机器学习—— Machine Learning
第三章 线性模型

中国地质大学 计算机学院
主讲：刘超

授课章节安排（主教材）

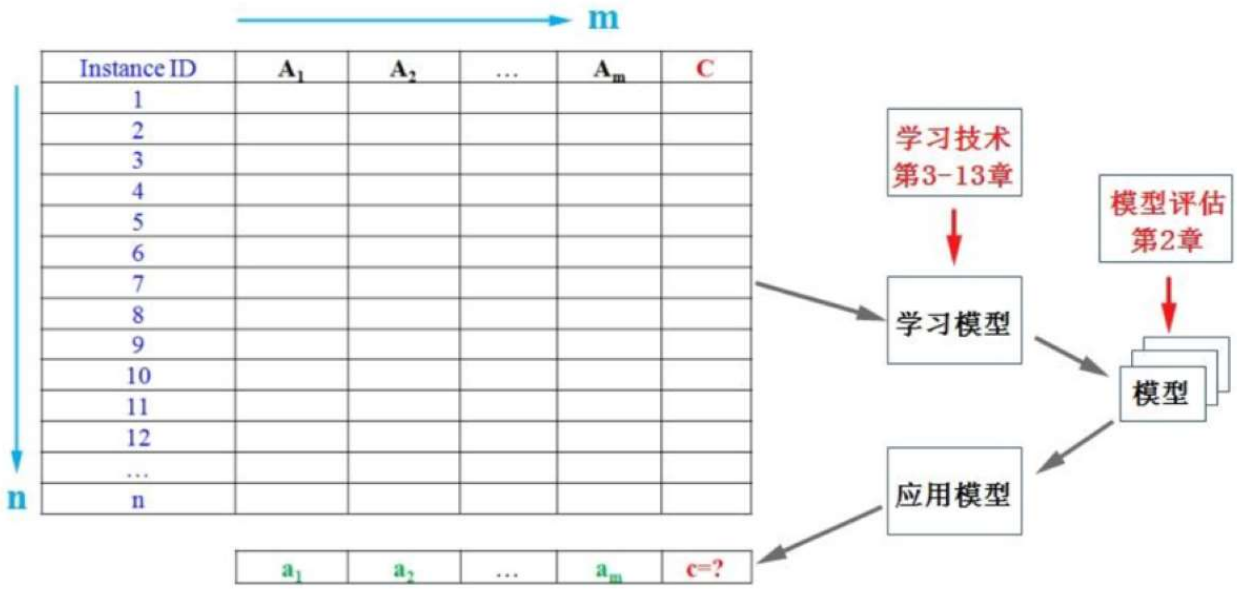
- 第1章：绪论
- 第2章：模型评估与选择
- 第3章：线性模型
- 第4章：决策树
- 第5章：神经网络
- 第6章：支持向量机
- 第7章：贝叶斯分类器
- 第8章：集成学习
- 第9章：聚类
- 第10章：降维与度量学习
- 第11章：特征选择与稀疏学习
- 第12章：计算学习理论
- 第13章：半监督学习
- 第14章：概率图模型
- 第15章：规则学习
- 第16章：强化学习

第三章　线性模型

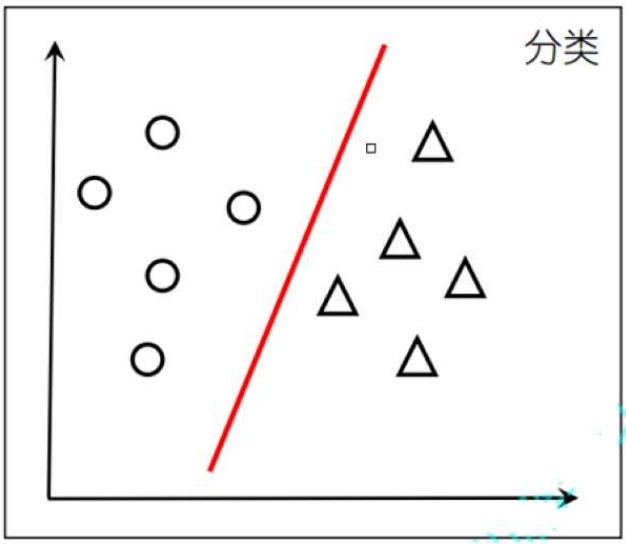
- 3.1 线性模型概念
- 3.2 线性回归
- 3.3 对数几率回归
- 3.4 线性判别分析
- 3.5 多分类学习
- 3.6 类别不平衡

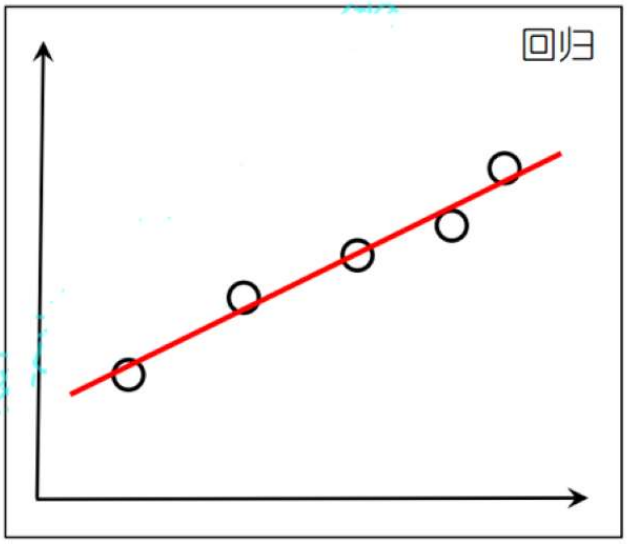
机器学习授课思路

分类：定义与过程



3.1线性模型概念





- 线性模型(linear model)包括：线性回归，二分类和多分类等。

3.2、线性回归

- **提及线性学习，我们首先会想到线性回归。回归跟分类的区别在于要预测的目标函数是连续值。**
- 给定由m个属性描述的样本$\boldsymbol{x}=(x_1;x_2;\ldots;x_m)$，其中$x_i$是$\boldsymbol{x}$在第$i$个属性上的取值，线性回归（linear regression）试图学得一个通过属性值的线性组合来进行预测的函数：

$$f(\boldsymbol{x})=w_1x_1+w_2x_2+\ldots+w_mx_m+b$$

- 一般用向量的形式写成：

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$$

其中$\boldsymbol{w}=(w_1;w_2;\ldots;w_m)$。

3.2、线性回归

- 给定训练数$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)\}$
  其中$\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{im})$，$y_i \in \mathbb{R}$

- 可用最小二乘法 (least square method) 对$\mathbf{w}$和b进行估计。

3.2、线性回归

- 下面以一元线性回归为例，来详细讲解 w 和 b 的最小二乘法估计

$$f(x_i) = wx_i + b, \text{ 使得 } f(x_i) \simeq y_i$$

- 最小二乘法就是基于预测值和真实值的均方差最小化的方法来估计参数 w 和 b：

$$\begin{aligned}(w^*, b^*) &= \arg\min_{(w,b)} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \\ &= \arg\min_{(w,b)} \sum_{i=1}^{n} (y_i - wx_i - b)^2\end{aligned}$$

3.2、线性回归

- 最小化均方误差

$$E_{(w,b)} = \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

- 分别对 w 和 b 求偏导，可得

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right), \tag{3.5}$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right), \tag{3.6}$$

推导公式(3.5)

已知 $E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$，所以

$$\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial w} &= \frac{\partial}{\partial w}\left[\sum_{i=1}^{m}(y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m}\frac{\partial}{\partial w}\left[(y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m}[2\cdot(y_i - wx_i - b)\cdot(-x_i)] \\
&= \sum_{i=1}^{m}\left[2\cdot\left(wx_i^2 - y_ix_i + bx_i\right)\right] \\
&= 2\cdot\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}y_ix_i + b\sum_{i=1}^{m}x_i\right) \\
&= 2\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right)
\end{aligned}$$

推导公式（3.6）

已知 $E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$，所以

$$\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial b} &= \frac{\partial}{\partial b}\left[\sum_{i=1}^{m} (y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m} \frac{\partial}{\partial b}\left[(y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m} [2 \cdot (y_i - wx_i - b) \cdot (-1)] \\
&= \sum_{i=1}^{m} [2 \cdot (b - y_i + wx_i)] \\
&= 2 \cdot \left[\sum_{i=1}^{m} b - \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} wx_i\right] \\
&= 2\left(mb - \sum_{i=1}^{m} (y_i - wx_i)\right)
\end{aligned}$$

3.2、线性回归

$$2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b)\, x_i\right), \tag{3.5}$$

$$2\left(mb - \sum_{i=1}^{m} (y_i - wx_i)\right), \tag{3.6}$$

令式(3.5)和(3.6)为零可得到 w 和 b 最优解的闭式(closed-form)解

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}, \tag{3.7}$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i), \tag{3.8}$$

其中 $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ 为 x 的均值.

推导公式(3.7)

$$0 = w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i$$

$$w\sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} y_i x_i - \sum_{i=1}^{m} b x_i$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - w x_i)$$

$\frac{1}{m}\sum_{i=1}^{m} y_i = \bar{y}$, $\frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}$, 则 $b = \bar{y} - w\bar{x}$

$$w\sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} y_i x_i - \sum_{i=1}^{m}(\bar{y} - w\bar{x})x_i$$

$$w\sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} y_i x_i - \bar{y}\sum_{i=1}^{m} x_i + w\bar{x}\sum_{i=1}^{m} x_i$$

$$w(\sum_{i=1}^{m} x_i^2 - \bar{x}\sum_{i=1}^{m} x_i) = \sum_{i=1}^{m} y_i x_i - \bar{y}\sum_{i=1}^{m} x_i$$

$$w = \frac{\sum_{i=1}^{m} y_i x_i - \bar{y}\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i^2 - \bar{x}\sum_{i=1}^{m} x_i}$$

$$\bar{y}\sum_{i=1}^{m} x_i = \frac{1}{m}\sum_{i=1}^{m} y_i \sum_{i=1}^{m} x_i = \bar{x}\sum_{i=1}^{m} y_i$$

$$\bar{x}\sum_{i=1}^{m} x_i = \frac{1}{m}\sum_{i=1}^{m} x_i \sum_{i=1}^{m} x_i = \frac{1}{m}(\sum_{i=1}^{m} x_i)^2$$

代入 可得公式 (3.7)

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum_{i=1}^{m} x_i)^2}$$

多元(multi-variate)线性回归

更一般的情形是如本节开头的数据集 D, 样本由 d 个属性描述. 此时我们试图学得

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b, \text{ 使得 } f(\boldsymbol{x}_i) \simeq y_i \ ,$$

这称为“多元线性回归”(multivariate linear regression).

类似的, 可利用最小二乘法来对 $\boldsymbol{w}$ 和 b 进行估计. 为便于讨论, 我们把 $\boldsymbol{w}$ 和 b 吸收入向量形式 $\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$, 相应的, 把数据集 D 表示为一个 $m \times (d+1)$ 大小的矩阵 $\mathbf{X}$, 其中每行对应于一个示例, 该行前 d 个元素对应于示例的 d 个属性值, 最后一个元素恒置为 1, 即

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix} ,$$

多元线性回归

- 同样采用最小二乘法求解，

$$\hat{\boldsymbol{w}}^* = \arg\min_{\hat{\boldsymbol{w}}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) \ . \tag{3.9}$$

令 $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$, 对 $\hat{\boldsymbol{w}}$ 求导得到

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\,\mathbf{X}^{\mathrm{T}} (\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) \ . \tag{3.10}$$

当 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 为满秩矩阵(full-rank matrix)或正定矩阵(positive definite matrix)时, 令式(3.10)为零可得

$$\hat{\boldsymbol{w}}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y} \ , \tag{3.11}$$

线性回归模型为

$$f(\hat{\boldsymbol{x}}_i) = \hat{\boldsymbol{x}}_i^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y} \ . \tag{3.12}$$

推导公式（3.10）

将 $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$ 展开可得

$$E_{\hat{\boldsymbol{w}}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \hat{\boldsymbol{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}}$$

对 $\hat{\boldsymbol{w}}$ 求导可得

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = \frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{\partial \hat{\boldsymbol{w}}} - \frac{\partial \boldsymbol{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}} - \frac{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\partial \hat{\boldsymbol{w}}} + \frac{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}}$$

由矩阵微分公式 $\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}, \frac{\partial \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\boldsymbol{x}$ 可得

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 0 - \mathbf{X}^{\mathrm{T}}\boldsymbol{y} - \mathbf{X}^{\mathrm{T}}\boldsymbol{y} + (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{w}}$$

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y})$$

对数线性回归"log-linear regression

线性模型虽简单, 却有丰富的变化. 例如对于样例 $(\boldsymbol{x}, y)$, $y \in \mathbb{R}$, 当我们希望线性模型(3.2) 的预测值逼近真实标记 y 时, 就得到了线性回归模型. 为便于观察, 我们把线性回归模型简写为

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \,. \tag{3.13}$$

可否令模型预测值逼近 y 的衍生物呢? 譬如说, 假设我们认为示例所对应的输出标记是在指数尺度上变化, 那就可将输出标记的对数作为线性模型逼近的目标, 即

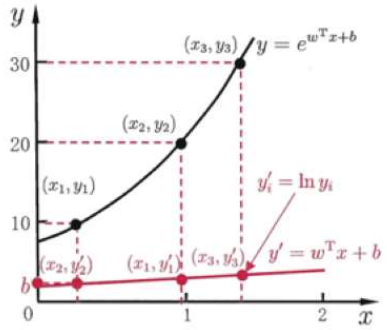


图 3.1 对数线性回归示意图

$$\ln y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \,. \tag{3.14}$$

广义(generalized)线性模型

- 只要学到w 和b，模型就可以确定；对于任意的测试样例$\boldsymbol{x}$，只要输入它的属性值，就可以输出它的预测值。
- 线性回归假定输入空间到输出空间的函数映射成线性关系，但现实应用中，很多问题都是非线性的。为拓展其应用场景，我们可以将线性回归的预测值做一个非线性的函数变化去逼近真实值，这样得到的模型统称为广义线性回归 (generalized linear regression)：

$$y = g^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b) \,, \tag{3.15}$$

  - g(.)单调可微的 联系函数 (link function)

3.3、对数几率回归

- 前面的内容都是在讲解如何利用线性模型进行回归学习，完成回归任务。但如果我们要做的是分类任务该怎么办？
- 为了简化，我们先考虑二分类任务，其输出标记 $y \in \{0,1\}$ ，但线性回归模型产生的预测值 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 是实值，因此，我们需将实值z转换为0/1值。最容易想到的联系函数 $g(\cdot)$ 当然是单位阶跃函数：

$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

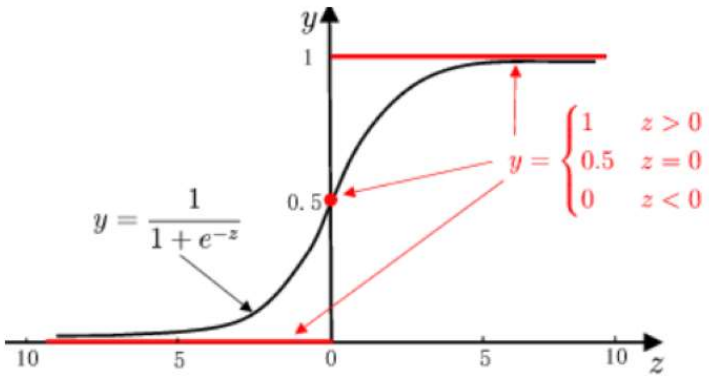
如预测值大于零就判为正例，小于零就判为反例，预测值为临界值零则可任意判别

3.3、对数几率回归

- 但单位阶跃函数不连续，因此不能直接用作联系函数 $g(\cdot)$。于是我们希望找到能在一定程度上近似单位阶跃函数的替代函数，并希望它在临界点连续且单调可微。逻辑斯蒂函数(logistic function) 正是这样一个常用的替代函数：

$$y = \frac{1}{1 + e^{-z}}$$

$$y = \frac{1}{1 + e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)}}$$



单位阶跃函数与对数几率函数的比较

3.3、对数几率回归

- 对数几率回归”（logistic regression），也称为逻辑斯蒂回归。
- 逻辑斯蒂(logistic function) 函数形似s，是Sigmoid函数的典型代表，它将z值转化为一个接近0或1的y值，并且其输出值在z=0附近变化很陡。
- 其对应的模型称为逻辑斯蒂回归(logistic regression)。需要特别说明的是，虽然它的名字是“回归”，但实际上却是一种分类学习方法。
- 逻辑斯蒂回归有很多优点：1）可以直接对分类可能性进行预测，将y视为样本x作为正例的概率；2）无需事先假设数据分布，这样就避免了假设分布不准确所带来的问题；3）是任意阶可导的凸函数，可直接应用现有数值优化算法求取最优解。

3.3、对数几率回归

- 将y视为样本x属于正例的概率 $p(y=1\mid \boldsymbol{x})$，根据逻辑斯蒂函数很容易得到：

$$y=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

$$p(y=1\mid \boldsymbol{x})=\frac{e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}$$

$$p(y=0\mid \boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}$$

- 给定训练数据集 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$，可通过“极大似然法”（maximum likelihood method)来估计ω和b，即最大化样本属于其真实标记的概率（对数似然）：

$$\ell(\boldsymbol{w},b)=\sum_{i=1}^{m}\ln p(y_i\mid \boldsymbol{x}_i;\boldsymbol{w},b)\,, \tag{3.25}$$

3.3、对数几率回归

即令每个样本属于其真实标记的概率越大越好. 为便于讨论, 令 $\boldsymbol{\beta} = (\boldsymbol{w}; b)$, $\hat{\boldsymbol{x}} = (\boldsymbol{x}; 1)$, 则 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 可简写为 $\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$. 再令 $p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) = p(y = 1 \mid \hat{\boldsymbol{x}}; \boldsymbol{\beta})$, $p_0(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) = p(y = 0 \mid \hat{\boldsymbol{x}}; \boldsymbol{\beta}) = 1 - p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta})$, 则式(3.25) 中的似然项可重写为

$$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) \ . \tag{3.26}$$

将式(3.26)代入(3.25), 并根据式(3.23)和(3.24)可知, 最大化式(3.25)等价于最小化

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i + \ln \left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i} \right) \right) \ . \tag{3.27}$$

- 高阶可导连续凸函数，可用经典的数值优化方法如梯度下降法/牛顿法求解.

式(3.27)是关于 $\boldsymbol{\beta}$ 的高阶可导连续凸函数, 根据凸优化理论 [Boyd and Vandenberghe, 2004], 经典的数值优化算法如梯度下降法(gradient descent method)、牛顿法(Newton method)等都可求得其最优解, 于是就得到

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \ . \tag{3.28}$$

以牛顿法为例, 其第 $t+1$ 轮迭代解的更新公式为

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^{\mathrm{T}}} \right)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \ , \tag{3.29}$$

其中关于 $\boldsymbol{\beta}$ 的一阶、二阶导数分别为

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i (y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})) \ , \tag{3.30}$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}} p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) (1 - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})) \ . \tag{3.31}$$

推导公式（3.27）

[推导]：将公式 (3.26) 代入公式 (3.25) 可得

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \ln\left(y_i p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) + (1-y_i)p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)$$

其中 $p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) = \dfrac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}, p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) = \dfrac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}$，代入上式可得

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{m} \ln\left(\frac{y_i e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i} + 1 - y_i}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right) \\ &= \sum_{i=1}^{m} \left(\ln(y_i e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i} + 1 - y_i) - \ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})\right)\end{aligned}$$

由于 y_i=0 或 1，则

$$\ell(\boldsymbol{\beta}) = \begin{cases} \sum_{i=1}^{m}(-\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})), & y_i = 0 \\ \sum_{i=1}^{m}(\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i - \ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})), & y_i = 1 \end{cases}$$

两式综合可得

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i - \ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})\right)$$

3.4 线性判别分析 (LDA)

- 线性判别分析(Linear Discriminant Analysis, LDA)是一种经典的判别学习方法。

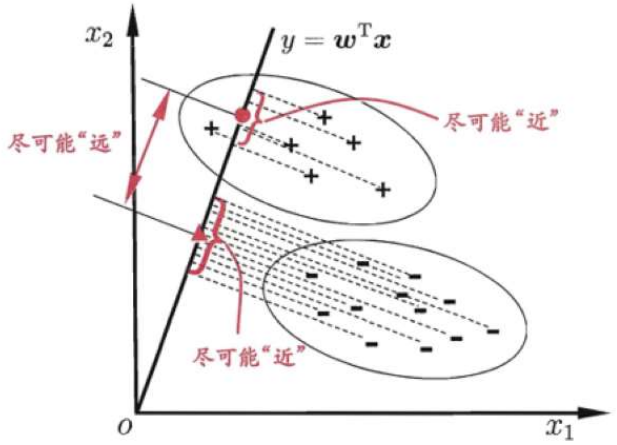
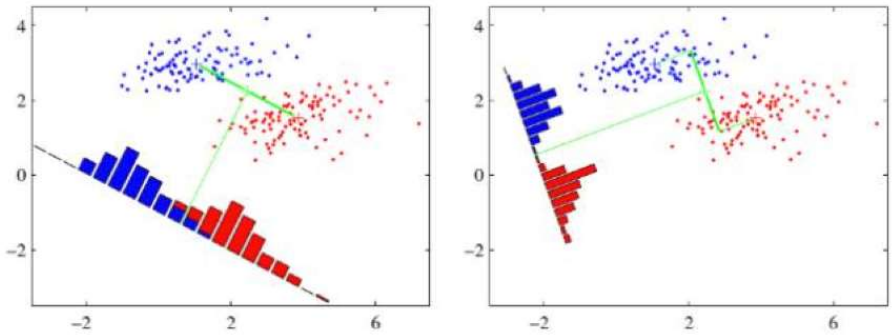


图 3.3 LDA 的二维示意图. "+"、"−"分别代表正例和反例, 椭圆表示数据簇的外轮廓, 虚线表示投影, 红色实心圆和实心三角形分别表示两类样本投影后的中心点.

3.4 线性判别分析 (LDA)—比较示例



- 从直观上可以看出，右图要比左图的投影效果好，因为右图的黑色数据和蓝色数据各个较为集中，且类别之间的距离明显。

3.4 线性判别分析 (LDA)

给定数据集 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, $y_i \in \{0,1\}$, 令 X_i、$\boldsymbol{\mu}_i$、$\boldsymbol{\Sigma}_i$ 分别表示第 $i \in \{0,1\}$ 类示例的集合、均值向量、协方差矩阵. 若将数据投影到直线 $\boldsymbol{w}$ 上, 则两类样本的中心在直线上的投影分别为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$; 若将所有样本点都投影到直线上, 则两类样本的协方差分别为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$.

欲使同类样例的投影点尽可能接近, 可以让同类样例投影点的协方差尽可能小, 即 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小; 而欲使异类样例的投影点尽可能远离, 可以让类中心之间的距离尽可能大, 即 $\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2$ 尽可能大. 同时考虑二者, 则可得到欲最大化的目标

$$
\begin{aligned}
J &= \frac{\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}} \\
&= \frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}}.
\end{aligned} \tag{3.32}
$$

投影点到原点的距离解释

设$\mu_0 = (x_0, y_0)$为样本中心点，LDA直线为$y = \omega^T x$，

投影直线的斜率与LDA直线垂直，斜率为$-\frac{1}{\omega^T}$，求得直线：

$$y - y_0 = -\frac{1}{\omega^T}(x - x_0)$$

直线与LDA直线的交点，即是样本中心在直线上的投影点：

$$\omega^T x = -\frac{1}{\omega^T}x + \frac{1}{\omega^T}x_0 + y_0$$

解得投影点：

$$x_\mu = \frac{x_0 + y_0\omega^T}{(\omega^T)^2 + 1}$$

$$y_\mu = \omega^T \cdot x_\mu$$

所以，距离为：

$$d = \sqrt{{x_\mu}^2 + {y_\mu}^2} = \sqrt{{x_\mu}^2 \cdot (1 + (\omega^T)^2)}$$

$$d = x_\mu \cdot \sqrt{(1 + (\omega^T)^2)} = \frac{x_0 + y_0\omega^T}{\sqrt{(\omega^T)^2 + 1}}$$

$$d = \frac{(1, \omega^T) \cdot (x_0, y_0)}{|(1, \omega^T)|} = \frac{\omega^T}{1} \cdot \mu_0 = \omega^T \mu_0$$

或者，直接用两个向量的夹角计算：

$$d = |\mu_0| \cdot \cos\theta = |\mu_0| \cdot \frac{\mu_0 \cdot \omega}{|\mu_0| \cdot |\omega|} = \omega^T \mu_0$$

推导公式(3.32)

$$\begin{aligned}
J &= \frac{\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}} \\
&= \frac{\|(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1)^{\mathrm{T}}\|_2^2}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}} \\
&= \frac{\|(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}\|_2^2}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}} \\
&= \frac{[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}]^{\mathrm{T}}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}} \\
&= \frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}}
\end{aligned}$$

LDA目标

定义“类内散度矩阵”(within-class scatter matrix)

$$\begin{aligned}\mathbf{S}_w &= \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 \\ &= \sum_{\boldsymbol{x}\in X_0} (\boldsymbol{x}-\boldsymbol{\mu}_0)(\boldsymbol{x}-\boldsymbol{\mu}_0)^{\mathrm{T}} + \sum_{\boldsymbol{x}\in X_1} (\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\end{aligned} \tag{3.33}$$

“类间散度矩阵”(between-class scatter matrix)

$$\mathbf{S}_b = (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\ , \tag{3.34}$$

- LDA 欲最大化的目标，即 S_b 与 S_w的“广义瑞利商”(generalized Rayleigh quotient). J值仅考虑w的方向。

$$J = \frac{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}}\ . \tag{3.35}$$

3.4 线性判别分析 (LDA)

不失一般性, 令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1$, 最大化广义瑞利商等价形式为

$$\begin{aligned}&\min_{\boldsymbol{w}} \quad -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w} \\ &\text{s.t.} \quad \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1\ .\end{aligned} \tag{3.36}$$

由拉格朗日乘子法，上式等价于，

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w}\ , \tag{3.37}$$

其中 λ 是拉格朗日乘子. 注意到 $\mathbf{S}_b\boldsymbol{w}$ 的方向恒为 $\boldsymbol{\mu}_0-\boldsymbol{\mu}_1$, 不妨令

$$\mathbf{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\ , \tag{3.38}$$

代入式(3.37)即得

$$\boldsymbol{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\ . \tag{3.39}$$

推导公式（3.37）

$$\min_{\boldsymbol{w}} \quad -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}$$
$$\text{s.t.} \quad \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1\ .$$

由公式 (3.36) 可得拉格朗日函数为

$$L(\boldsymbol{w},\lambda)=-\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}+\lambda(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}-1)$$

对 $\boldsymbol{w}$ 求偏导可得

$$\begin{aligned}\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}}&=-\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w})}{\partial \boldsymbol{w}}+\lambda\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}-1)}{\partial \boldsymbol{w}}\\&=-(\mathbf{S}_b+\mathbf{S}_b^{\mathrm{T}})\boldsymbol{w}+\lambda(\mathbf{S}_w+\mathbf{S}_w^{\mathrm{T}})\boldsymbol{w}\end{aligned}$$

由于 $\mathbf{S}_b=\mathbf{S}_b^{\mathrm{T}},\mathbf{S}_w=\mathbf{S}_w^{\mathrm{T}}$，所以

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}}=-2\mathbf{S}_b\boldsymbol{w}+2\lambda\mathbf{S}_w\boldsymbol{w}$$

令上式等于 0 即可得

$$-2\mathbf{S}_b\boldsymbol{w}+2\lambda\mathbf{S}_w\boldsymbol{w}=0$$

$$\mathbf{S}_b\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$$

推导公式（3.38-3.39）

$$\mathbf{S}_b\boldsymbol{w}=(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}$$

令 λ 恒等于 $(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}$，

$$\mathbf{S}_b\boldsymbol{w}=\lambda(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$$

将其代入 $\mathbf{S}_b\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$ 即可解得

$$\boldsymbol{w}=\mathbf{S}_w^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$$

推广到多类

假定有 N 个类

- 全局散度矩阵 $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$
- 类内散度矩阵 $\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}$ $\mathbf{S}_{w_i} = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$
- 类间散度矩阵 $\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$

多分类LDA有多种实现方法：采用 $\mathbf{S}_b$, $\mathbf{S}_w$, $\mathbf{S}_t$ 中的任何两个

例如，$\max_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W})}{\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W})} \Rightarrow \mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W}$

$\mathbf{W} \in \mathbb{R}^{d \times (N-1)}$

$\mathbf{W}$ 的闭式解是 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的 N-1 个最大广义特征值所对应的特征向量组成的矩阵

3.5、多分类学习

- 前面讲到的都是二分类学习任务，现实应用中常常会遇到多分类学习任务。
- 多分类学习方法
  - 二分类学习方法推广到多类
  - 利用二分类学习器解决多分类问题（常用）
    - 对问题进行拆分，为拆出的每个二分类任务训练一个分类器
    - 对每个分类器的预测结果进行集成以获得最终的多分类结果
- 拆分策略
  - 一对一（One vs. One, OvO）
  - 一对其余（One vs. Rest, OvR）
  - 多对多（Many vs. Many, MvM）

3.5、多分类学习

- 一对一拆分：

拆分阶段

- N个类别两两配对
  - N(N-1)/2 个二类任务
- 各个二类任务学习分类器
  - N(N-1)/2 个二类分类器

测试阶段

- 新样本提交给所有分类器预测
  - N(N-1)/2 个分类结果
- 投票产生最终分类结果
  - 被预测最多的类别为最终类别

3.5、多分类学习

- 一对其余拆分：

拆分阶段

- ✓ 某一类作为正例，其余类作为反例
  - ◆N 个二类任务
- ✓ 各个二类任务学习分类器
  - ◆N 个二类分类器

测试阶段

- ✓ 新样本提交给所有分类器预测
  - ◆N 个分类结果
- ✓ 比较各分类器的预测置信度
  - ◆仅有一个分类器预测为正类，则对应的类别标记作为最终分类结果；若有多个分类器预测为正类，选择置信度最大类别作为最终类别

3.5、多分类学习

- 拆解法：将一个多分类任务拆分为若干个二分类任务求解

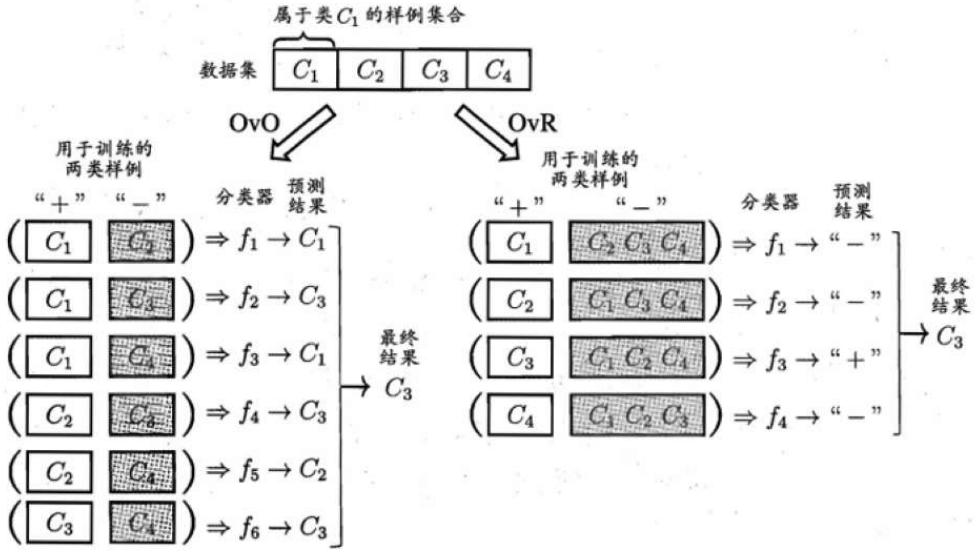


图 3.4 OvO 与 OvR 示意图

- 预测性能取决于具体数据分布，多数情况下两者差不多

3.5、多分类学习

一对一

- 训练N(N-1)/2个分类器，存储开销和测试时间大
- 训练只用两个类的样例，训练时间短

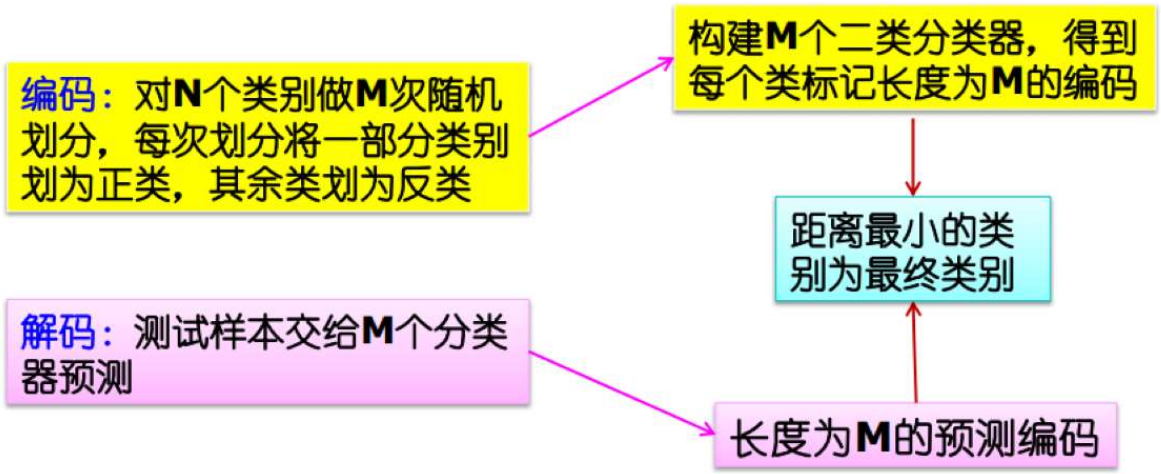
一对其余

- 训练N个分类器，存储开销和测试时间小
- 训练用到全部训练样例，训练时间长

预测性能取决于具体数据分布，多数情况下两者差不多

3.5、多分类学习

- 多对多（Many vs Many, MvM）
  - ✓ 若干类作为正类，若干个其他类作为反类
  - ✓ 纠错输出码（**Error Correcting Output Code, ECOC**）



3.5、多分类学习

- 纠错输出码：二元码和三元码

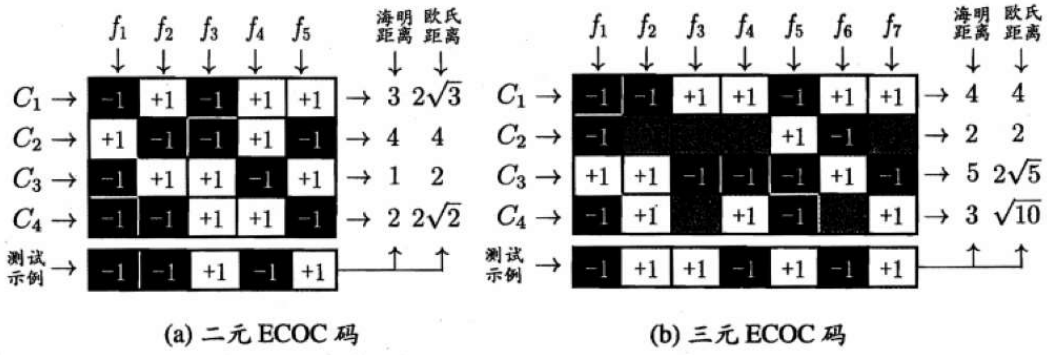


图 3.5 ECOC 编码示意图. “+1”、“−1”分别表示学习器 f_i 将该类样本作为正、反例; 三元码中“0”表示 f_i 不使用该类样本

- 对分类器错误有一定容忍和修正能力，编码越长、纠错能力越强
- 对同等长度的编码，理论上来说，任意两个类别之间的编码距离越远，则纠错能力越强

3.6、类别不平衡 (class-imbalance)

- 不同类别的样本比例相差很大；“小类”往往更重要

基本思路：(3.46)反映正例可能性与反例可能性之比值

若 $\frac{y}{1-y} > 1$ 则 预测为正例. (3.46)

若 $\frac{y}{1-y} > \frac{m^+}{m^-}$ 则 预测为正例. (3.47)

- 基本策略：再缩放(rescaling)

$$\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+} . \quad (3.48)$$

- 常见类别不平衡学习方法：
  - 过采样 (oversampling) 如： SMOTE
  - 欠采样 (undersampling)如：EasyEnsemble
  - 阈值移动 (threshold-moving)