Sigma_w \).

- A non-convex problem, because \( \beta^T \hat{\Sigma}_b \beta \) isn’t concave in \( \beta \).

- Can we find a local optimum?

Citation: Witten and Tibshirani 2011, JRSSB
Maximizing a function via minorization
Maximizing a function via minorization
Maximizing a function via minorization
Maximizing a function via minorization
Maximizing a function via minorization
Maximizing a function via minorization
Maximizing a function via minorization
Maximizing a function via minorization
Minorization

- Key point: Choose a minorizing function that is easy to maximize.

- Minorization allows us to efficiently find a local optimum for Fisher’s discriminant problem with any convex penalty.
Connections between flavors of penalized LDA

**Connections**

- **Fisher’s Discriminant Problem** + $\ell_1$
- **Normal Model** + $\ell_1$
- **Optimal Scoring Problem** + $\ell_1$
Connections between flavors of penalized LDA

1. **Normal Model + $\ell_1$**: use a diagonal estimate for $\Sigma_w$ and then apply $\ell_1$ penalty to the class mean vectors.
2. **Optimal scoring + $\ell_1$**: apply $\ell_1$ penalty to discriminant vectors.
3. **Fisher’s discriminant problem + $\ell_1$**: apply $\ell_1$ penalty to discriminant vectors.

So are (1) and (3) different? And are (2) and (3) the same?
Normal Model + $\ell_1$ and Fisher’s + $\ell_1$
Normal Model $+ \ell_1$ and Fisher’s $+ \ell_1$

- Normal model $+ \ell_1$ penalizes the elements of this matrix.
- Fisher’s $+ \ell_1$ penalizes the left singular vectors.
- Clearly these are different...
- ...but if $K = 2$, then they are (essentially) the same.
Normal Model + $\ell_1$ and Fisher’s + $\ell_1$

FISHER’S DISCRIMINANT PROBLEM + $L_1$

Very closely related if $K=2$

NORMAL MODEL + $L_1$

OPTIMAL SCORING PROBLEM + $L_1$
Fisher’s $+ \ell_1$ and Optimal Scoring $+ \ell_1$

Both problems involve “penalizing the discriminant vectors” so they must be the same, right?
Theorem: For any value of the tuning parameter for FD$+\ell_1$, there exists some tuning parameter for OS$+\ell_1$ such that the solution to one problem is a critical point of the other.

▶ In other words – there is a correspondence between the critical points, though not necessarily the solutions.

▶ So the resulting “sparse discriminant vectors” may be different!
Connections

Fisher’s Discriminant Problem + $L_1$

Very closely related if $K=2$

Normal Model + $L_1$

Correspondence between critical points

Optimal Scoring Problem + $L_1$
Pros and Cons

Penalized LDA via normal model:
► (+) In the case of a diagonal estimate for $\Sigma_w$ and $\ell_1$ penalties on mean vectors, it is well-motivated and simple.
► (-) No obvious extension to non-diagonal estimates of $\Sigma_w$.
► (-) Cannot obtain a “low-rank” classifier.

Penalized LDA via Fisher’s discriminant problem:
► (+) Any convex penalties can be applied to discriminant vectors.
► (+) Can use any full-rank estimate of $\Sigma_w$.
► (+) Can obtain a “low-rank” classifier.

Penalized LDA via optimal scoring:
► (+) An extension of regression.
► (+) Any convex penalties can be applied to discriminant vectors.
► (+) Can obtain a “low-rank” classifier.
► (-) Cannot use any estimate of $\Sigma_w$.
► (-) Usual motivation for OS is that it yields the same discriminant vectors as Fisher’s problem. Not true when penalized!
A sensible way to regularize regression when $p \gg n$:

$$\min_{\beta} \{||y - X\beta||^2 + P(\beta)\}.$$ 

One could argue that this is the way to penalize regression.

But as soon as we step away from regression, even to a closely related problem like LDA, the situation becomes much more complex – there is no longer a “single way” to approach the problem.

And the situation becomes only more complex for more complex statistical methods!

Need a principled framework to determine which penalized extension of established statistical methods is “best”.
References
