Generic Setup

Data: $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbb{R}^p$.

Categorical variables are coded using dummy variables.

We collect the x-values in a big matrix

$$\mathbf{X} = \left\{ \begin{array}{c} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{array} \right\} = \left\{ \begin{array}{ccc} x_{1,1} & \ldots & x_{1,p} \\ x_{2,1} & \ldots & x_{2,p} \\ \vdots & & \vdots \\ x_{N,1} & \ldots & x_{N,p} \end{array} \right\}$$

with dimensions $N \times p$.

If a we need to work with a categorical x-coordinate that can occur on K different levels we can encode the variable as a K-dimensional vector of zero's and one's containing just a single one. This coding is known as dummy variables. If $K = 2$ there are the two possible outcomes $(1, 0)$ and $(0, 1)$.

Figure 14.22 – Threes

In this example the resulting data matrix $\mathbf{X}$ is 130×256.

Linear Algebra - the Mean Value

Matrix computations and decompositions is the key to many theoretical results, and practical success relies heavily on efficient matrix computations.

With $\mathbf{1}$ the N-dimensional vector with one's at all positions, the *column means* can be computed as

$$\bar{x}^T = \frac{1}{N}\mathbf{1}^T\mathbf{X}$$

The *projection* in $\mathbb{R}^N$ onto $\mathbf{1}$ and the orthogonal complement $\mathbf{1}^\perp$ are given by the matrices

$$P = \frac{1}{N}\mathbf{1}\mathbf{1}^T, \quad I_N - P = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T,$$

respectively.

Linear Algebra - the Covariance Matrix

The empirical covariance matrix is

$$\begin{aligned}(N-1)\hat{\Sigma} &= (\mathbf{X}-\mathbf{1}\bar{x}^T)^T(\mathbf{X}-\mathbf{1}\bar{x}^T)\\ &= (\mathbf{X}-P\mathbf{X})^T(\mathbf{X}-P\mathbf{X})\\ &= ((I_N-P)\mathbf{X})^T(I_N-P)\mathbf{X}\\ &= \mathbf{X}^T(I_N-P)\mathbf{X}\end{aligned}$$

since $(I_N-P)^2 = I_N - P$.

Often we will use the augmented matrix $\{\mathbf{1}\ \mathbf{X}\}$ and often we will assume that $\mathbf{X}$ has then been orthogonalized with $\mathbf{1}$. This means that $\mathbf{X}$ has been replaced with $(I_N-P)\mathbf{X} = \mathbf{X}-\mathbf{1}\bar{x}^T$. This *does not change the column space* of $\{\mathbf{1}\ \mathbf{X}\}$.

Singular Value Decomposition

Theorem 1. *If* $\mathbf{X}$ *is an* $N \times p$ *matrix there exists an* $N \times p$ *matrix* U*, a* $p \times p$ *matrix* V *and a diagonal matrix*

$$D = \left\{\begin{array}{ccc} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_p \end{array}\right\}$$

such that $U^TU = I_p$, $V^TV = I_p$, $d_1 \geq \ldots \geq d_p \geq 0$ *and*

$$\mathbf{X} = UDV^T.$$

We call $d_1, \ldots, d_p$ the *singular values.* V is an orthogonal matrix with $V^{-1} = V^T$. The columns in U with corresponding $d_i > 0$ form an orthonormal basis for the column space of $\mathbf{X}$.

Figure 14.20 – Dimension Reduction

A one dimensional representation of 2D data points is sought.

The natural idea is to minimize the sum of squared distances from the line to the data points *perpendicular to the line.*

This differs from linear regression where we consider the sum of distances *parallel to the 2nd coordinate axis.*

Dimension Reduction and Projections

How can we visualize the data in $\mathbf{X}$? What is a good *low-dimensional projection* $P : \mathbb{R}^p \rightarrow \mathbb{R}^p$ with rank 1, 2 or 3?

With

$$V = \{V_q\ V_{p-q}\}$$

where V_q is $p \times q$, the projection onto the columns of V_q is

$$P_q = V_q V_q^T.$$

Then P_q *minimizes* among all rank q projections the *reconstruction error*

$$\sum_{i=1}^{N} ||x_i - P_q x_i||^2 = \text{trace}((\mathbf{X} - \mathbf{X}P_q)(\mathbf{X} - \mathbf{X}P_q)^T)$$

Note that a computation of the reconstruction error by computing the $N \times N$ matrix $(\mathbf{X} - \mathbf{X}P_q)(\mathbf{X} - \mathbf{X}P_q)^T$ and then computing the trace is a computational waste. All the non-diagonals in the matrix product are not needed.

We will generally always replace $\mathbf{X}$ by $\mathbf{X} - \mathbf{1}\bar{x}^T$ before attempting a projection onto a subspace. Because the p coordinates we measure by no means need to be measured on a common scale it is often also most relevant to normalize the columns to have unit length before we attempt a dimension reduction. That is, we divide each column by its empirical standard error. If there are other ways to bring all variables measured on a common scale that might be preferred. We should note that the projections obtained from the singular value decomposition is not invariant to marginal scaling of the columns in $\mathbf{X}$.

Figure 14.21 – Dimension Reduction and PC

The coordinates for the P_q projections of the data points in the V_q basis are called the q first principal components.

The coordinates are

$$\begin{aligned} XV_q &= UDV^T V_q \\ &= UD\text{diag}(1, \ldots, 1, 0, \ldots, 0) \\ &= U_q D_q \end{aligned}$$

with U_q and D_q the matrices with the q first columns from U and D, respectively.

Figure 14.23 – Two First Principal Components for Threes

The first principal component shows primarily the variation in how wide the hand written threes are. The second shows primarily the variation in how thick the drawn line is.

Figure 14.23 – Two First Principal Components for Threes

All pixel values are measured on the same scale so we would only centralize – not scale – the columns.

One-dimensional Normal Variables

Let X be real valued and $X|Y=k$ be $N(\mu_k,\sigma^2)$ for $k=1,2$. If $\Pr(Y=k)=\pi_k$ the Bayes classifier is

$$f(x)=\begin{cases} 1 & \text{if } \pi_1 f_1(x) \geq \pi_2 f_2(x) \\ 2 & \text{if } \pi_1 f_1(x) < \pi_2 f_2(x) \end{cases}$$

Or

$$f(x)=\begin{cases} 1 & \text{if } \log(f_1(x)/f_2(x)) \geq \log(\pi_2/\pi_1) \\ 2 & \text{if } \log(f_1(x)/f_2(x)) < \log(\pi_2/\pi_1) \end{cases}$$

Or

$$f(x)=\begin{cases} 1 & \text{if } 2x(\mu_1-\mu_2) \geq 2\sigma^2\log(\pi_2/\pi_1)-\mu_2^2+\mu_1^2 \\ 2 & \text{if } 2x(\mu_1-\mu_2) < 2\sigma^2\log(\pi_2/\pi_1)-\mu_2^2+\mu_1^2 \end{cases}$$

Linear Discriminant Analysis

Let Y take values in $\{1,\ldots,K\}$ with

$$\Pr(Y=k)=\pi_k$$

with $\pi_1+\ldots+\pi_K=1$, and let the conditional distribution of $X|Y=k$ be $N(\mu_k,\Sigma)$ on $\mathbb{R}^p$ with Σ regular. That is, the density for $X|Y=k$ is

$$g_k(x)=\frac{1}{\sqrt{2\pi\det(\Sigma)}^p}e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}.$$

The conditional probability of $Y=k|X=x$ is

$$\Pr(Y=k|X=x)=\frac{\pi_k g_k(x)}{\pi_1 g_1(x)+\ldots+\pi_K g_1(x)}$$

The Bayes Classifier

$$\begin{aligned}\log\frac{\Pr(Y=k|X=x)}{\Pr(Y=l|X=x)} &= \log\frac{\pi_k}{\pi_l}+\log\frac{g_k(x)}{g_l(x)} \\ &= \log\frac{\pi_k}{\pi_l}+\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}(x-\mu_l)-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) \\ &= \log\frac{\pi_k}{\pi_l}+\frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l-\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k+x^T\Sigma^{-1}(\mu_k-\mu_l)\end{aligned}$$

The boundary – the x's where $\Pr(Y=k|X=x)=\Pr(Y=l|X=x)$ – is a hyperplane. We call this a *linear classifier* as we can determine the classification by the computation of the finite number of linear functions $x^T\Sigma^{-1}(\mu_k-\mu_l)$, $k,l=1,\ldots,K$.

Linear Discriminant Functions

Introducing

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^T\mu_k + \log\pi_k$$

we see that

$$\log\frac{\Pr(Y=k|X=x)}{\Pr(Y=l|X=x)} = \delta_k(x) - \delta_l(x)$$

The decision boundaries are the solutions to the linear equations

$$\delta_k(x) = \delta_l(x)$$

and the Bayes classifier is

$$f(x) = \text{argmax}_k\delta_k(x).$$

Figure 4.5 – Linear Discrimination

Estimation

We use the *the plug-in principle* for estimation. That is, maximum likelihood estimation of all the parameters in the full model for (X, Y)

$$\begin{aligned}\hat{\pi}_k &= \frac{N_k}{N}, \quad N_k = \sum_{i=1}^{N} 1(y_i = k)\\ \hat{\mu}_k &= \frac{1}{N_k}\sum_{i:y_i=k} x_i\\ \hat{\Sigma} &= \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \hat{\mu}_k)^2\end{aligned}$$

– with the usual centralized estimate of the covariance matrix.

Estimation

If A is the $N \times K$ design matrix, the projection onto its column space is $P = A(A^TA)^{-1}A^T$ and we can write

$$\begin{aligned}\hat{\mu}^T &= (A^TA)^{-1}A^T\mathbf{X}\\ \hat{\Sigma} &= \frac{1}{N-K}(\mathbf{X} - P\mathbf{X})^T(\mathbf{X} - P\mathbf{X})\\ &= \frac{1}{N-K}\mathbf{X}^T(I_N - P)\mathbf{X}\end{aligned}$$

The design matrix A is given by

$$A_{i,j} = \begin{cases} 1 & \text{if } y_i = j\\ 0 & \text{if } y_i \neq j\end{cases}$$

Parameter Functions

Fixing the last group K as a reference group we have for $k = 1, \ldots, K-1$ that

$$\begin{aligned}\log \frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)} &= \underbrace{\log \frac{\pi_k}{\pi_K} + \frac{1}{2}\mu_K^T \Sigma^{-1} \mu_K - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k}_{\beta_{k0}} \\ &\quad + x^T \underbrace{\Sigma^{-1}(\mu_k - \mu_K)}_{\beta_k}\end{aligned}$$

Thus

$$\Pr(Y=k|X=x) = \frac{\exp(\beta_{k0} + x^T \beta_k)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + x^T \beta_l)}$$

for $k = 1, \ldots, K-1$. The conditional distribution depends upon $\pi_1, \ldots, \pi_{K-1}, \mu_1, \ldots, \mu_k, \Sigma$ through the parameter function

$$\tau : \mathbb{R}^{K-1} \times \mathbb{R}^{Kp} \times \mathrm{PD}_p \to \mathbb{R}^{K-1} \times \mathbb{R}^{(K-1)p}$$

$$\tau(\pi_1, \ldots, \pi_{K-1}, \mu_1, \ldots, \mu_K, \Sigma) = (\beta_{10}, \ldots, \beta_{(K-1)0}, \beta_1, \ldots, \beta_{K-1}).$$

We use PD_p to denote the set of $p \times p$ positive definite matrices. Note that it is more problematic to define any simple parameter functions ρ such that the marginal distribution of X depends upon ρ only.

Quadratic Discriminant Analysis

What if $\Sigma_1 \neq \Sigma_2$ $(K=2)$?

$$\log \frac{\Pr(Y=k|X=x)}{\Pr(Y=l|X=x)} = \bar{\delta}_k(x) - \bar{\delta}_l(x)$$

where

$$\bar{\delta}_k(x) = -\frac{1}{2} \log \det \Sigma_k - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k).$$

is a *quadratic function.* The decision boundaries are the solutions to the quadratic equations $\bar{\delta}_k(x) = \bar{\delta}_l(x)$ and the Bayes classifier is

$$f(x) = \mathrm{argmax}_k \bar{\delta}_k(x).$$

Figure 4.6 – Quadratic Discrimination

To get quadratic boundaries one can either do QDA (right) or one can transform the bivariate variable $X = (X_1, X_2)^T$ to the five dimensional variable $X' = (X_1, X_2, X_1^2, X_1X_2, X_2^2)$ and do LDA in $\mathbb{R}^5$ (left). The linear boundary in $\mathbb{R}^5$ shows up as a quadratic boundary in $\mathbb{R}^2$.

If the linear boundary $\mathbb{R}^5$ is given by

$$\beta^T X' + \beta_0 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2 + \beta_5 X_2^2 = 0$$

we see that in terms of X_1 and X_2 this is a quadratic equation. Note that due to the transformation $X \mapsto X'$ there is no chance that X' can be 5-dimensional, regular normal distribution. The methodology can, however, still be useful.

Regularized Estimation

Can we estimate Σ – or Σ_k – to the needed precision? What if p is large?

Regularization or *shrinkage estimation* may be a solution.

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

for $\alpha \in [0,1]$ when we do QDA.

$$\gamma\hat{\Sigma} + (1-\gamma)\text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$$

for $\gamma \in [0,1]$ when we do LDA.

Or even

$$\gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I_p$$

for $\gamma \in [0,1]$.

Bias-Variance Tradeoff

For the vowel data we try the use of different convex combinations of the estimated covariance matrices $\hat{\Sigma}_k$ and the common estimated covariance matrix $\hat{\Sigma}$.

Figure 4.4 – Dimension Reduction

Linear discriminant analysis provides a direct dimension reduction to the K-dimensional space. The above figure shows a further reduction to a 2D projection chosen to *maximize the spread of the group means.*

Figure 4.9 – Discrimination and Dimension Reduction

How to project to maximize the spread of group means? The usual inner product in Euclidean space is not optimal – we should use the inner product given by Σ^{-1}

Change of Basis Point of View

If $\Sigma = cVD^2V^T$ with D a diagonal matrix with strictly positive entries and $c > 0$ we let $\tilde{x} = D^{-1}V^T x$ and $\tilde{\mu}_k = D^{-1}V^T\mu_k$. This is a *change of basis* given by the matrix $D^{-1}V^T$. With R a constant not depending on k we have

$$\begin{aligned}\log \Pr(Y=k|X=x) &= \log \pi_k - \frac{1}{2c}(x-\mu_k)^T V D^{-2} V^T (x-\mu_k) + R \\ &= \log \pi_k - \frac{||\tilde{x}-\tilde{\mu}_k||^2}{2c} + R.\end{aligned}$$

Hence

$$\text{argmax}_k \Pr(Y=k|X=x) = \underset{k}{\text{argmin}}\, ||\tilde{x}-\tilde{\mu}_k||^2 - 2c\log \pi_k.$$

If A denotes the affine space spanned by $\tilde{\mu}_1, \ldots, \tilde{\mu}_K$ and Q the projection onto that space we have that $Q\tilde{x} - \tilde{\mu}_k \perp \tilde{x} - Q\tilde{x}$ and we see that

$$\text{argmax}_k \Pr(Y = k | X = x) = \underset{k}{\text{argmin}} \, ||Q\tilde{x} - \tilde{\mu}_k||^2 - 2c \log \pi_k.$$

Assuming that $A = \mu_0 + \text{span}\{v_1^*, \ldots, v_K^*\}$ where $v_1^*, \ldots, v_K^*$ constitute and othonormal basis in the usual inner product and $\mu_0 \in A$ we find that

$$Q\tilde{x} = \mu_0 + \sum_{k=1}^{K} (\tilde{x}^T v_k^*) v_k^*$$

and

$$\begin{aligned} ||Q\tilde{x} - \tilde{\mu}_k||^2 &= \sum_{k=1}^{K} ((\tilde{x}^T v_k^* - \tilde{\mu}_k^T v_k^*)^2 \\ &= \sum_{k=1}^{K} (x^T D^{-1} V v_k^* - \mu_k^T D^{-1} V v_k^*)^2 \end{aligned}$$

Thus a practical solution for computing the classifier is to first compute one such orthonormal basis $v_1^*, \ldots, v_K^*$ and then compute the vectors

$$\text{LD}_k = D^{-1} V v_k^*.$$

For a given x is we use the formula above to compute the distance from $\tilde{x}$ to $\tilde{\mu}_k$ for each $k = 1, \ldots, K$ and classify to the group k with the smallest distance – modulo the correction given by $-2c \log \pi_k$. If all groups are equally probably we can ignore this correction, and otherwize it has the effect of adding a larger number to the least probable groups.

The next construction for a given dataset provides one such choice of orthonomal basis where we seek (for plotting purposes) to sequentially maximize the discrimination of the group means for each choice of basis vector. Thus the first vectors provide the best discrimination of the group means and the last provide the least discrimination.

LDA as Dimension Reduction Technique

If $\mathbf{X} - P\mathbf{X} = UDV^T$ is the singular value decomposition we get that the estimated covariance matrix is

$$\hat{\Sigma} = \frac{1}{N-K} V D^2 V^T$$

Choosing a *change of basis* given by the matrix $D^{-1}V^T$ the resulting data matrix will have empirical covariance matrix $\frac{1}{N-K} I$.

In the changed basis we let

$$(P\mathbf{X} - \mathbf{1}\bar{x}^T) V D^{-1} = U^* D^* V^{*T}$$

denote the singular value decomposition. The minimal rank q reconstruction error – measured using the usual Euclidean norm – for the deviations of the group means to the total mean is spanned by the first q columns in V^*. Then the columns of $VD^{-1}V^*$ are the *canonical coordinates.*

The so-called “sphering” of the data is not a unique operation. If we apply any orthogonal transformation to the data after one “sphering”, we still get a data matrix with the empirical covariance matrix proportional to the unit matrix. Therefore, we could also apply the change of basis given by $VD^{-1}V^T$, which is equal to $\sqrt{N-K}\hat{\Sigma}^{1/2}$.

Figure 4.8 – Dimension Reduction