4 Linear Models for Classification

分类目标：将输入变量x分到K个离散的类别 $\mathcal{C}_k$ 中的某一类.

输入空间被划分为不同的决策区域，它的边界被称为决策边界/决策面. 输出用one-hot方式编码.

第一章中提出的三种方法：

- 构造判别函数，直接把向量x分到具体的类别中.
- 推断阶段对条件概率分布 $p(\mathcal{C}_k \mid x)$ 直接建模，然后使用该概率分布进行最优决策：
  - 直接对条件概率分布建模，将条件概率分布表示为参数模型，用训练集优化.
  - 生成式方法，对类条件概率密度 $p(x \mid \mathcal{C}_k)$ 以及类先验概率分布 $p(\mathcal{C}_k)$ 建模，然后使用贝叶斯定理计算后验.

4.1 判别函数

输入 x，输出 $\mathcal{C}_k$.

4.1.1 二分类

线性判别函数：

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

$\boldsymbol{x} \in \mathbb{R}^D$，则 $y(\boldsymbol{x}) = 0$ 对应一个 $D-1$ 维的超平面.

- $\boldsymbol{w}$ 决定了决策面的方向：

$$\boldsymbol{w}^T (\boldsymbol{x}_A - \boldsymbol{x}_B) = 0$$

- w_0 决定了决策面的位置：

$$\frac{\boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|} = -\frac{w_0}{\|\boldsymbol{x}\|}$$

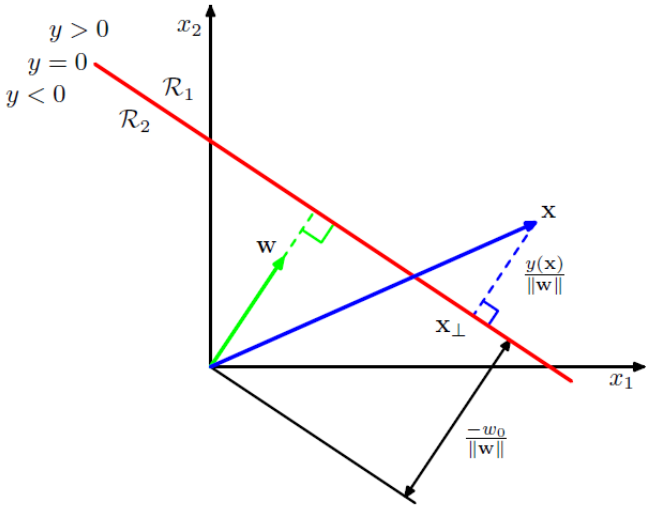

Figure 4.1 Illustration of the geometry of a linear discriminant function in two dimensions. The decision surface, shown in red, is perpendicular to $\mathbf{w}$, and its displacement from the origin is controlled by the bias parameter w_0. Also, the signed orthogonal distance of a general point $\mathbf{x}$ from the decision surface is given by $y(\mathbf{x})/\|\mathbf{w}\|$.

- 任意一点 $\boldsymbol{x}$ 到决策面的距离 r：

$$\boldsymbol{x} = \boldsymbol{x}_\perp + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

$$r = \frac{y(\boldsymbol{x})}{\|\boldsymbol{w}\|}$$

4.1.2 多分类

- 第一种方法：one-versus-one：$\frac{K(K-1)}{2}$ 个二分类器，但是仍然会造成无法分类的区域：

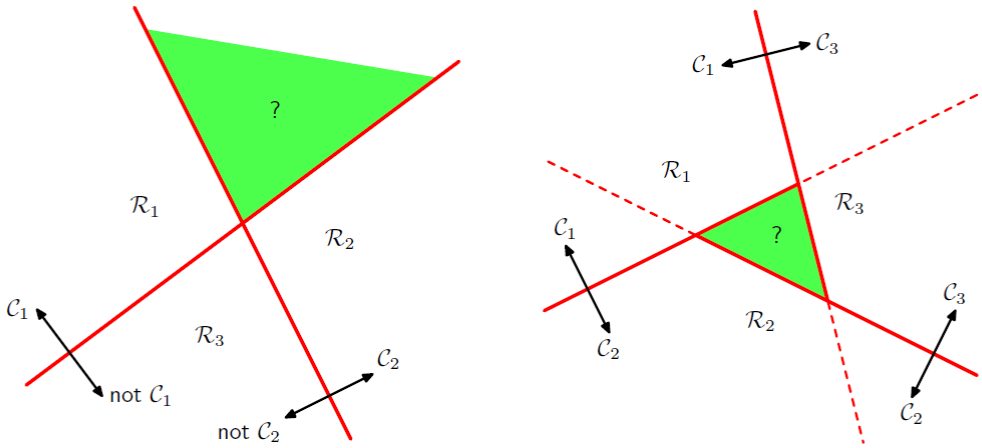

Figure 4.2 Attempting to construct a K class discriminant from a set of two class discriminants leads to ambiguous regions, shown in green. On the left is an example involving the use of two discriminants designed to distinguish points in class $\mathcal{C}_k$ from points not in class $\mathcal{C}_k$. On the right is an example involving three discriminant functions each of which is used to separate a pair of classes $\mathcal{C}_k$ and $\mathcal{C}_j$.

- 第二种方法：引入 K 类判别函数，K 个，判别属于 $\mathcal{C}_k$ 的概率.

  此时类别 $\mathcal{C}_k$ 和 $\mathcal{C}_j$ 之间的决策面为 $y_k(\boldsymbol{x}) = y_j(\boldsymbol{x})$，并且对应于一个 $(D-1)$ 维超平面，形式为：

$$(\boldsymbol{w}_k - \boldsymbol{w}_j)^T\boldsymbol{x} + (w_{k0} - w_{j0}) = 0$$

  - 此时决策区域是单连通的，并且是凸的.

    证明如下，$\boldsymbol{x}_A, \boldsymbol{x}_B$ 是位于决策区域 $\mathcal{R}_k$ 中任意两点，$0 \leq \lambda \leq 1$：

$$\widehat{\boldsymbol{x}} = \lambda\boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B$$

其中$0 \leq \lambda \leq 1$，根据判别函数的线性性质，有：

$$y_k(\widehat{\boldsymbol{x}}) = \lambda y_k(\boldsymbol{x}_A) + (1-\lambda) y_k(\boldsymbol{x}_B)$$

由于 x_A 和 x_B 位于 $\mathcal{R}_k$ 内部，因此对于所有 $j \neq k$，都有 $y_k(\boldsymbol{x}_A) > y_j(\boldsymbol{x}_A)$ 以及 $y_k(\boldsymbol{x}_B) > y_j(\boldsymbol{x}_B)$，因此 $y_k(\widehat{\boldsymbol{x}}) > y_j(\widehat{\boldsymbol{x}})$，从而 $\hat{x}$ 也位于 $\mathcal{R}_k$ 内部，即 $\mathcal{R}_k$ 是单连通的并且是凸的.

4.1.3 用于分类的最小平方方法

沿用上述第二种方法，每个类别 $\mathcal{C}_k$ 由自己的线性模型表述.

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$
$$\boldsymbol{y}(\boldsymbol{x}) = \tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{x}}$$

如上整理成带有 x 全一，并且带有 w_0 的全部矩阵表述，同理目标输出 t（注意这里目标输出就是**one-hot类型的，属于某个类的向量**）也整理成矩阵：

$$E_D(\tilde{\boldsymbol{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} - \boldsymbol{T})^T(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} - \boldsymbol{T})\right\}$$

$$\tilde{\boldsymbol{W}} = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{T} = \tilde{\boldsymbol{X}}^{\dagger}\boldsymbol{T}$$

其中 $\tilde{X}^{\dagger}$ 是矩阵 $\tilde{X}$ 的伪逆矩阵，这样得到了判别函数：

$$y(\boldsymbol{x}) = \tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{x}} = \boldsymbol{T}^T\left(\tilde{\boldsymbol{X}}^{\dagger}\right)^T\tilde{\boldsymbol{x}}$$

+ 最小平方方法缺陷

对于离群点缺少鲁棒性，与logistic回归模型比较：

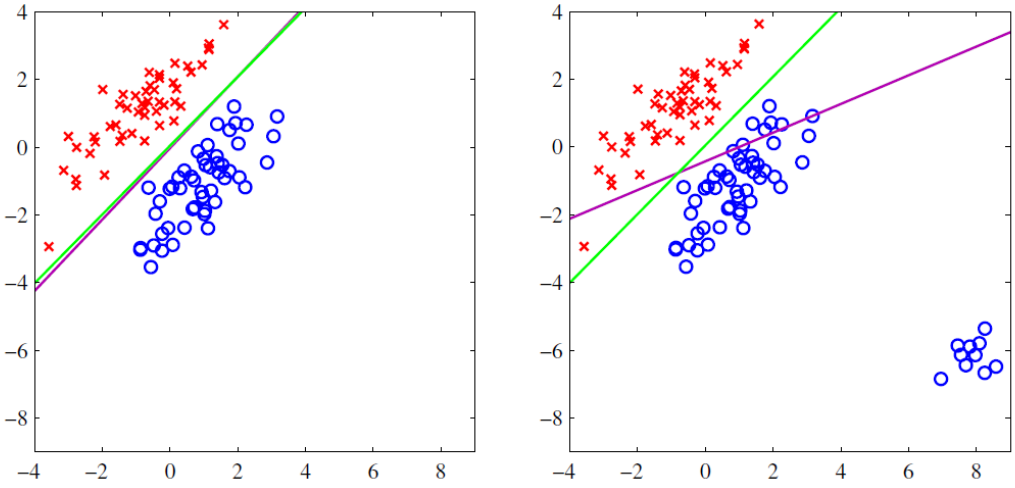

Figure 4.4 The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

而且，因为最小平方 对应于 高斯条件分布假设下的最大似然(在第1，2章有证明)，但是目标向量的概率分布不是高斯分布：

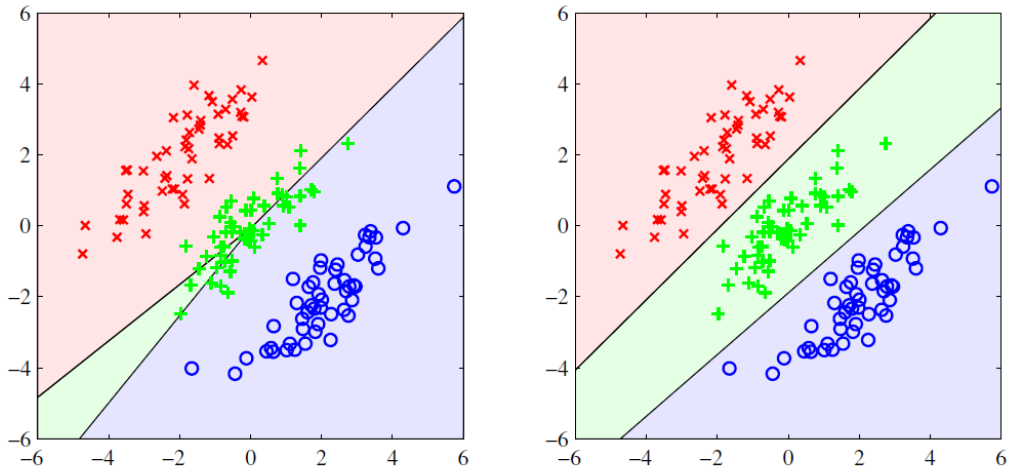

Figure 4.5 Example of a synthetic data set comprising three classes, with training data points denoted in red (×), green (+), and blue (○). Lines denote the decision boundaries, and the background colours denote the respective classes of the decision regions. On the left is the result of using a least-squares discriminant. We see that the region of input space assigned to the green class is too small and so most of the points from this class are misclassified. On the right is the result of using logistic regressions as described in Section 4.3.2 showing correct classification of the training data.

4.1.4 Fisher 线性判别分析

- 度量类别之间分开程度的方式(v1.0)：类别均值投影后的距离 $\boldsymbol{w}^T(\boldsymbol{m}_2-\boldsymbol{m}_1)$ 尽量大，限制 $\sum w_i^2=1$，则可以用拉格朗日乘数法求解.
- (v2.0) 仅如上分类会产生一个问题：

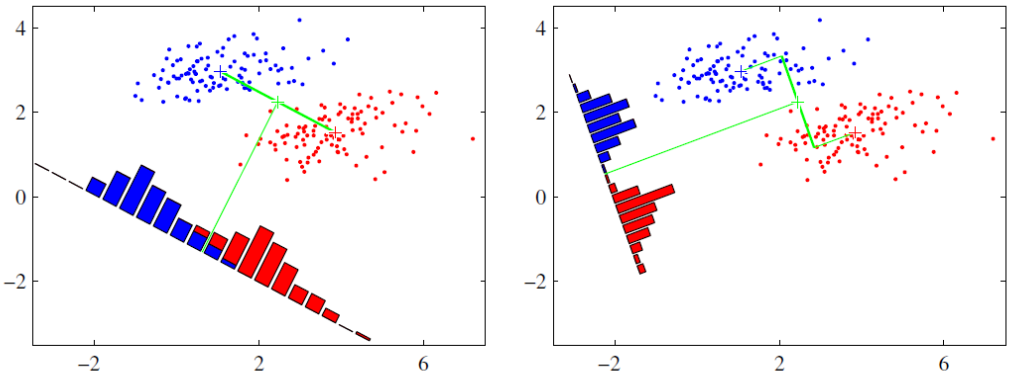

Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on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

所以引入 同时让类内方差最小，从而最小化类间的重叠.

$$J(\boldsymbol{w})=\frac{(m_2-m_1)^2}{s_1^2+s_2^2}$$

$$J(\boldsymbol{w})=\frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}$$

$$w\propto S_W^{-1}(m_2-m_1)$$

上式即 Fisher 线性判别函数. 到这里(式(4.30))，求解和西瓜书几乎一样呀，上式立即给出了 w 最优的投影方向.

4.1.5 与最小平方的关系

二分类问题，Fisher准则是最小平方的一个特例．要使用一种稍微不同的表达方法：

- $t_n = \frac{N}{N_1}$，如果该样本属于 $\mathcal{C}_1$．N_1 是第1类中样本数量．
- $t_n = -\frac{N}{N_2}$，如果该样本属于 $\mathcal{C}_2$．

$$E = \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n\right)^2$$

令E关于w_0和$\boldsymbol{w}$的导数等于零：

$$\sum_{n=1}^{N}\left(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n\right) = 0$$
$$\sum_{n=1}^{N}\left(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n\right)\boldsymbol{x}_n = 0$$

1．综合 t_n 的表达，得到偏置的表达式，$\boldsymbol{m}$ 为均值．

$$w_0 = -\boldsymbol{w}^T\boldsymbol{m}$$

2．同理，由上述第二个式子推出：

$$\left(S_W + \frac{N_1N_2}{N}S_B\right)w = N\left(m_1 - m_2\right)$$

上式移项，将 $\frac{N_1N_2}{N}S_Bw$ 移到等式对面，发现 S_Bw 总是在 $(m_2 - m_1)$ 的方向上：

$$\boldsymbol{w} \propto \boldsymbol{S}_W^{-1}\left(\boldsymbol{m}_2 - \boldsymbol{m}_1\right)$$

4.1.6 多分类的Fisher判别函数

推广，类内协方差矩阵：

$$S_W = \sum_{k=1}^{K} S_k$$
$$S_k = \sum_{n\in\mathcal{C}_k}\left(\boldsymbol{x}_n - \boldsymbol{m}_k\right)\left(\boldsymbol{x} - \boldsymbol{m}_k\right)^T$$
$$\boldsymbol{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\boldsymbol{x}_n$$

$$S_B = \sum_{k=1}^{K} N_k\left(m_k - m\right)\left(m_k - m\right)^T$$

最大化：

$$J(\boldsymbol{W}) = \mathrm{Tr}\left\{\left(\boldsymbol{W}^T\boldsymbol{S}_W\boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}^T\boldsymbol{S}_B\boldsymbol{W}\right)\right\}$$

注意，S_B 是K个类，K个矩阵的和，每个矩阵都是外积，所以秩最大为1，但是因为(4.44) 均值 $\boldsymbol{m}$ 的定义：$m = \frac{1}{N}\sum_{n=1}^{N}x_n = \frac{1}{N}\sum_{k=1}^{K}N_km_k$，所以最多只有 $K-1$ 个相互独立的，因此 S_B 的秩最大是 $K-1$．