Linear Discriminant Analysis (线性判别分析)

1. LDA 判别

(1) Model Assumption.

Let π_k be the prior probability of class k, i.e., $\sum_k \pi_k = 1$. Suppose $f_k(x)$ is the conditional density function of X given the class $G = k$ (we use G here to denote the class label). By Bayes Theorem,

$$P(G = k|X = x) = \frac{\pi_k f_k(x)}{\sum_k \pi_k f_k(x)}$$

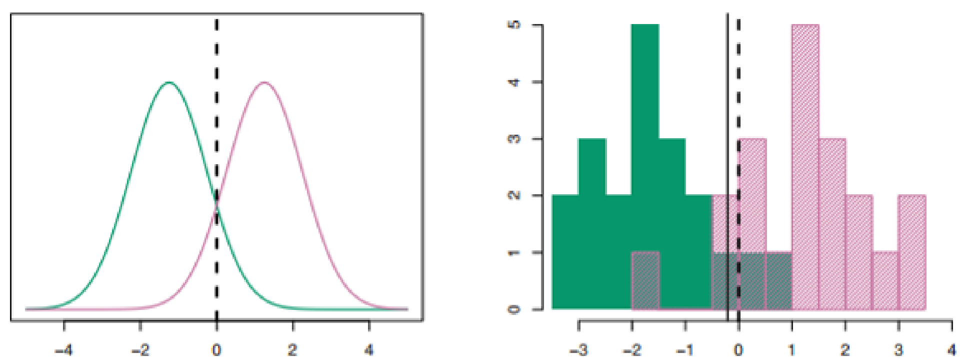


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

LDA uses Gaussian densities, i.e.,

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\Big\} \tag{1}$$

The LDA assumes $\Sigma_k = \Sigma$. It suffices to look at the log-ratio,

$$\begin{aligned}&\log\frac{P(G=k|X=x)}{P(G=l|X=x)} = \log\frac{f_k(x)}{f_l(x)} + \log\frac{\pi_k}{\pi_l},\\ &= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^\top\Sigma^{-1}(\mu_k-\mu_l) + (\mu_k-\mu_l)^\top\Sigma^{-1}x\end{aligned}$$

(2) Linear discriminant functions:

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

(3) Parameter Estimation:

$$\widehat{\pi}_k = N_k/N,\quad \widehat{\mu}_k = \sum_{g_i=k} x_i/N_k,\quad \widehat{\Sigma} = \sum_{k=1}^K\sum_{g_i=k}(x_i-\widehat{\mu}_k)(x_i-\widehat{\mu}_k)^\top/(N-K)$$

(4) 几何解释

考虑二分类问题，当 $\delta_1(x) > \delta_2(x)$ 时，将 x 判别为第一类。此时，

$$x^\top\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1 + \log\pi_1 \ge x^\top\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^\top\Sigma^{-1}\mu_2 + \log\pi_2,$$

等价于

$$x^\top\Sigma^{-1}(\mu_1-\mu_2) \ge \frac{1}{2}(\mu_1+\mu_2)^\top\Sigma^{-1}(\mu_1-\mu_2) - \log\frac{\pi_1}{\pi_2}.$$

注意到

$$\begin{aligned}&x^\top\Sigma^{-1}(\mu_1-\mu_2) = (\Sigma^{-1/2}x)^\top\{\Sigma^{-1/2}(\mu_1-\mu_2)\},\\ &\frac{1}{2}(\mu_1+\mu_2)^\top\Sigma^{-1}(\mu_1-\mu_2) = \Big\{\Sigma^{-1/2}\frac{1}{2}(\mu_1+\mu_2)\Big\}^\top\{\Sigma^{-1/2}(\mu_1-\mu_2)\},\end{aligned}$$

$\Sigma^{-1/2}x$ 将一般多维正态分布缩放旋转成了标准多维正态分布, 对应旋转后的两个类别中心为 $\Sigma^{-1/2}\mu_1$ 和 $\Sigma^{-1/2}\mu_2$，对应两个类别中心的中点为 $\frac{1}{2}\Sigma^{-1/2}(\mu_1+\mu_2)$.

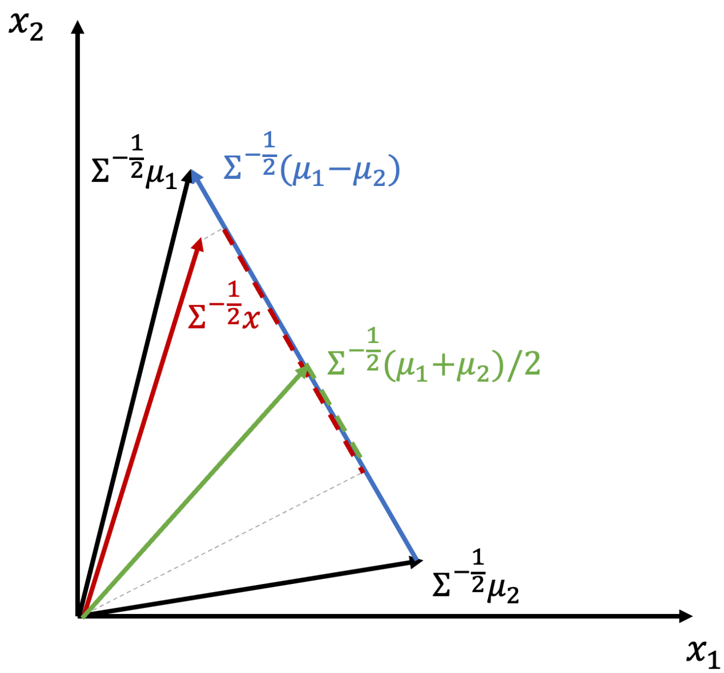
考虑 $(\Sigma^{-1/2}x)^\top\Sigma^{-1/2}(\mu_1-\mu_2)$:

这是变换后空间中 $\Sigma^{-1/2}x$ 和 $\Sigma^{-1/2}(\mu_1-\mu_2)$ 的内积，将其标准化，得到：

$$\frac{(\Sigma^{-1/2}x)^\top\Sigma^{-1/2}(\mu_1-\mu_2)}{\|\Sigma^{-1/2}(\mu_1-\mu_2)\|},$$

这个标准化的表达式衡量了 $\Sigma^{-1/2}x$ 在 $\Sigma^{-1/2}(\mu_1-\mu_2)$ 方向上的数量投影。这个投影可以用来衡量变换后的点 $\Sigma^{-1/2}x$ 相对于两个类别中心的位置，通过分析这个数量投影，我们可以判断观测点 x 在变换后空间中更接近哪个类别中心，从而为分类决策提供依据。

若 "$\geq$" 成立, 说明旋转后，x 比中点离第 1 类的类别中心更近, 也就是离第 2 类的类别中心更远。但是还要考虑一个因素, 如果第 1 类的个体比第 2 类的个体数目多, 就要加入一定的偏好, 比如这里的对数几率 $\log\frac{\pi_1}{\pi_2}$, 当 $\pi_1>\pi_2$ 时是正数, 就更容易判断为第 1 类.



2. Fisher 判别

Fisher 判别的核心思想将多元观测值 x 变换成一元观测值 y, 使得由类别 1 和类别 2 导出的 y 尽可能地分离开。可以用 x 的线性组合来建立 y, Fisher 判别并未假定

总体具有正态性，但是隐含有总体协方差矩阵 Σ 相等的假定。

(1) 分离度

假定 x 的一个固定线性组合对来自类别 1 的观测值来说其取值为 $y_{11}, y_{12}, \cdots, y_{1n_1}$，对来自类别 2 的观测值来说其取值为 $y_{21}, y_{22}, \cdots, y_{2n_2}$。这两组单变量数据之间的分离度用标准化后的 $\bar{y}_1$ 与 $\bar{y}_2$ 之间的差别来表示, 即

$$\text{分离度} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \quad \text{其中} s_y^2 = \frac{\sum_{j=1}^{n_1} \left(y_{1j} - \bar{y}_1\right)^2 + \sum_{j=1}^{n_2} \left(y_{2j} - \bar{y}_2\right)^2}{n_1 + n_2 - 2}$$

为方差的联合估计量. Fisher 判别量的目标是选择适当的 x 的线性组合, 使得样本均值 $\bar{y}_1$ 与 $\bar{y}_2$ 之间的分离度达到最大。

(2) 线性组合的选择

线性组合 $\widehat{y} = \widehat{a}^\top x = \left(\bar{x}_1 - \bar{x}_2\right)^\top \widehat{\Sigma}^{-1} x$ 对所有可能的线性系数向量 $\widehat{a}$ 使下述比值达到最大：

$$\frac{(y \text{ 的样本均值之间的距离平方})}{(y \text{ 的样本方差})} = \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_y^2} = \frac{\left(\widehat{a}^\top \bar{x}_1 - \widehat{a}^\top \bar{x}_2\right)^2}{\widehat{a}^\top \widehat{\Sigma} \widehat{a}} = \frac{\left\{\widehat{a}^\top \left(\bar{x}_1 - \bar{x}_2\right)\right\}^2}{\widehat{a}^\top \widehat{\Sigma} \widehat{a}}. \tag{3}$$

证明.

Lemma 1. *(**极大化引理**)* 令 $S \in \mathbb{R}^{p \times p}$ 为正定矩阵且 $x \in \mathbb{R}^p$ 为给定向量，则对任意非零向量 $a \in \mathbb{R}^p$ 有

$$\max_{a \neq 0} \frac{\left(a^\top x\right)^2}{a^\top S a} = x^\top S^{-1} x$$

对任意非零常数 c ，当 $a = cS^{-1}x$ 时，达到最大值。

由柯西-施瓦茨不等式: $(x^\top y)^2 \leq (x^\top x)(y^\top y)$, 对于正定矩阵 S 和向量 a,x 有 $(a^\top x)^2 = (S^{\frac{1}{2}} a)^\top (S^{-\frac{1}{2}} x) \leq (a^\top S a)(x^\top S^{-1} x)$。因为 $a \neq \mathbf{0}$ 且 S 为正定, $a^\top S a > 0$ 。不等式两边同时除以 $a^\top S a$，得出上界

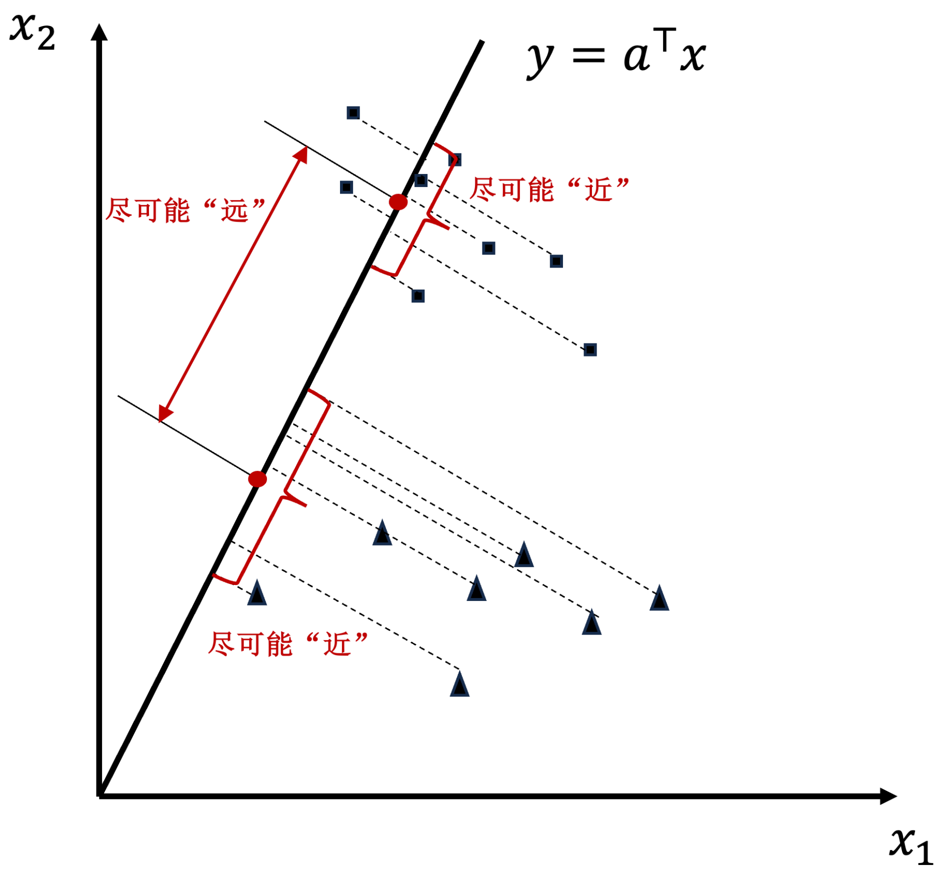
$$\frac{\left(a^\top x\right)^2}{a^\top S a} \leq x^\top S^{-1} x$$

由于 $a = cS^{-1}x$ 时达到上界, 因此, 通过此 a 取得最大值。

令 $x = \bar{x}_1 - \bar{x}_2$，$S = \widehat{\Sigma}$，有 $\widehat{a} = \widehat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2)$ 时，(3)取得最大值。 □

.

(3) 几何意义



(4) Fisher 判别

若

$$\widehat{y} = (\bar{x}_1 - \bar{x}_2)^\top \Sigma^{-1} x \geq \frac{1}{2} (\bar{x}_1 - \bar{x}_2)^\top \Sigma^{-1} (\bar{x}_1 + \bar{x}_2),$$

则将 x 分到第 1 类，否则分到第 2 类