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,

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + P(\boldsymbol{\beta})\},$$

  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, $P(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$
  - 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.

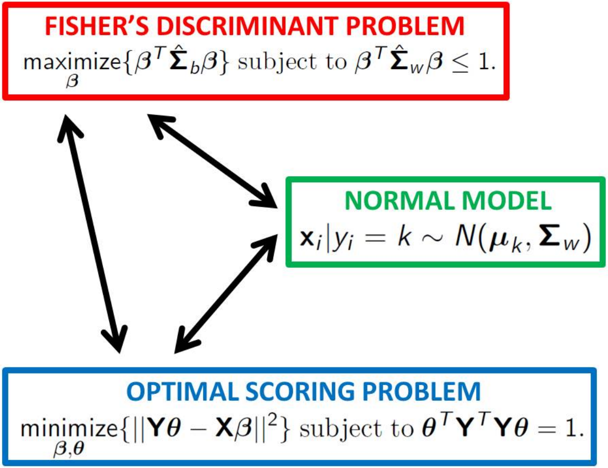
The classification problem

- The Set-up:
  - We are given n training observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, each of which falls into one of K classes.
  - Let $\mathbf{y} \in \{1, \ldots, K\}^n$ contain class memberships for the training observations.
  - Let $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$.
  - Each column of $\mathbf{X}$ (feature) is centered to have mean zero.

The classification problem

- The Set-up:
  - We are given n training observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, each of which falls into one of K classes.
  - Let $\mathbf{y} \in \{1, \ldots, K\}^n$ contain class memberships for the training observations.
  - Let $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$.
  - 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, that we can use to classify a test observation $\mathbf{x}^* \in \mathbb{R}^p$.
  - A classical approach: linear discriminant analysis.

Linear discriminant analysis



LDA via the normal model

- Fit a simple normal model to the data:

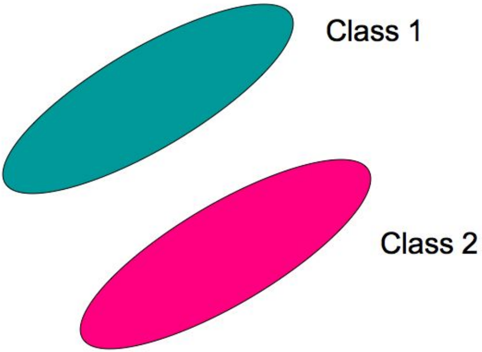
$$\mathbf{x}_i | y_i = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_w)$$

- Apply Bayes' Theorem to obtain a classifier: assign $\mathbf{x}^*$ to the class for which $\delta_k(\mathbf{x}^*)$ is largest:

$$\delta_k(\mathbf{x}^*) = \mathbf{x}^{*T} \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_k + \log \pi_k$$

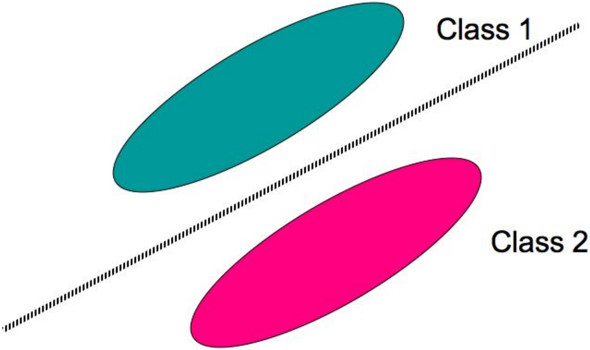
Fisher's discriminant

A geometric perspective: project the data to achieve good classification.



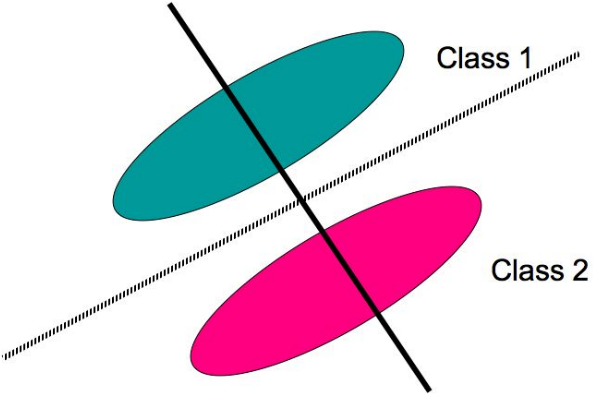
Fisher's discriminant

A geometric perspective: project the data to achieve good classification.



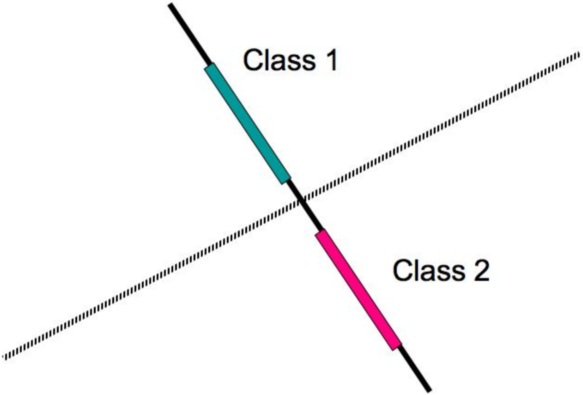
Fisher's discriminant

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 $\boldsymbol{\beta} \in \mathbb{R}^p$ that maximizes

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

- $\hat{\boldsymbol{\Sigma}}_b$ is an estimate for the between-class covariance matrix.
- $\hat{\boldsymbol{\Sigma}}_w$ 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

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\{||\mathbf{y} - \mathbf{X}\beta||^2\}.$$

- But can we formulate classification as a regression problem in some other way?

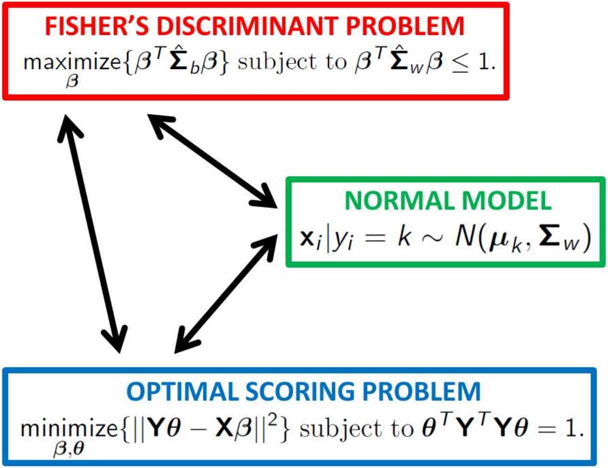
LDA via optimal scoring

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

$$\underset{\boldsymbol{\beta},\boldsymbol{\theta}}{\text{minimize}}\{||\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\beta||^2\} \text{ subject to } \boldsymbol{\theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta} = 1.$$

- We are choosing the optimal scoring of the class labels in order to recast the classification problem as a regression problem.
- The resulting β is proportional to the discriminant vector in Fisher's discriminant problem.
- Can obtain the LDA classification rule, or reduced-rank LDA.

Linear discriminant analysis



LDA when $p \gg n$

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

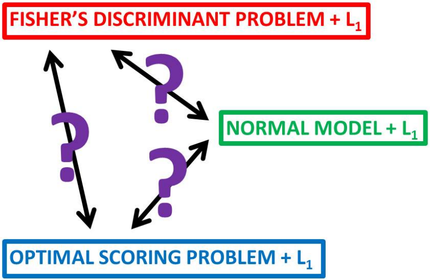
There is also an interpretability issue:

- 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 ℓ_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

$$\mathbf{x}^{*T}\hat{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\mu}}_k - \frac{1}{2}\hat{\boldsymbol{\mu}}_k^T\hat{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\mu}}_k,$$

  where $\hat{\boldsymbol{\Sigma}}_w$ and $\hat{\boldsymbol{\mu}}_k$ denote MLEs for $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\mu}_k$.
- When $p \gg n$, we cannot invert $\hat{\boldsymbol{\Sigma}}_w$.
- Can use a regularized estimate of $\boldsymbol{\Sigma}_w$, such as

$$\boldsymbol{\Sigma}_w^D = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \hat{\sigma}_p^2 \end{pmatrix}.$$

Interpretable class centroids in the normal model

- For a sparse classifier, we need zeros in estimate of $\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_k$.
- An interpretable classifier:
  - Use $\boldsymbol{\Sigma}_w^D$, and estimate $\boldsymbol{\mu}_k$ according to

$$\underset{\boldsymbol{\mu}_k}{\text{minimize}} \left\{ \sum_{j=1}^{p} \sum_{i:y_i=k} \frac{(X_{ij} - \mu_{kj})^2}{\sigma_j^2} + \lambda ||\boldsymbol{\mu}_k||_1 \right\}.$$

  - 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:

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

- 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

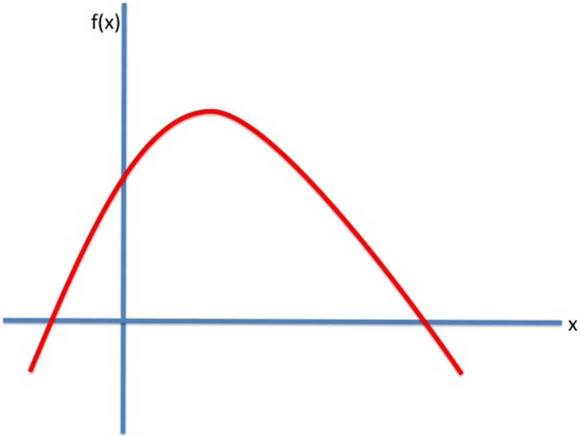
- A simple formulation:

$$\underset{\boldsymbol{\beta}}{\text{maximize}}\{\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta} - \lambda||\boldsymbol{\beta}||_1)\} \text{ subject to } \boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} \leq 1$$

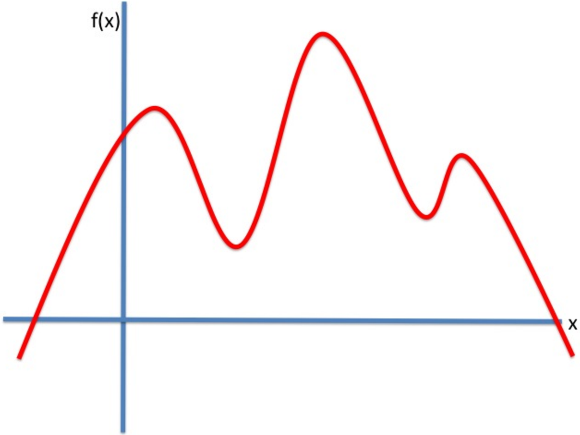
  where $\tilde{\boldsymbol{\Sigma}}_w$ is some full rank estimate of $\boldsymbol{\Sigma}_w$.
- A non-convex problem, because $\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}$ isn't concave in $\boldsymbol{\beta}$.
- Can we find a local optimum?

Citation: Witten and Tibshirani 2011, JRSSB

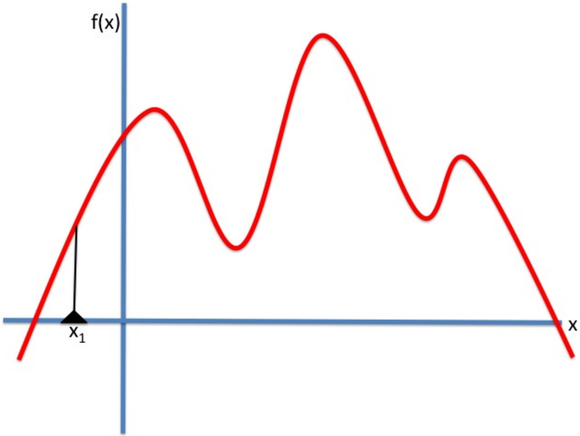
Maximizing a function via minorization



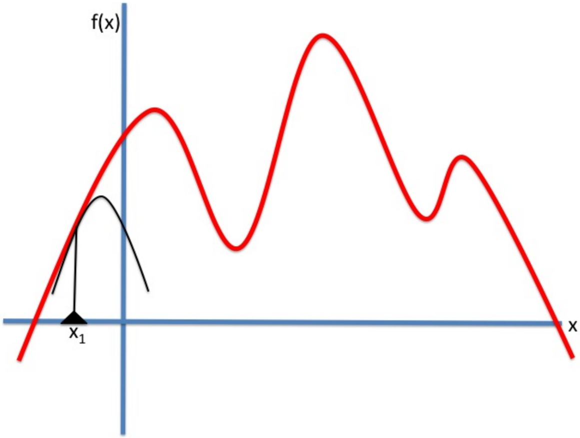
Maximizing a function via minorization



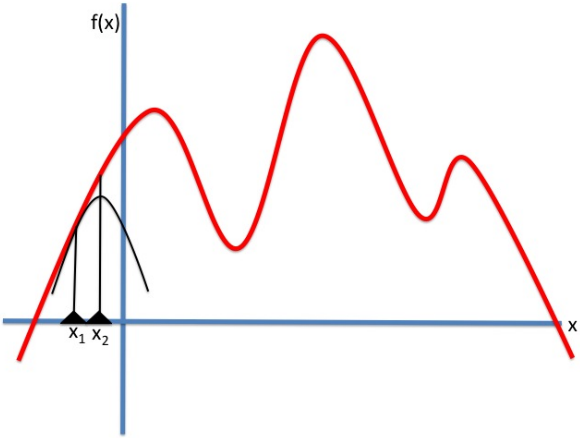
Maximizing a function via minorization



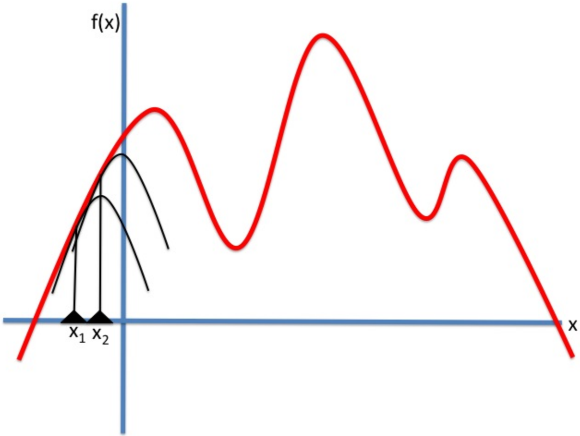
Maximizing a function via minorization



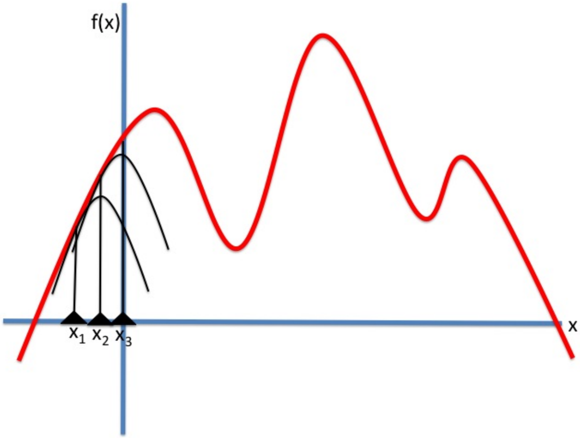
Maximizing a function via minorization



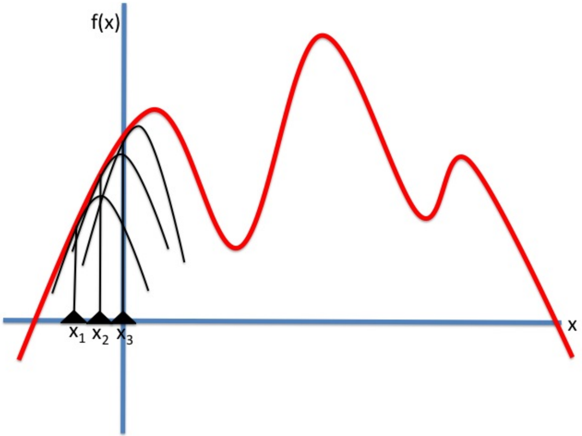
Maximizing a function via minorization



Maximizing a function via minorization

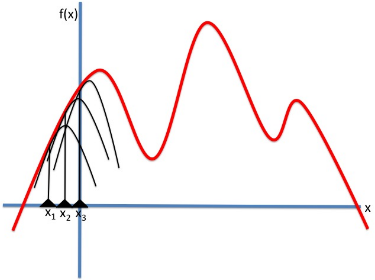


Maximizing a function via minorization



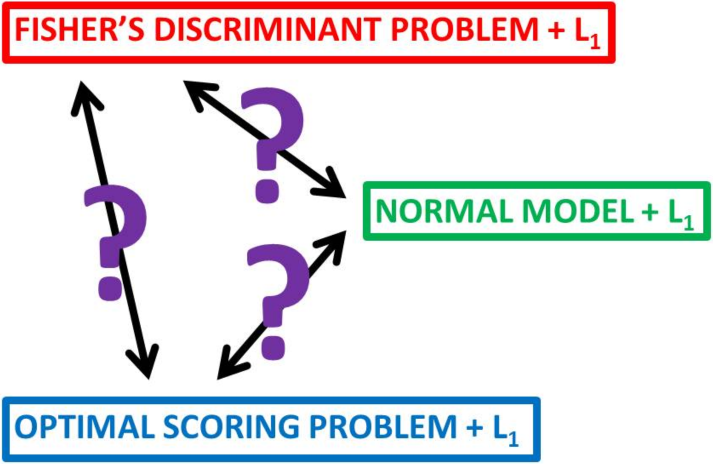
Minorization

- 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

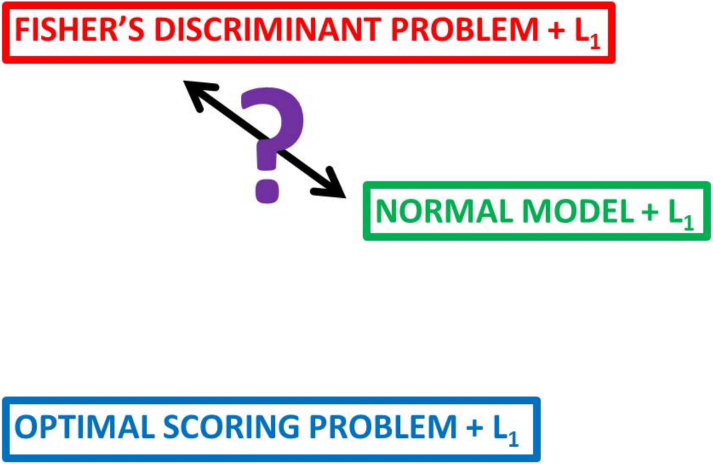


Connections between flavors of penalized LDA

1. Normal Model + ℓ_1: use a diagonal estimate for $\boldsymbol{\Sigma}_w$ and then apply ℓ_1 penalty to the class mean vectors.
2. Optimal scoring + ℓ_1: apply ℓ_1 penalty to discriminant vectors.
3. Fisher's discriminant problem + ℓ_1: apply ℓ_1 penalty to discriminant vectors.

So are (1) and (3) different? And are (2) and (3) the same?

Normal Model + ℓ_1 and Fisher's + ℓ_1

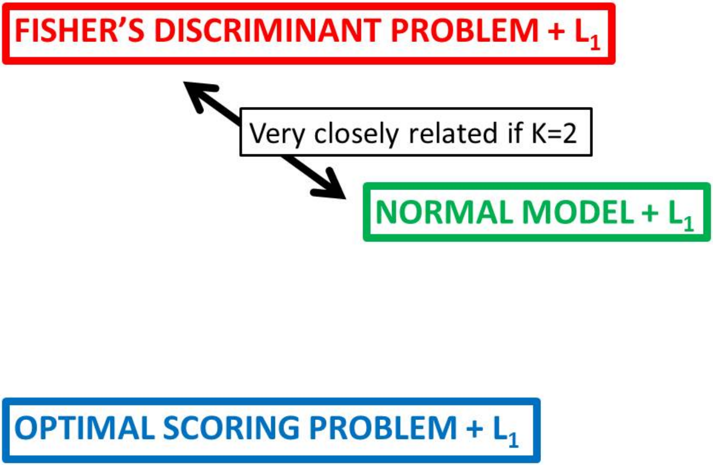


Normal Model + ℓ_1 and Fisher's + ℓ_1

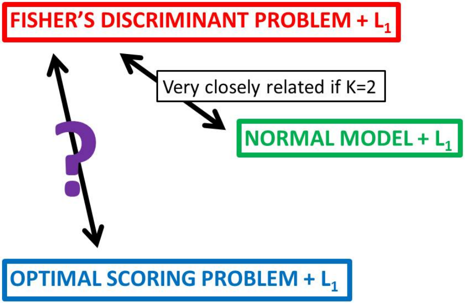
$$\begin{matrix} \mu_{11} & \mu_{21} & & \mu_{K1} \\ \mu_{12} & \mu_{22} & & \mu_{K2} \\ \mu_{13} & \mu_{23} & & \mu_{K3} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \mu_{1p} & \mu_{2p} & & \mu_{Kp} \end{matrix}$$

- Normal model + ℓ_1 penalizes the elements of this matrix.
- Fisher's + ℓ_1 penalizes the left singular vectors.
- Clearly these are different...
- ...but if $K = 2$, then they are (essentially) the same.

Normal Model + ℓ_1 and Fisher's + ℓ_1



Fisher's+ℓ_1 and Optimal Scoring+ℓ_1



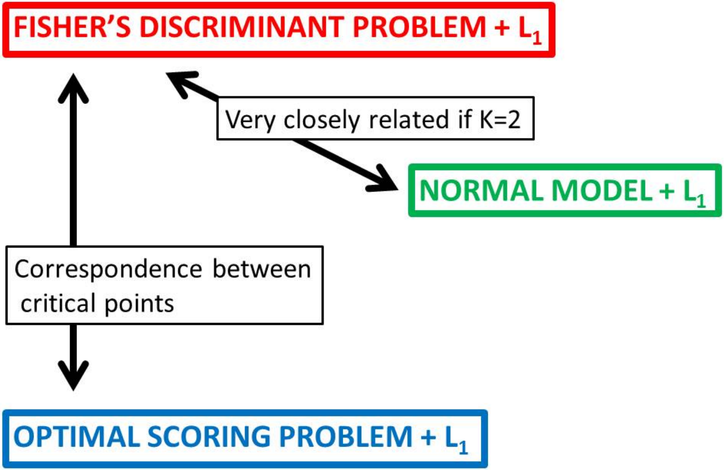
Both problems involve "penalizing the discriminant vectors" so they must be the same, right?

Fisher's+ℓ_1 and Optimal Scoring+ℓ_1

Theorem: For any value of the tuning parameter for FD+ℓ_1, there exists some tuning parameter for OS+ℓ_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



Pros and Cons

Penalized LDA via normal model:

- (+) In the case of a diagonal estimate for $\mathbf{\Sigma}_w$ and ℓ_1 penalties on mean vectors, it is well-motivated and simple.
- (-) No obvious extension to non-diagonal estimates of $\mathbf{\Sigma}_w$.
- (-) 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 $\mathbf{\Sigma}_w$.
- (+) 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 $\mathbf{\Sigma}_w$.
- (-) Usual motivation for OS is that it yields the same discriminant vectors as Fisher's problem. Not true when penalized!

Conclusions

- A sensible way to regularize regression when $p \gg n$:

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\{||\mathbf{y} - \mathbf{X}\beta||^2 + P(\beta)\}.$$

- 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

- Witten and Tibshirani (2011) Penalized classification using Fisher's linear discriminant. To appear in *Journal of the Royal Statistical Society, Series B.*
- Clemmensen, Hastie, Witten, and Ersboll (2011) Sparse discriminant analysis. *Submitted.*