Course: Math 6397 Spring 2015
Instructor: Robert Azencott
Lecture Date:2/3/2015
Notes prepared by Peixin Zhang
Draft 3

In order to use Linear discrimination classification, we assume that the data have an approximate Gaussian distribution.

1. Gaussian distribution

   Let $X \in \mathbb{R}^k$ be a random vector. For any event $A \subseteq \mathbb{R}^k$, probability distribution of X is defined by

$$Pty(x \in A) = \int_A f(x)\,dx = \underbrace{\int \cdots \int}_{k} \{1\}_A(x_1, \ldots, , x_k) f(x_1, \ldots, x_k)\,dx_1 \ldots dx_k$$

   with $f(x) \geq 0$ and $\int_{\mathbb{R}^k} f(x)\,dx = 1$. $f(x)$ is the density function of X.

   Example: uniform density function over cube $[-1, 1]^k$ is $g(x) = \frac{1}{2^k}$.

1.1 Gaussian density in dimension 1:

$$f(x) \sim \mathcal{N}(m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

   where m is fixed and is equal to the average value of X.

   σ^2 is variance and is equal to the average value of $(X - m)^2$

1.2 Gaussian density with multiple dimension N:

   X be a random vector as before.

   Average of X: $M = \int_{\mathbb{R}^k} x f(x)\,dx$.

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \text{ implies } \mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix}$$

   Covariance matrix of X: average value of $(X - M)(X - M)^*$, defined as

$$\Sigma = \mathbb{E}[\underbrace{(X - M)}_{k\times 1}\underbrace{(X - M)^*}_{k\times 1}]$$

   which is a symmetric and positive definite matrix.

   Multidimensional Gaussian density:

$$f(x) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} \exp[-\frac{(x - M)^*\Sigma^{-1}(x - M)}{2}]$$

with $f(x) \geq 0$ and $\int_{\mathbb{R}^k} f(x)\,dx = 1$.

Each coordinate of vector x is also has a Gaussian distribution.

Elements of covariance matrix Σ_{ij} are defined as follow:

$$\Sigma_{ij} = Cov(X_i, X_j) = \mathbb{E}[(X_i - M_i)(X_j - M_j)^*]$$

which also can be written as integral form

$$\Sigma_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [x_i - \mathbb{E}(X_i)][x_j - \mathbb{E}(X_j)]\,dx_i\,dx_j.$$

In order to quantify the level of linear relationship between two random vector X_i and X_j, correlation has been defined as follows

$$\rho_{X_i,X_j} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}$$

Remark: Random variable mean is estimated by sample average, accuracy is of the order of $\frac{1}{\sqrt{n}}$ and n is sample size. Since the covariance matrix has $\frac{k(k+1)}{2}$ elements, estimate accuracy is even worse due to both the fact that mean sample estimation introduces errors and that a large group of covariance matrix elements have to be estimated.

2. Discrimination between two classes of data

   Assume all objects in the dataset has k coordinates.

   Suppose $i.i.d$ random vectors $X(1), X(2), \cdots, X(N))$ belong to first class, while random vectors $(Y(1), Y(2), \cdots, Y(M))$ belong to second class. Given a new object z, we want to make decision about which class z belongs to.

   Object: to find a linear separator $w_1 z_1 + \cdots + w_k z_k + b$. And we will have follow as a decision rule: when $w_1 z_1 + \cdots + w_k z_k + b > 0$, object z belongs to class 1, otherwise it will come from class 2.

   Based on maximum likelihood principle, we will derive a classification rule.

   Look at arbitrary $\mathfrak{z} \in \mathbb{R}^k$. If $\mathfrak{z} \in$ population 0, then it has density function $f_0(\mathfrak{z})$. $\mathfrak{z}$ may come from population 1 with density function $f_1(\mathfrak{z})$.To classify our object $\mathfrak{z}$, we make decision as follow:

   $$argmax[f_0(\mathfrak{z}), f_1(\mathfrak{z})]$$

   Suppose $f_0(\mathfrak{z}) > f_1(\mathfrak{z})$, $\mathfrak{z}$ will be classified as an element of class 0 . Instead, when we have $f_0(\mathfrak{z}) < f_1(\mathfrak{z})$, $\mathfrak{z}$ will be classified as an element of class 1.

3. Hyperplane computation

   Assume X,Y are Gaussian random vectors with $\mathbb{E}(X) = A$, $\mathbb{E}(Y) = B$ and $Cov(X) = Cov(Y) = \Sigma$.

   Usually, we can estimate mean and covariance from our dataset.

$$\hat{A} = \frac{1}{N}\sum_{1}^{N} X_i \text{ and } \hat{B} = \frac{1}{M}\sum_{1}^{M} Y_i$$

$$\hat{\Sigma_1} = \mathbb{E}[(X - \hat{A})(X - \hat{A})^*] \text{ and } \hat{\Sigma_2} = \mathbb{E}[(X - \hat{B})(X - \hat{B})^*]$$

i.e. fixed i, j,

$$\hat{\Sigma_{ij}} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - \hat{C}_i)(x_{kj} - \hat{C}_j)$$

and

$$\hat{\Sigma} = \{\hat{\Sigma_{ij}}\}$$

with

$$\hat{C}_i = \frac{1}{N}\sum_{1}^{N} x_i \text{ and } \hat{C}_j = \frac{1}{N}\sum_{1}^{N} x_j$$

being random variable mean estimation.

By taking weighted average value:

$$\hat{\Sigma} = \frac{N}{N+M}\hat{\Sigma_1} + \frac{M}{N+M}\hat{\Sigma_2}$$

Define

$$f_0(\mathfrak{z}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} \exp[-\frac{(\mathfrak{z} - \hat{A})^*\Sigma^{-1}(\mathfrak{z} - \hat{A})}{2}]$$

$$f_1(\mathfrak{z}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} \exp[-\frac{(\mathfrak{z} - \hat{B})^*\Sigma^{-1}(\mathfrak{z} - \hat{B})}{2}]$$

Take ratio and compute logarithm

$$\log \frac{f_0}{f_1} = -\frac{(\mathfrak{z} - \hat{A})^*\Sigma^{-1}(\mathfrak{z} - \hat{A})}{2} + \frac{(\mathfrak{z} - \hat{B})^*\Sigma^{-1}(\mathfrak{z} - \hat{B})}{2}$$

$$2\log \frac{f_0}{f_1} = -(\mathfrak{z} - \hat{A})^*\Sigma^{-1}(\mathfrak{z} - \hat{A}) + (\mathfrak{z} - \hat{B})^*\Sigma^{-1}(\mathfrak{z} - \hat{B})$$

Decision rule

If $2\log \frac{f_0}{f_1} > 0$, $z \in$ population 0, otherwise $z \in$ population 1.

Next, simplify decision rule

Let $T = \Sigma^{-1}$, and $T^* = T$,

$$-[\mathfrak{z}^*T\mathfrak{z} - \hat{A}^*T\mathfrak{z} - \mathfrak{z}^*T\hat{A} + \hat{A}^*T\hat{A})] + [\mathfrak{z}^*T\mathfrak{z} - \hat{B}^*T\mathfrak{z} - \mathfrak{z}^*T\hat{B} + \hat{B})^*T\hat{B})]$$

$$-[\mathfrak{z}^*T\mathfrak{z} - 2\hat{A}^*T\mathfrak{z} + \hat{A}^*T\hat{A})] + [\mathfrak{z}^*T\mathfrak{z} - 2\hat{B}^*T\mathfrak{z} + \hat{B}^*T\hat{B})]$$

$$2(\hat{A}^* - \hat{B}^*)T\mathfrak{z} - \hat{A}^*T\hat{A} + \hat{B}^*T\hat{B}$$

Define a vector W and b by $2(\hat{A}^* - \hat{B}^*)T = W^*$ and $-\hat{A}^*T\hat{A} + \hat{B}^*T\hat{B} = b$, our new decision rule becomes:

If $W^*\mathfrak{z} + b > 0$,$z \in$ population 0, otherwise $z \in$ population 1.

And since $W = \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}$, we have a hyperplane of form $Sep(\mathfrak{z}) = w_1 z_1 + \cdots + w_k z_k + b = 0$ to classify our above two classes.

Evaluate performance

On training data set X, define R_0 as the number of objects be correctly classified as elements of population 0, then percentage of correct decision is $\frac{R_0}{N}$; same method, $\frac{R_1}{M}$ is percentage of correct decision on data set Y.

Next, In order to generalize our LDC result to real problems, we wants to estimate how accurately our hyperplane will perform in practice. Our hyperplane is tested against testing dataset.Go to testing dataset $(x(N+1), x(N+2), \cdots, x(N+R))$ and $(y(M+1), y(M+2), \cdots, y(M+R)$ to check the percentage of correct classification.

Remark: Drawback of Linear discrimination classification:

1. We assume that the datasets are Gaussian distributed.

2. Estimation of mean and covariances does introduce some errors.

3. Using linear hyperplane to do classify may be a poorly performing separator. Dataset classes have to be concentrated on the space. Suppose one dataset class breaks into several distinct pieces, we have to use more sophisticated classification methods. LDC can not be a good separator.