Regularized Discriminant Analysis

Daniela Birkel

09.06.2006

Institut für Statistik münchen

Part I

Linear and Quadratic Discriminant Analysis

Discriminant Analysis

The **purpose of discriminant analysis** is to assign objects to one of several (K) groups based on a set of measurements $X = (X_1, X_2, ..., X_p)$ which are obtained from each object

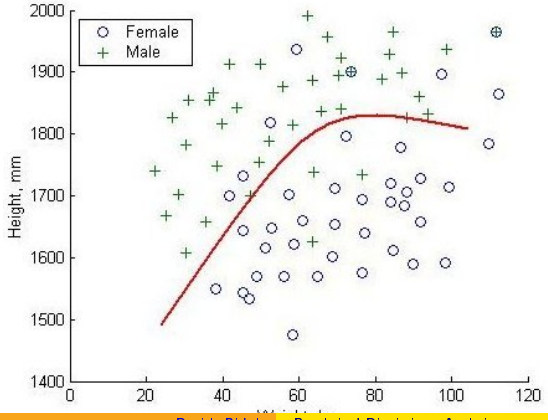
- each object is assumed to be a member of one (and only one) group $1 \leq k \leq K$
- an error is incurred if the object is attached to the wrong group
- the measurements of all objects of one class k are characterized by a probability density $f_k(\mathbf{X})$
- we want to find a rule to decide for every object to which class it belongs to

Example

A group of people consist of male and female persons
$\Rightarrow K = 2$

- from each person the data of their weight and height is collected
  $\Rightarrow p = 2$
- the gender is unknown in the data set
- we want to classify the gender for each person from the weight and height
  $\Rightarrow$ discriminant analysis
- a classification rule is needed (discriminant function) to choose the group for each person
- to construct this function a training sample is used in which the gender is already known

Example



Class distribution

- the distribution of the measurements X are seldom identical in each class
  $\Rightarrow$ **conditional distribution** for each class k
- most often applied classification rules are based on the multivariate normal distribution

$$f_k(\mathbf{X}) = f(\mathbf{X} \mid k) = \frac{1}{(2\pi)^{p/2} |\sum_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \sum_k^{-1} (x-\mu_k)}$$

  where μ_k and Σ_k are the class k $(1 \leq k \leq K)$ population mean vector and covariance matrix
- $f_k(\mathbf{X})$ are seldom known

Prior and unconditional distribution

There might be some prior knowledge about the probability of observing a member of class k

- **prior probability**

$$\pi_k \qquad \text{with} \qquad \pi_1 + ... + \pi_k = 1$$

- if the prior probabilities are equal for each k
  $\Rightarrow$ leads to Maximum-Likelihood
- to estimate the class conditional densities $f_k(X)$ and the prior probability π_k a **training sample** with already correctly classified classes is used

the **unconditional distribution** of X is given by

$$f(\mathbf{X}) = \sum_{k=1}^{K} \pi(k) f(\mathbf{X} \mid k)$$

Posterior distribution

Probability for one object with given vector $X^T = (X_1, ..., X_p)$ to belong to class k is calculated by the Bayes formula

$$p(k \mid \mathbf{X}) = \frac{\overbrace{f(\mathbf{X} \mid k)}^{\text{class distribution}} \cdot \overbrace{\pi(k)}^{\text{prior probablitiy}}}{\underbrace{f(\mathbf{X})}_{\text{unconditional distribution}}} \propto f(\mathbf{X} \mid k) \cdot \pi(k)$$

posterior distribution

an object is assigned to **class** $\hat{k}$, if it has the biggest posterior probability $p(\hat{k} \mid \mathbf{X})$
$\Rightarrow$ this is equal to minimizing the expected loss

Log posterior distribution

For easier calculation we take the **logarithm of the posterior distribution**

$$\log p(k \mid \mathbf{X}) = \log f(\mathbf{X} \mid k) + \log \pi(k)$$

- with the multivariate normal distribution it leads to

$$\begin{aligned}\log p(k \mid \mathbf{X}) &= \log((2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}) + \log \pi_k \\ &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| \\ &\quad + \log \pi_k + \text{constant} \end{aligned} \tag{1}$$

- the constant term $-\frac{p}{2}\log(2\pi)$ can be omitted as it is the same for each class k

Quadratic discriminant analysis

multiplication with -2 leads to the **discriminant function**

$$d_k(\mathbf{X}) = \underbrace{(\mathbf{X}-\mu_k)^T\Sigma_k^{-1}(\mathbf{X}-\mu_k)}_{Mahalanobis-distance} + \log|\Sigma_k| - 2\log\ \pi_{k)}$$

and to the **classification rule**

$$d_{\hat{k}}(X) = \min_{1\le k\le K} d_k(X) \quad \Leftrightarrow \quad \max_{1\le k\le K} p(k \mid X)$$

Using this rule is called the
Quadratic Discriminant Analysis (QDA)

Linear discriminant analysis

A special case occurs when all k class covariance matrices are identical

$$\Sigma_k = \Sigma$$

The discriminant function

$$d_k(x) = (x-\mu_k)^T\Sigma^{-1}(x-\mu_k) - 2\log\pi(k)$$

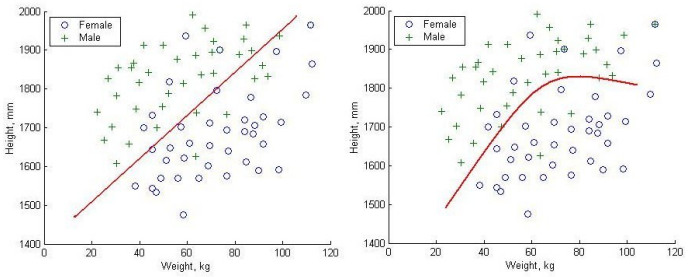
simplifies to

$$d_k(x) = 2\mu_k^T\Sigma^{-1}X - \mu_k^T\Sigma^{-1}\mu_k - 2\log\pi(k)$$

This is called the **Linear Discriminant Analysis (LDA)** because the quadratic terms in the discriminant function cancel:

- $x^T\Sigma^{-1}x$ is the same in every class k and can be left out
- the decision boundaries are now linear

Linear and Quadratic Boundaries



Estimation

In most applications of linear and quadratic discriminant analysis the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated by their sample analogs

$$\hat{\mu}_k = \overline{X}_k = \frac{1}{N_k}\begin{bmatrix} \sum_{i=1}^{N} X_{n1} \\ \vdots \\ \sum_{i=1}^{N} X_{np} \end{bmatrix} = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

and

$$\hat{\Sigma}_k = \frac{S_k}{N_k} = \frac{1}{N_k}\sum_{c(v)=k}(X-\overline{X}_k)(X-\overline{X}_k)^T$$

with $c(v) =$ class of vth observation

Part II

Regularized Discriminant Analysis

Small sample sizes

These estimates are straightforward to compute and represent the corresponding maximum likelihood estimates.

Problem: they are only optimal for $n \to \infty$ and not for small n

Small sample sizes

- the $(p \times p)$ covariance matrix estimates become highly variable
- not all of the parameters are even identifiable
- Σ is singular
- the inverse Σ^{-1} does not exist

Small sample sizes

- **poorly-posed**
  $\Rightarrow$ the number of parameters to be estimated is comparable to the number of observations
- **ill-posed**
  $\Rightarrow$ that number exceeds the sample size

	QDA	LDA
poorly-posed	$N_k \approx p$	$N \approx p$
ill-posed	$N_k \le p$	$N \le p$
parameters to be estimated	$k \cdot p^2 + p$	$p^2 + p$

QDA requires generally larger samples size than LDA

Regularization

For ill- or poorly-posed situations:

- parameter estimates can be highly unstable
- high variance

The **aim of regularization** is to improve the estimates by biasing them away from their sample-based values

- reduction of variance at the expense of potentially increased bias
- the bias variance trade-off is regulated by two parameters
- these parameters control the strength of the biasing

Regularization for Quadratic Discriminant Analysis

Strategy 1: If QDA is ill- or poorly-posed

- Replacing the individual class sample covariance matrices by their average (pooled covariance matrix)

$$\hat{\Sigma} = \frac{\sum_{k=1}^{K} S_k}{\sum_{k=1}^{K} N_k}$$

- regularization by reducing the number of parameters to be estimated
- this can result in superior performance, especially in small-sample settings
- leads to LDA

$\Rightarrow$ the choice between Linear and Quadratic Discriminant Analysis is quite restrictive

Regularization with parameter λ

Strategy 2: A less limited approach is represented by

$$\hat{\Sigma}_k(\lambda) = (1-\lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma}$$

with $0 \leq \lambda \leq 1$

- λ controls the degree of shrinkage of the individual class covariance matric estimates toward the pooled estimate
- $\lambda = 0$ gives rise to QDA
- $\lambda = 1$ gives rise to LDA

$\Rightarrow$ still fairly limited
$\Rightarrow$ cannot be used if LDA is ill- or poorly posed

Eigenvalues

If $N \leq p$ then even LDA is poorly- or ill-posed

- $\hat{\Sigma}$ is singular
- some eigenvalues are 0

decomposing Σ with the **spectral decomposition** leads to

$$\Sigma^{-1} = \sum_{i=1}^{p} \frac{v_{ik} v_{ik}^T}{e_{ik}}$$

e_{ik} ith eigenvalue of Σ_k
v_{ik} ith eigenvector of Σ_k

$\Rightarrow \hat{\Sigma}^{-1}$ does not exist

Further regularization

Strategy 3 If LDA is ill- or poorly-posed

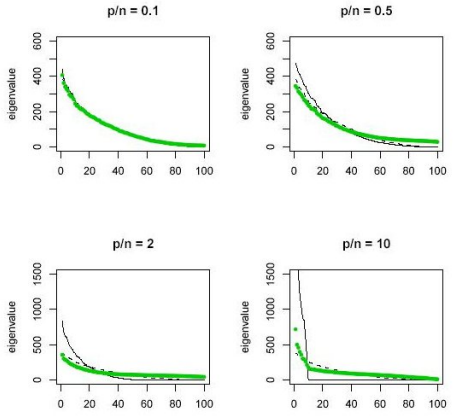
$$\hat{\Sigma}_k(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \gamma \underbrace{\frac{\mathrm{tr}\,[\hat{\Sigma}_k(\lambda)]}{p}}_{average\ \ eigenvalue} \mathbb{I}$$

with $0 \leq \gamma \leq 1$
tr A = sum of eigenvalues

- the additional regularization parameter γ controls shrinkage toward a multiple of the identity matrix for a given value of λ
- decreasing the larger eigenvalues and increasing the smaller ones

$\Rightarrow$ shrinkage toward the average eigenvalue of $\hat{\Sigma}_k(\lambda)$

Eigenvalues



Regularized Discriminant Analysis

The discriminant function for the **Regularized Discriminant Analysis (RDA)** is

$$d_k(\mathbf{X}) = (\mathbf{X}-\bar{\mathbf{X}}_k)^T\hat{\Sigma}_k^{-1}(\lambda,\gamma)(\mathbf{X}-\bar{\mathbf{X}}_k) + \log|\hat{\Sigma}_k(\lambda,\gamma)| - 2\log\ \pi(k)$$

- the values for λ and γ are not likely to be known in advance
  $\Rightarrow$ we have to estimate them
- the aim is to find values for λ and γ that jointly minimize the future misclassification risk

Methods:

- bootstrapping
- cross-validation

Model selection

Idea of **cross-validation** (leave-one-out)

- one particular observation X_v is removed from the model
- the classification rule is developed on the $N-1$ training observations without X_v
- then this classification rule is used to classify X_v and to calculate the loss which occurs if classified to the wrong group
- this is repeated for every observation

$\Rightarrow$ the future misclassification risk is estimated by the average of the resulting misclassification loss
this is done for a number of combinations for λ and γ

Conclusion

The potential for **RDA** to improve misclassification risk over that of **QDA** or **LDA** depend on the situation

- $\mathbf{N_k \gg p}$
  no regularization is needed and QDA can be used
  $\Rightarrow$ model-selection procedure should tend to small values of λ and γ
- $\mathbf{N \approx p}$
  LDA has been the method of choice in the past $\Rightarrow$ regularization can substantially improve the misclassification risk when
  - Σ_k are not close to being equal
  - N is even too small for LDA

Part III

Prediction Analysis with Microarrays (PAM)

Class prediction with gene expression data

The aim is to assign people to one of $K (1 \leq k \leq K)$ diagnostic categories based on their gene expression profile

The classification by DNA microarrays is challenging because:

- there is a very large number of genes (p) from which to predict classes and only a relatively small number of samples (N)
  $\Rightarrow$ again a restricted form of Discriminant Analysis
- identify the genes which contribute most to the classification
  $\Rightarrow$ reduction of p

Example

Small round blue cell tumors (SRBCT) of childhood can be divided in four groups

- Burkitt lymphome (BL)
- Ewing sarcoma (EWS)
- neuroblastoma (NB)
- rhabdomyosarcoma (RMS)

The DNA microarrays of 88 children with SRBCT were obtained

- 63 of them were already classified right and their data were used as the training sample to estimate the classification rule
- the category for the other 25 children (of which 5 were not SRBCT) was then predicted by this rule
- the aim is to correctly classify the test samples

Example

The data consist of expression measurements on **2,308 genes**

- the mean expression value (centroid) was calculated from the training sample for each of the four classes
- then the squared distance from the gene expression profile to each class centroids was calculated for each test sample
- the predicted class for a child was the one with the closest centroid
  $\Rightarrow$ **nearest centroid classification**

It would be more attractive if fewer genes were needed
$\Rightarrow$ modification to **nearest shrunken centroid classification** where the genes which don't contribute for the class prediction are eliminated

Shrunken centroids

$\bar{x}_{ik} = \sum_{j \epsilon C_k} \frac{x_{ij}}{n_k}$ mean value in class k for gene i

$\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$ overall centroid for gene i

d_{ik} is a t **statistic for gene** i, comparing the mean of class k to the overall centroid

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)}$$

$m_k = \sqrt{1/n_k + 1/n}$
s_i pooled standard deviation for gene i
s_0 same value for every gene, positive constant

Shrunken Centroids

The distance for gene i between the mean in class k and the overall mean is shrunk toward zero $\Rightarrow$ **soft thresholding**

$$d'_{ik} = sign(d_{ik})(|d_{ik}| - \Delta)_+$$

with
Δ shrinkage parameter, also called threshold

$$t_+ = \begin{cases} t & \text{if t} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- many of the genes are eliminated as Δ increases
- Δ is again chosen by cross-validation ($\Delta = 4.34$ in the example)

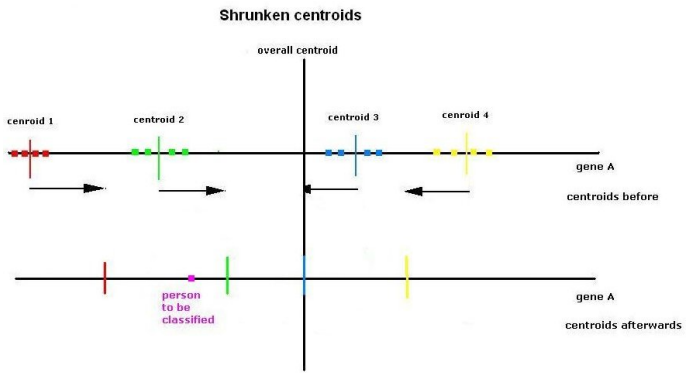
Shrunken centroids

The centroids are shrunk towards the overall centroid

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik}$$

- if $d'_{ik} = 0$ for every class k
  $\Rightarrow \bar{x}'_{i1} = ... = \bar{x}'_{iK}$
  the centroid for each class is the same
- if each group has the same mean for one gene, this gene does not help to predict a class and can be left out
- in the example only 43 genes are needed for the class prediction and 2,275 are not needed to distinguish between the groups

Shrunken centroids



Discriminant Function

The discriminant function uses the Mahalanobis metric in computing distances to centroids:

$$\delta_k^{LDA}(x^*) = \underbrace{(x^* - \bar{x}_k)^T}_{1\times p} \underbrace{\Sigma^{-1}}_{p\times p} \underbrace{(x^* - \bar{x}_k)}_{p\times 1} - 2log\pi_k$$

with

π_k class prior probability
W pooled within-class covariance matrix

- Σ is huge as $p \gg n$ and any sample estimate will be singular
- to cope with this problem a **heavily restricted form of LDA** is used
  $\Rightarrow \Sigma$ is assumed to be a **diagonal matrix**

Classification rule

The discriminant function can be rewritten as

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}_{ik}')^2}{(s_i + s_0)^2} - 2log\pi_k$$

- again standardized by $s_i + s_0$
- the discriminant function is for one person with p genes $x^* = (x_1^*, x_2^*, ..., x_p^*)$
- the person is assigned to group $\hat{k}$ if

$$\delta_{\hat{k}}(x^*) = \min_{1\leq k\leq K} \delta_k(x^*)$$

Literature

- Friedmann, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*
- Tibshirani R., Hastie T., Nasasimhan B., Chu G. (2002), "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *PNAS*