Chapter 4. Classification methods
Part 1

March 24, 2007

1 A brief review of Linear classification method

Suppose each sample $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ belongs to one of two classes, denoted by 0 and 1 respectively. We need to set up a method. For each new sample, we can easily discriminate its class.

Classic statistical method assumes that there are two populations: A and B, with mean and variance matrix are respectively μ_A and Σ_A and μ_B and Σ_B.

For any new sample X_{new}, the basic criterion is the probability of $P(X_{new} \in A)$ and $P(X_{new} \in B)$.

$$\text{if } P(X_{new} \in A) > P(X_{new} \in B), \quad \text{then } X_{new} \in A$$

$$\text{if } P(X_{new} \in B) > P(X_{new} \in A), \quad \text{then } X_{new} \in B$$

If further the normal distribution is assumed, then the distributions of A and B are respectively

$$f_A(x) = (2\pi)^{-p/2}|\Sigma_A|^{-1/2}\exp\{-\frac{1}{2}(x-\mu_A)^\top\Sigma_A^{-1}(x-\mu_A)\}$$

and

$$f_B(x) = (2\pi)^{-p/2}|\Sigma_B|^{-1/2}\exp\{-\frac{1}{2}(x-\mu_B)^\top\Sigma_B^{-1}(x-\mu_B)\}$$

The criteria is

$$\text{if } f_A(X_{new}) > f_B(X_{new}), \quad \text{then } X_{new} \in A$$

$$\text{if } f_B(X_{new}) > f_A(X_{new}), \quad \text{then } X_{new} \in B$$

[please note the difference between the two criteria above]

If normal distribution is not assumed, we can simply calculate $(x-\mu_A)^\top \Sigma_A (x-\mu_A)$ or $(x-\mu_B)^\top \Sigma_B (x-\mu_B)$, and compare them with a threshold b_0,

$$\text{if } (X_{new}-\mu_A)^\top \Sigma_A (X_{new}-\mu_A) < b_0, \quad \text{then } X_{new} \in A$$

$$\text{if } (X_{new}-\mu_A)^\top \Sigma_A (X_{new}-\mu_A) \geq b_0, \quad \text{then } X_{new} \in B$$

In practice, we need to estimated μ_A and Σ_A and μ_B and Σ_B based on what we known (a learning procedure). Suppose We have n_A samples $X_{A,1}, ..., X_{A,n_A}$ from A and n_B samples $X_{B,1}, ..., X_{B,n_B}$ from B. These samples are called training set. Then we can estimate

$$\hat{\mu}_A = n_A^{-1} \sum_{i=1}^{n_A} X_{A,i}, \quad \hat{\Sigma}_A = n_A^{-1} \sum_{i=1}^{n_A} (X_{A,i}-\hat{\mu}_A)(X_{A,i}-\hat{\mu}_A)^\top$$

and

$$\hat{\mu}_B = n_B^{-1} \sum_{i=1}^{n_B} X_{B,i}, \quad \hat{\Sigma}_B = n_B^{-1} \sum_{i=1}^{n_B} (X_{B,i}-\hat{\mu}_B)(X_{B,i}-\hat{\mu}_B)^\top$$

1.1 Fisher's linear discriminant

In practice, we have have problem with the estimation of covariance matrices when p is large.

Fisher's linear discriminant is a classification method that projects high-dimensional data X onto a line and performs classification in this one-dimensional space $\beta^\top X = \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p$. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. This defines the Fisher criterion, which is maximized over all linear projections β

$$\frac{(m_A - m_B)^2}{s_A^2 + s_B^2} \quad \text{a function of } \beta$$

where

$$m_A = \beta^\top \mu_A, \quad m_B = \beta^\top \mu_B, \quad s_A^2 = \beta^\top \Sigma_A \beta, \quad s_B^2 = \beta^\top \Sigma_B \beta$$

In signal theory, this criterion is also known as the signal-to-interference ratio. Maximizing this criterion yields a closed form solution that involves the inverse of a covariance-like matrix.

Based on the training set. We need to find a linear combination of X:

$$\beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p$$

Let $Z_{Ai} = \beta_1 \mathbf{x}_{A,i,1} + ... + \beta_p \mathbf{x}_{A,i,p}$. We need to maximize

$$\frac{(\beta^\top \hat{\mu}_A - \beta^\top \hat{\mu}_B)^2}{\beta^\top \hat{\Sigma}_B \beta + \beta^\top \hat{\Sigma}_B \beta}$$

with respect to β. Suppose the solution to the above minimization problem is $\hat{\beta}$. We need to find a threshold b_0' and classify a new sample X_{new}

$$\text{if } \hat{\beta}_1 \mathbf{x}_{1,new} + ... + \hat{\beta}_p \mathbf{x}_{p,new} < \hat{b}_0', \quad \text{then } X_{new} \in \text{A (or B)}$$
$$\text{if } \hat{\beta}_1 \mathbf{x}_{1,new} + ... + \hat{\beta}_p \mathbf{x}_{p,new} \geq \hat{b}_0', \quad \text{then } X_{new} \in \text{B (or A)}.$$

which is equivalent to (by setting $\beta_0' = \hat{\beta}_0'$)

$$\text{if } \hat{\beta}_1 \mathbf{x}_{1,new} + ... + \hat{\beta}_p \mathbf{x}_{p,new} + \mathbf{b}_0 < 0, \quad \text{then } X_{new} \in \text{A (or B)}$$
$$\text{if } \hat{\beta}_1 \mathbf{x}_{1,new} + ... + \hat{\beta}_p \mathbf{x}_{p,new} + b_0 \geq 0, \quad \text{then } X_{new} \in \text{B (or A)}.$$

We call

$$\beta_1 \mathbf{x}_{1,new} + ... + \hat{\beta}_p \mathbf{x}_{p,new} + b_0$$

A separating hyperplane.

We learn the threshold by minimizing classification errors on the training set.

1.2 estimating the separating hyperplane

Let Y denote the class of X. A simple way is to assign 1 to Y for the samples in A, and 0 in B. Therefore, we can consider the linear regression model

$$Y_i = \beta_0 + \beta^\top X_i + \varepsilon_i, \quad i = 1, 2, ..., n.$$

By the least squares estimation method, we can estimate the parameters as

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

(what are $\mathbf{X}$ and $\mathbf{Y}$) The predicted Y of X_{new} is

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1^\top X_{new}$$

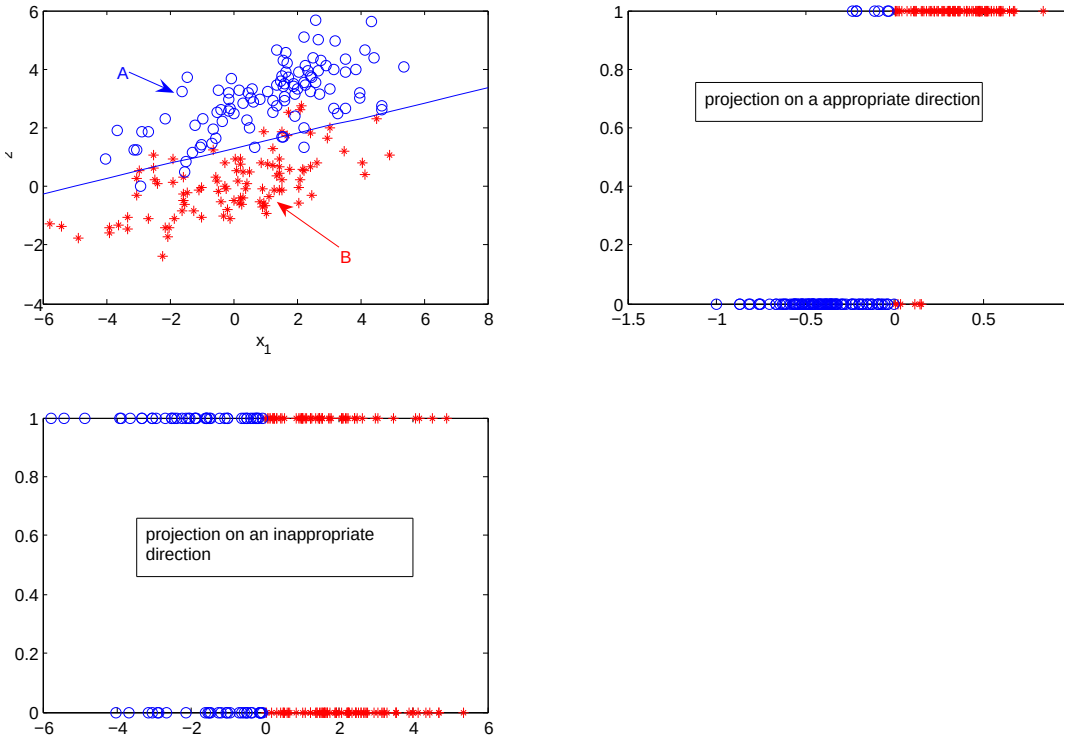

Figure 1: An example

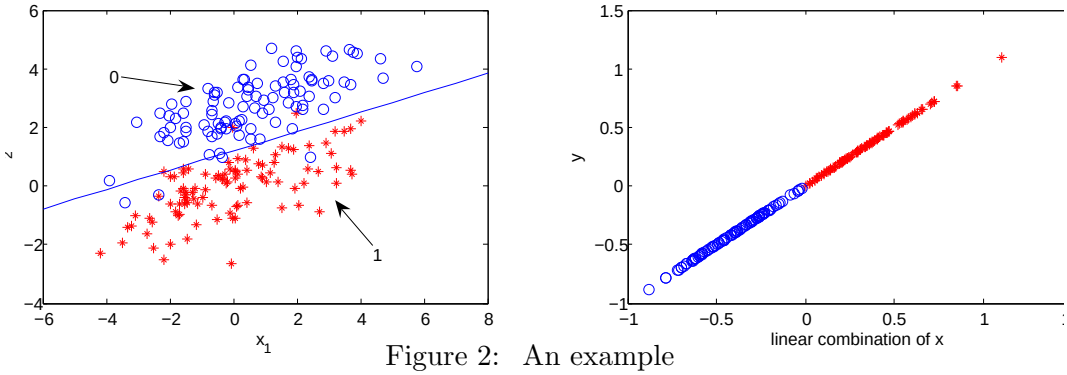

Figure 2: An example

We also need a threshold c such that if $\hat{Y}_{new} > c$, $X_{new} \in A$ otherwise $X_{new} \in B$. c can be selected by minimizing classification errors on the training set. See the illustration in figure 2.

Note that our separating hyperplane is now actually the

$$f(x) = \hat{\beta}_0 - c + \hat{\beta}_1^\top x.$$

Note that this linear model is not a suitable model for the problem (because Y takes values 0 and 1 only). Statisticians immediately have a solution to this: the logistic regression.

$$y = \frac{\exp(\beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p)}{1 + \exp(\beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p)}$$

1.3 Handling nonlinear separating hyperplane (feature space)

In the above discussion, we use a linear hyperplane to separate the classes see figure 3. However, this might not be always true. To include nonlinear separating surface, we can change the representation of the data

$$x = (x_1, ..., x_p) \longmapsto \phi(x) = (\phi_1(x), ..., \phi_m(x)).$$

We call $\{\phi(X)\}$ the feature space. with appropriately selected feature space, we can usually find a separating hyperplane; see figure 4. That is our separating hyperplane is

$$f(x) = \beta_1 \phi_1(x) + ... + \beta_m \phi_m(x) + \beta_0$$

Again, we can estimate $\beta_1, ..., \beta_m$ by minimizing

$$\sum_{i=1}^{n} \{Y_i - f(X_i)\}^2 = \sum_{i=1}^{n} \{Y_i - [\beta_1 \phi_1(x) + ... + \beta_m \phi_m(x) + \beta_0]\}^2$$

with respect to $\beta_0, ...\beta_m$. The solution is

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{pmatrix} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y},$$

where

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1 & \phi_1(X_1) & ... & \phi_m(X_1) \\ 1 & \phi_1(X_2) & ... & \phi_m(X_2) \\ ... & & & \\ 1 & \phi_1(X_n) & ... & \phi_m(X_n) \end{pmatrix}$$

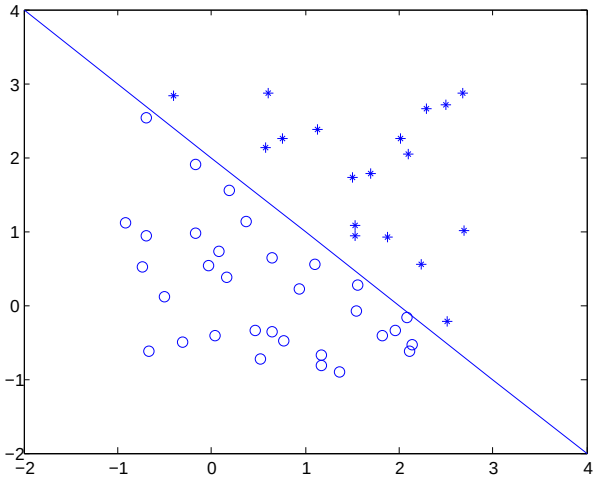

Figure 3: An example

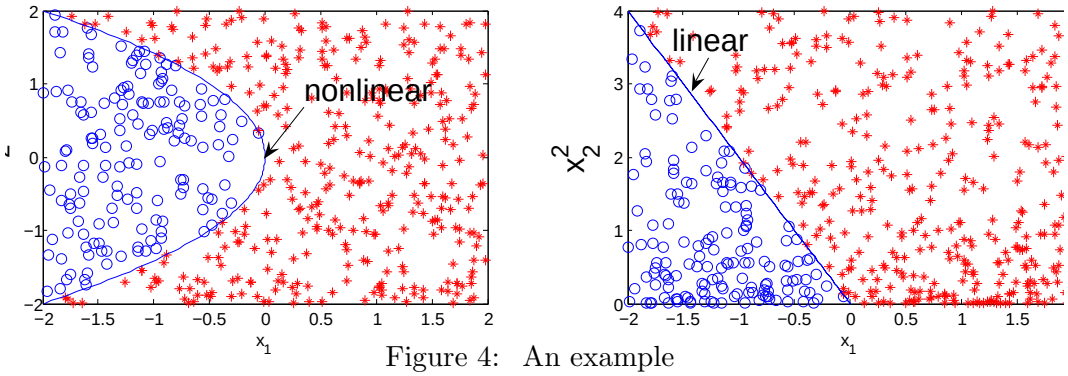

Figure 4: An example

As we noticed from the previous discussion, m could be very large. To estimate $\beta_0, \beta_1, ..., \beta_m$ might have overfitting problem. i.e. $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ does not have inverse matrix. A simple way to avoid this problem is to estimate $\beta_1, ..., \beta_m$ by the so-called ridge regression,

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{pmatrix} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \lambda I)^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y},$$

where I is the identity matrix and $\lambda > 0$ is small and can be chosen by the CV method. The solution is equivalent to that of

$$\sum_{i=1}^{n} \{Y_i - [\beta_1 \phi_1(x) + ... + \beta_m \phi_m(x) + \beta_0]\}^2 + \lambda(\beta_0^2 + \beta_1^2 + ... + \beta_m^2)$$

The second term is called the penalty term. The method is called the penalized least squares estimation.

References

N. Cristianini and J. Shawe-Taylor (2000) *AN INTRODUCTION TO SUPPORT VECTOR MACHINES (and other kernel-based learning methods)* Cambridge University Press 2000