机器学习导论

第2章 线性方法
2.3 线性判别分析

谢茂强

南开大学软件学院

线性判别分析（Linear Discriminate Analysis）

- Fisher 1936年提出，用来线性分类、降维
- 核心思想：使用超平面切分两类样本，使得两类样本类内距最小，类间距最大（不严谨）

Sir Ronald Aylmer Fisher

线性判别分析（Linear Discriminate Analysis）

- 给定训练样本集合，设法将其投影到分割面的法线上，在法线上使得同类样本的投影点尽可能接近，非同类样本投影尽可能远。
- 分类时，将待分类样本同样投影到该直线上，根据投影位置确定样本类别

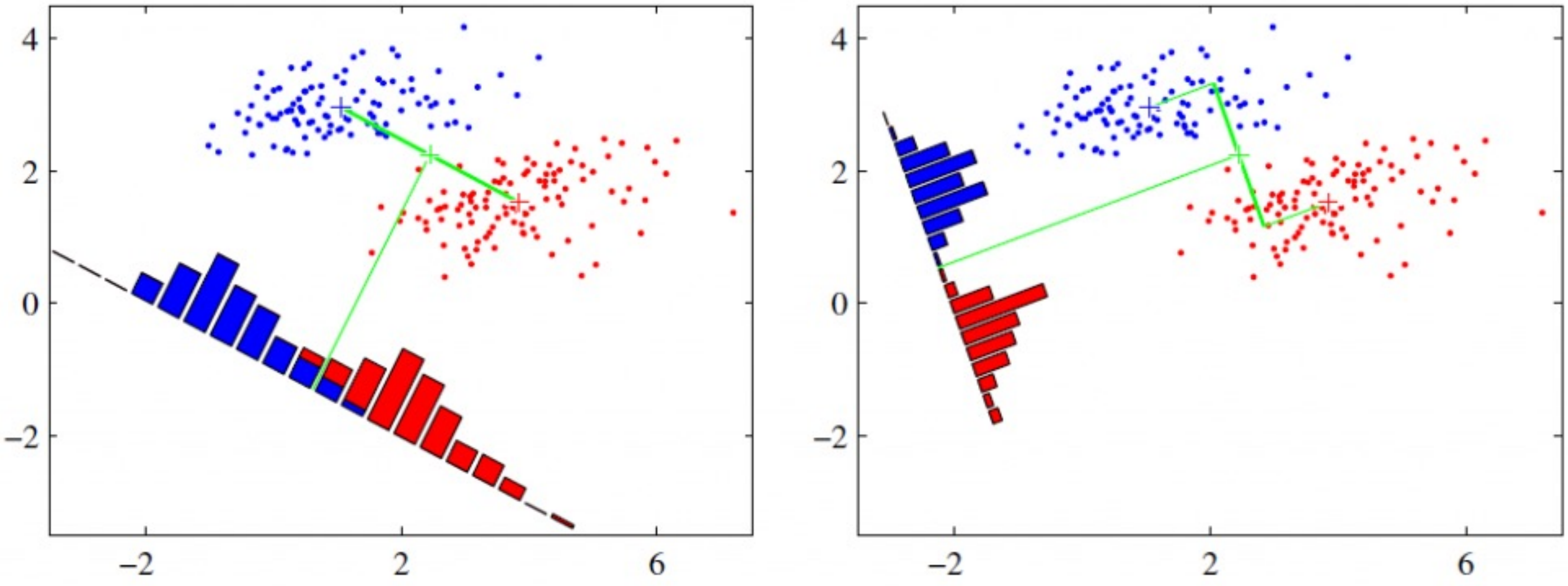


Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

图来自C. M. Bishop. Pattern Recognition and Machine Learning. 2006

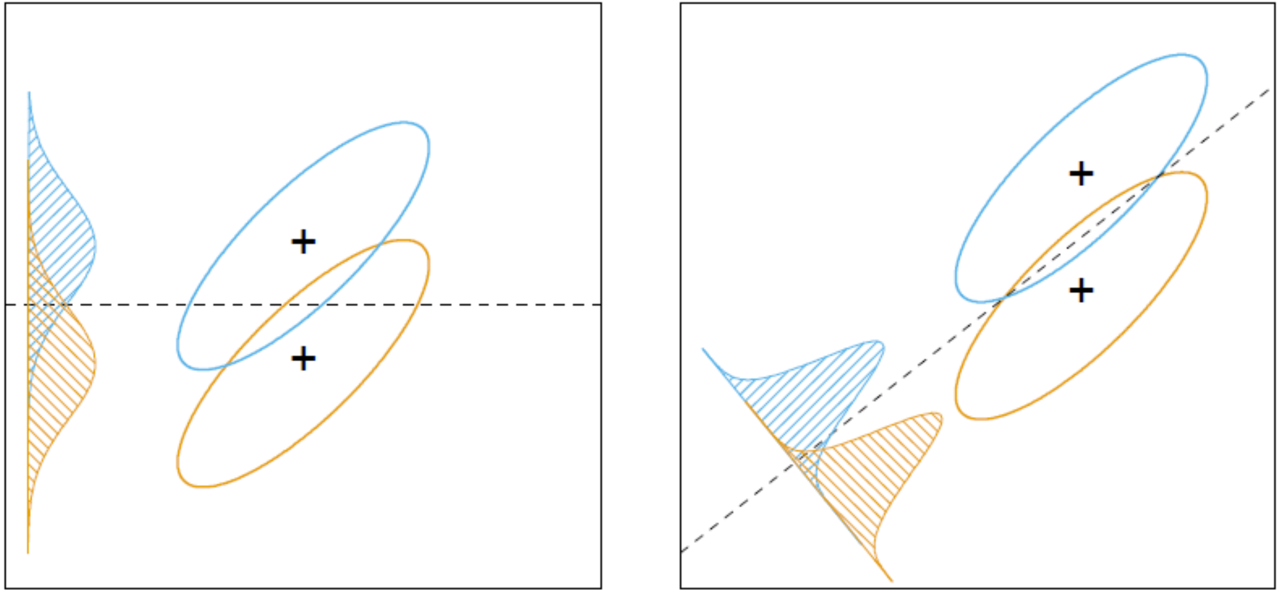


FIGURE 4.9. *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

图来自T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistic Learning: Data Mining, Inference, and Prediction. 2017

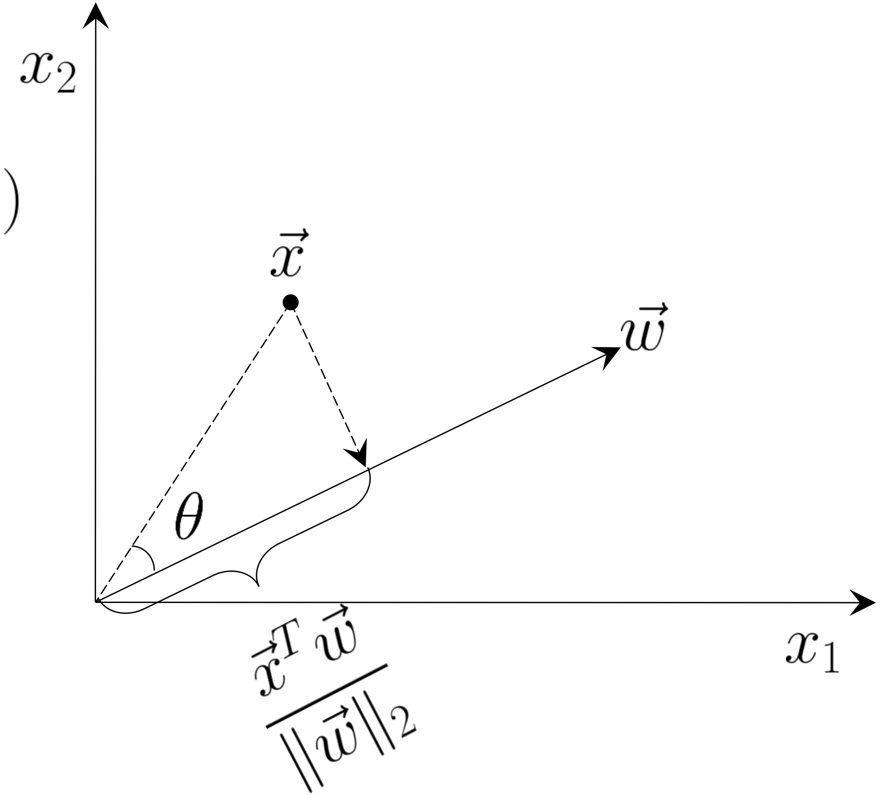
向量在另一向量上的投影

- 向量内积（点积）

$$\vec{x}^T\vec{w} = \|\vec{x}\|_2 \cdot \|\vec{w}\|_2 \cdot cos(\theta)$$

- 向量在另一向量上的投影

$$\|\vec{x}\|_2 \cdot cos(\theta) = \frac{\vec{x}^T\vec{w}}{\|\vec{w}\|_2}$$



- 给定训练样本集合$D=\{(\boldsymbol{x}_i,y_i)\}$ $i=1\ldots m$，$y_i=1$或0。

μ_0, μ_1 ：分别是两类样本的均值向量。

Σ_0, Σ_1 ：分别是两类的协方差矩阵。

$w^T\mu_0, w^T\mu_1$ ：分别是两类样本中心在w上的投影。

$w^T\Sigma_0 w, w^T\Sigma_1 w$ ：分别是两类样本中心在w上的投影长度的协方差。

LDA的代价函数（优化目标）

- 给定训练样本集合$D=\{(\boldsymbol{x}_i,y_i)\}$ $i=1\ldots m$，$y_i=1$或0。

最大化类间距： $\mathbf{max}\|w^T\mu_0 - w^T\mu_1\|_2^2$

最小化类内距： $\mathbf{min}\{w^T\Sigma_0 w + w^T\Sigma_1 w\}$

优化目标：

$$J = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w} = \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)w}{w^T(\Sigma_0+\Sigma_1)w}$$

优化目标： $$J = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w} = \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)w}{w^T(\Sigma_0+\Sigma_1)w} = \frac{w^T\mathbf{S}_b w}{w^T\mathbf{S}_w w}$$

类间散度矩阵(Between-class scatter matrix):

$$\mathbf{S}_b = (\mu_0-\mu_1)(\mu_0-\mu_1)^T$$

类内散度矩阵(Within-class scatter matrix):

$$\begin{aligned}\mathbf{S}_w &= \Sigma_0 + \Sigma_1 \\ &= \sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T + \sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T\end{aligned}$$

最大化： $J = \frac{w^T \mathbf{S}_b w}{w^T \mathbf{S}_w w}$

因为如果w是一个解的话，cw也是一个解，所以优化结果与w的长度无关，只与方向有关。不失一般性，令 $w^T \mathbf{S}_w w = 1$

等价优化目标：
$$\begin{aligned} \min_{w} \quad & -w^T \mathbf{S}_b w \\ s.t. \quad & w^T \mathbf{S}_w w = 1 \end{aligned}$$

定义拉格朗日函数：$L(w, \lambda) = -w^T \mathbf{S}_b w + \lambda (w^T \mathbf{S}_w w - 1)$

定义拉格朗日函数：$L(w,\lambda) = -w^T \mathbf{S}_b w + \lambda(w^T \mathbf{S}_w w - 1)$

求导：$$\frac{\partial L(w,\lambda)}{\partial w} = -\mathbf{S}_b w + \lambda \mathbf{S}_w w$$

令偏导得零可得：$\mathbf{S}_b w = \lambda \mathbf{S}_w w$ 这里的w没法直接求。

展开左式：$\mathbf{S}_b w = (\mu_0 - \mu_1)\underline{(\mu_0 - \mu_1)^T w} = \lambda \mathbf{S}_w w$

又因：$(\mu_0 - \mu_1)^T w$ 是个标量，设其为R

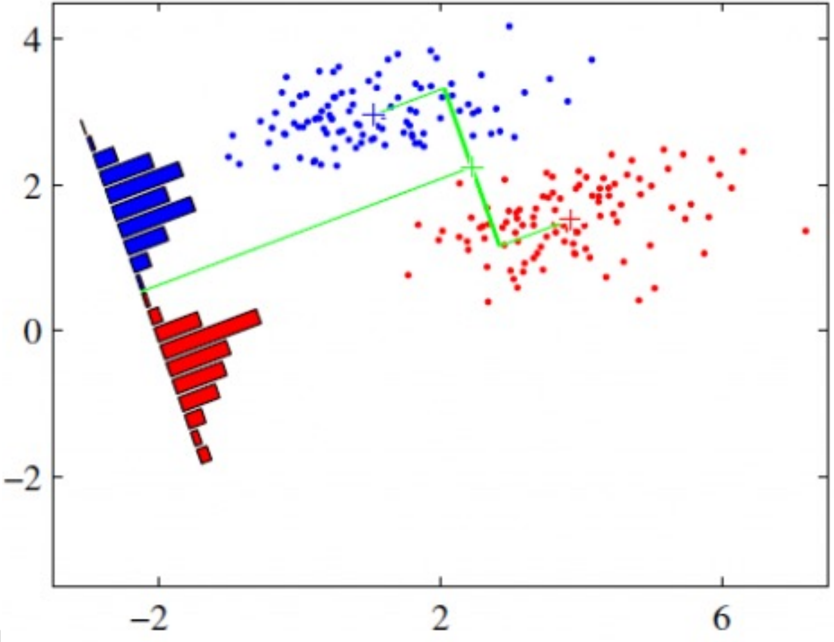
所以：$$w = \frac{R}{\lambda} \mathbf{S}_w^{-1} (\mu_0 - \mu_1)$$

最优方向： $w = \frac{R}{\lambda}\mathbf{S}_w^{-1}(\mu_0 - \mu_1)$

分割平面： $y = w^T x$

判别式：

$$y \lessgtr y_0 \quad \rightarrow \begin{cases} 0 \\ 1 \end{cases}$$



$$y_0 = \frac{1}{2} w^T(\mu_0 + \mu_1)$$

或 $$y_0 = \frac{m_0 w^T \mu_0 + m_1 w^T \mu_1}{m_0 + m_1}$$

LDA算法

1. 计算训练数据集中两类样本的均值(中心点) μ_0 、μ_1 方差 Σ_0和 Σ_1

2. 计算类内散度S_w;

3. 计算判决模型参数 $$\boldsymbol{w} = \mathbf{S}_w^{-1}(\mu_0 - \mu_1)$$

4. 计算判决模型阈值 $$y_0 = \frac{1}{2}\boldsymbol{w}^T(\mu_0 + \mu_1)$$

最后得判决模型：$y = \boldsymbol{w}^T\boldsymbol{x} \gtrless y_0 \rightarrow \begin{cases} 0 \\ 1 \end{cases}$

思考问题

1. 如何处理多类分类问题?

2. 如何度量两类样本集间差异?

   以Logistic回归和LDA为例

3. 判决模型的类别判别

   阈值 v.s. 大小

第2章 线性模型小测