电子科技大学
University of Electronic Science and Technology of China

《机器学习》
第三章 线性模型

任亚洲
yazhou.ren@uestc.edu.cn
英才学院

目录

- 线性回归
  - 最小二乘法
- 二分类任务
  - 对数几率回归
  - 线性判别分析
- 多分类任务
  - 一对一
  - 一对其余
  - 多对多
- 类别不平衡问题

基本形式

- **线性模型一般形式**

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

$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$ **是由属性描述的示例，其中** x_i **是** $\boldsymbol{x}$ **在第** i **个属性上的取值**

- **向量形式**

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

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

线性模型优点

- 形式简单、易于建模
- 可解释性
- 非线性模型的基础
  - 引入层级结构或高维映射

- 一个例子
  - 综合考虑色泽、根蒂和敲声来判断西瓜好不好
  - 其中根蒂的系数最大，表明根蒂最要紧；而敲声的系数比色泽大，说明敲声比色泽更重要

$$f_{\text{好瓜}}(\boldsymbol{x}) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.3 \cdot x_{\text{敲声}} + 1$$

线性回归

- 给定数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

  其中 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ ，$y_i \in \mathbb{R}$
- 线性回归（linear regression）目的
  - 学得一个线性模型以尽可能准确地预测实值输出标记
- 离散属性处理
  - 有“序”关系
    - 连续化为连续值：{高，中，低}-> {1，0.5，0}
  - 无“序”关系
    - 有k个属性值，则转换为k维向量
    - {西瓜,南瓜,黄瓜} -> {(0,0,1), (0,1,0), (1,0,0)}

线性回归

- **单一属性的线性回归目标**

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

- **参数/模型估计：最小二乘法（least square method）**

$$\begin{aligned}(w^*, b^*) &= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} \left(f(x_i) - y_i\right)^2 \\ &= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} \left(y_i - wx_i - b\right)^2\end{aligned}$$

线性回归 – 最小二乘法

- **最小化均方误差**

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

- **分别对** w **和** b **求导，可得**

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

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

线性回归 – 最小二乘法

$$\frac{\partial E}{\partial \omega} = \sum_{i=1}^{m}[2(y_i - \omega x_i - b)\cdot(-x_i)] = 2\left(\omega \sum x_i^2 - \sum (y_i - b)x_i\right) = 0$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{m}[2(y_i - \omega x_i - b)\cdot(-1)] = 2\left(mb - \sum (y_i - \omega x_i)\right) = 0$$

$$\omega \sum_i x_i^2 - \sum_i \left(y_i - \frac{1}{m}\sum (y_j - \omega x_j)\right) x_i = 0$$

$$\omega \sum_i x_i^2 - \sum_i x_i \left(y_i - \frac{1}{m}\sum y_j + \frac{\omega}{m}\sum x_j\right) = 0$$

$$\omega \sum_i x_i^2 - \sum_i x_i y_i + \frac{1}{m}\sum_i x_i \cdot \sum_j y_j - \frac{\omega}{m}\sum_i x_i \cdot \sum_j x_j = 0$$

$$\omega\left(\sum_i x_i^2 - \frac{1}{m}\left(\sum_i x_i\right)^2\right) = \sum_i y_i(x_i - \bar{x})$$

线性回归 – 最小二乘法

- 得到闭式（closed-form）解

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

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

其中， $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$

多元线性回归

- **给定数据集**

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

$$\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id}) \qquad y_i \in \mathbb{R}$$

- **多元线性回归目标**

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

多元线性回归

- **把** $\boldsymbol{w}$ **和** b **吸收入向量形式** $\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$ **，数据集表示为**

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

$$\boldsymbol{y} = (y_1; y_2; \ldots; y_m)$$

多元线性回归 – 最小二乘法

- **最小二乘法（least square method）**

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

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**

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

令上式为零可得 $\hat{\boldsymbol{w}}$ **最优解的闭式解**

多元线性回归－满秩讨论

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 是满秩矩阵或正定矩阵，则

$$\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

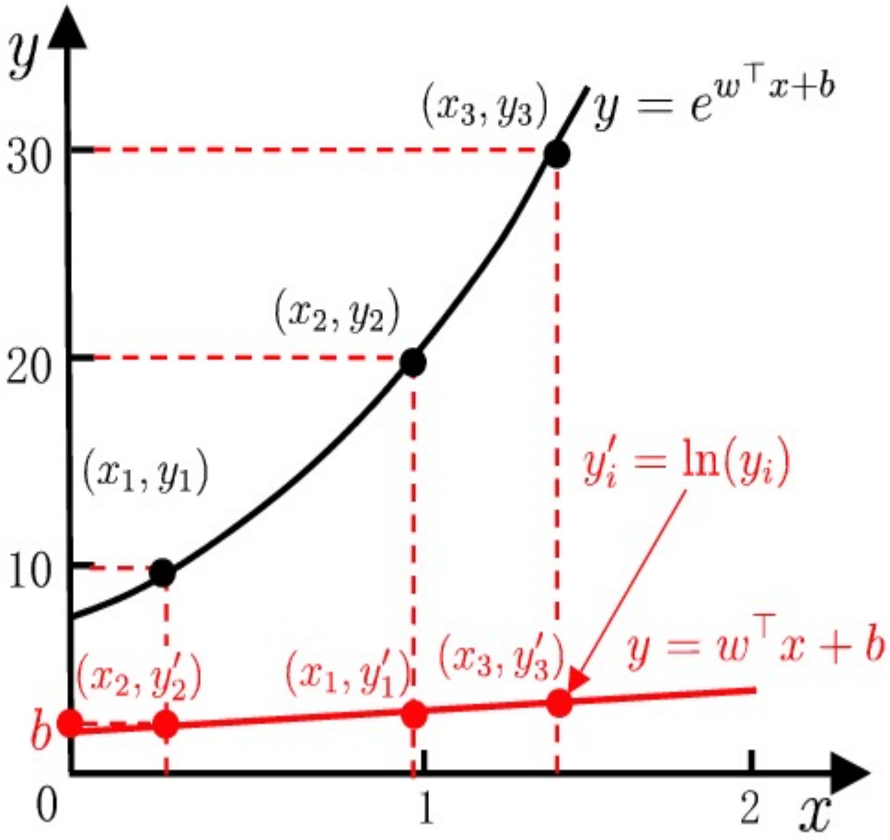
其中 $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ 是 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的逆矩阵，线性回归模型为

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

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 不是满秩矩阵
  - 根据归纳偏好选择解（参见1.4节）
  - 引入正则化（参见6.4节，11.4节）

对数线性回归

- **若样本输出标记是在指数尺度上变化，则将输出标记的对数为线性模型逼近的目标**



$$\ln y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$\Updownarrow$$

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

线性回归 – 广义线性模型

- **一般形式**

$$y = g^{-1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right)$$

□ $g(\cdot)$ **称为联系函数（link function）**

– **单调可微函数**

- **对数线性回归是** $g(\cdot) = \ln(\cdot)$ **时广义线性模型的特例**

二分类任务

- **预测值与输出标记**

$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \qquad y \in \{0, 1\}$$

- **寻找函数将分类标记与线性回归模型输出联系起来**
- **最理想的函数——单位阶跃函数**

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

  - **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**

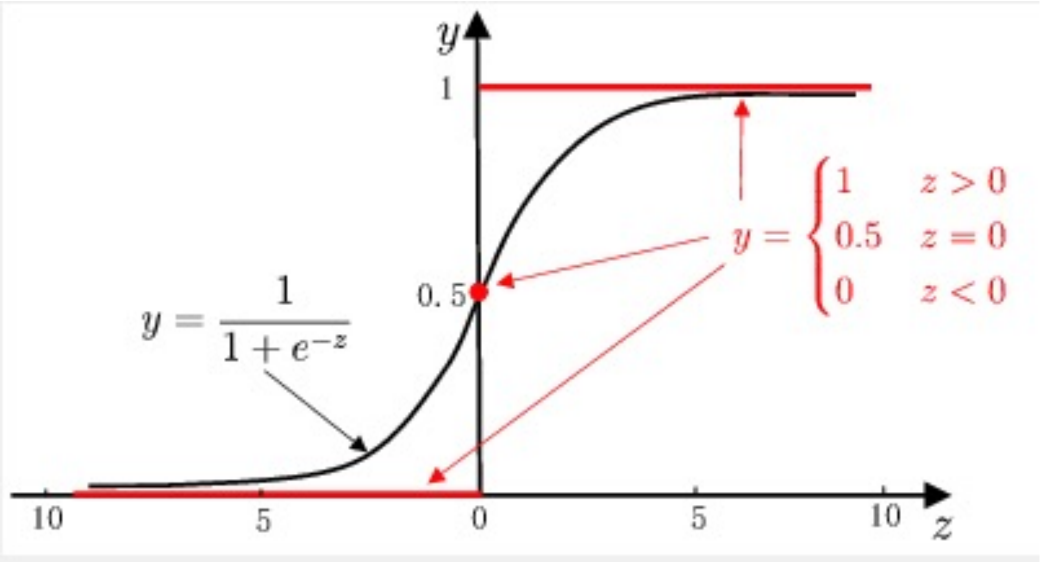
二分类任务

- **单位阶跃函数缺点**
  - 不连续
- **替代函数——对数几率函数（logistic/sigmoid function）**
  - 单调可微、任意阶可导

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

$$\begin{aligned} y' &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= y(1-y)' \end{aligned}$$

单位阶跃函数与对数几率函数的比较



对数几率回归-逻辑回归

- **运用对数几率函数**

$$y=\frac{1}{1+e^{-z}} \quad \text{变为} \quad y=\frac{1}{1+e^{-(w^{\mathrm{T}}x+b)}}$$

- **对数几率（log odds）**
  - **样本作为正例的相对可能性的对数**

$$\ln\frac{y}{1-y}$$

- **对数几率回归优点**
  - **无需事先假设数据分布**
  - **可得到“类别”的近似概率预测**
  - **可直接应用现有数值优化算法求取最优解**

对数几率回归 – 极大似然法

- **对数几率**

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

显然有

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

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

对数几率回归 – 极大似然法

- 极大似然法（maximum likelihood）
  - 给定数据集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$
  - 最大化样本属于其真实标记的概率
    - 最大化对数似然函数

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

对数几率回归 – 极大似然法

- **转化为最小化负对数似然函数求解**
  - 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 $$p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=p(y=1\mid\hat{\boldsymbol{x}};\boldsymbol{\beta})$$
  $$p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=p(y=0\mid\hat{\boldsymbol{x}};\boldsymbol{\beta})=1-p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})$$
  - 则似然项可重写为
  $$p(y_i\mid\boldsymbol{x}_i;\boldsymbol{w}_i,b)=y_ip_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})+(1-y_i)p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})$$
  $$\ln p= y_i\ln p_1+(1-y_i)\ln p_0$$
  - 故等价形式为要最小化

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

$$y_i\cdot\ln\frac{e^{\beta^T\widehat{x_i}}}{1+e^{\beta^T\widehat{x_i}}}+(1-y_i)\cdot\ln\frac{1}{e^{\beta^T\widehat{x_i}}}$$
$$y_i(\ln e^{\beta^T\widehat{x}_i}-\ln(1+e^{\beta^T\widehat{x}_i}))+(1-y_i)(-\ln(1+e^{\beta^T\widehat{x}_i}))$$
$$y_i\beta^T\widehat{x}_i-\ln(1+e^{\beta^T\widehat{x}_i})$$

对数几率回归

- **求解得**

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

- **牛顿法第t+1轮迭代解的更新公式**

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

- **其中关于 β 的一阶、二阶导数分别为**

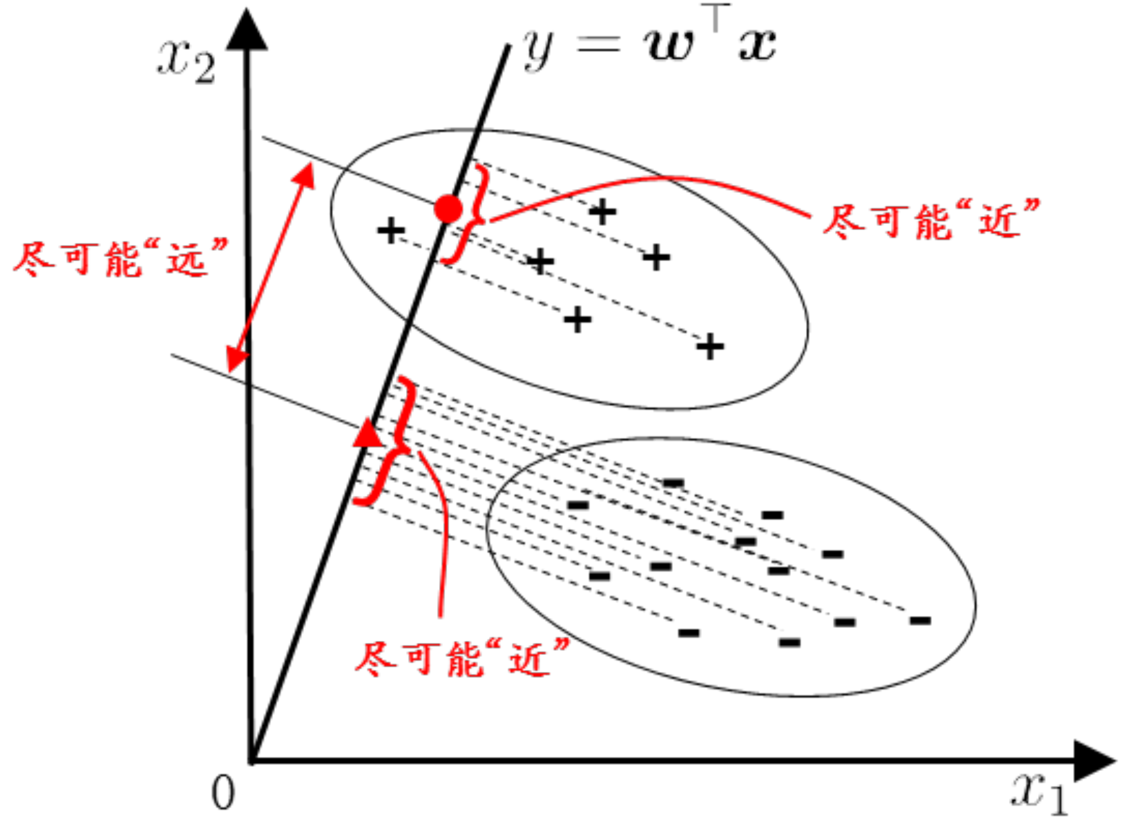
$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \left(y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})\right)$$

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

高阶可导连续凸函数，梯度下降法/牛顿法 [Boyd and Vandenberghe, 2004]

二分类任务 – 线性判别分析

- 线性判别分析（Linear Discriminant Analysis）[Fisher, 1936]
  - 将训练样本投影到一条直线上，使同类相近，异类相远
  - 预测时，对预测样本的投影点进行识别



LDA也可被视为一种
监督降维技术

二分类任务－ 线性判别分析

- **期望：** $\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$ $\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$
- **协方差矩阵**

$$\Sigma = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}]) (\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathrm{T}} \right]$$

$$= \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

- **其中第 (i,j)个元素是第*i*维数据X_i与第*j*维数据X_j的协方差**
- **第 (i,i)个元素是第*i*维数据X_i的方差**

二分类任务 – 线性判别分析

- **协方差矩阵**

$\Sigma = \mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{X}-\mathrm{E}[\mathbf{X}])^{\top}\right]$ 与$\mu = \mathrm{E}(\mathbf{X})$ 满足下边的基本性质:

1. $\Sigma = \mathrm{E}(\mathbf{X}\mathbf{X}^{\top}) - \mu\mu^{\top}$
2. Σ是半正定的和对称的矩阵。
3. $\mathrm{var}(\mathbf{a}^{\top}\mathbf{X}) = \mathbf{a}^{\top}\,\mathrm{var}(\mathbf{X})\mathbf{a}$
4. $\mathbf{\Sigma} \geq 0$
5. $\mathrm{var}(\mathbf{A}\mathbf{X}+\mathbf{a}) = \mathbf{A}\,\mathrm{var}(\mathbf{X})\mathbf{A}^{\top}$
6. $\mathrm{cov}(\mathbf{X},\mathbf{Y}) = \mathrm{cov}(\mathbf{Y},\mathbf{X})^{\top}$
7. $\mathrm{cov}(\mathbf{X_1}+\mathbf{X_2},\mathbf{Y}) = \mathrm{cov}(\mathbf{X_1},\mathbf{Y}) + \mathrm{cov}(\mathbf{X_2},\mathbf{Y})$
8. 若 $p = q$，则有$\mathrm{cov}(\mathbf{X}+\mathbf{Y}) = \mathrm{var}(\mathbf{X}) + \mathrm{cov}(\mathbf{X},\mathbf{Y}) + \mathrm{cov}(\mathbf{Y},\mathbf{X}) + \mathrm{var}(\mathbf{Y})$
9. $\mathrm{cov}(\mathbf{A}\mathbf{X},\mathbf{B}\mathbf{X}) = \mathbf{A}\,\mathrm{cov}(\mathbf{X},\mathbf{X})\mathbf{B}^{\top}$
10. 若$\mathbf{X}$ 与$\mathbf{Y}$ 是独立的，则有$\mathrm{cov}(\mathbf{X},\mathbf{Y}) = 0$
11. $\Sigma = \Sigma^{\top}$

其中 $\mathbf{X},\mathbf{X_1}$ 与$\mathbf{X_2}$ 是随机$(\mathbf{p}\times\mathbf{1})$向量, $\mathbf{Y}$ 是随机$(\mathbf{q}\times\mathbf{1})$向量, $\mathbf{a}$ 是$(\mathbf{p}\times\mathbf{1})$ 向量, $\mathbf{A}$ 与$\mathbf{B}$ 是$(\mathbf{q}\times\mathbf{p})$ 矩阵。

https://zh.wikipedia.org/zh-cn/%E5%8D%8F%E6%96%B9%E5%B7%AE%E7%9F%A9%E9%98%B5

二分类任务 – 线性判别分析

- **LDA的思想**
  - 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
  - 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

- **一些变量**
  - 第i类示例的集合 X_i
  - 第i类示例的均值向量 $\boldsymbol{\mu}_i$
  - 第i类示例的协方差矩阵 Σ_i
  - 两类样本的中心在直线上的投影：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$
  - 投影后，两类样本的协方差：$\boldsymbol{w}^{\mathrm{T}}\Sigma_0\boldsymbol{w}$ 和 $\boldsymbol{w}^{\mathrm{T}}\Sigma_1\boldsymbol{w}$

第0类

$$E[(\omega^T x - E(\omega^T x))(\omega^T x - E(\omega^T x))]$$

$$E\left[\left(\omega^T x - \omega^T\overrightarrow{\mu_0}\right)\left(\omega^T x - \omega^T\overrightarrow{\mu_0}\right)\right]$$

$$E[\omega^T(x - \overrightarrow{\mu_0})(x - \overrightarrow{\mu_0})^T\omega]$$

$$\omega^T E[(x - \overrightarrow{\mu_0})(x - \overrightarrow{\mu_0})^T]\omega$$

$$\Downarrow$$

$$\Sigma_0$$

二分类任务 – 线性判别分析

- **最大化目标**

$$J = \frac{\left\| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_1 \right\|_2^2}{\boldsymbol{w}^{\mathrm{T}} \mathbf{\Sigma}_0 \boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}} \mathbf{\Sigma}_1 \boldsymbol{w}} = \frac{\boldsymbol{w}^{\mathrm{T}} \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^{\mathrm{T}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \left(\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 \right) \boldsymbol{w}}$$

- **类内散度矩阵（对称）**

$$\mathbf{S}_w = \mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 = \sum_{\boldsymbol{x} \in X_0} \left(\boldsymbol{x} - \boldsymbol{\mu}_0 \right) \left(\boldsymbol{x} - \boldsymbol{\mu}_0 \right)^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_1} \left(\boldsymbol{x} - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{x} - \boldsymbol{\mu}_1 \right)^{\mathrm{T}}$$

- **类间散度矩阵（对称）** $\mathbf{S}_b = \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^{\mathrm{T}}$

二分类任务 – 线性判别分析

- **广义瑞利商（generalized Rayleigh quotient）**

$$J = \frac{\boldsymbol{w}^{\mathrm{T}} \mathrm{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \mathrm{S}_w \boldsymbol{w}}$$

- **令** $\boldsymbol{w}^{\mathrm{T}} \mathrm{S}_w \boldsymbol{w} = 1$**，最大化广义瑞利商等价形式为**

$$\min_{\boldsymbol{w}} \ -\boldsymbol{w}^{\mathrm{T}} \mathrm{S}_b \boldsymbol{w}$$

$$\text{s.t. } \boldsymbol{w}^{\mathrm{T}} \mathrm{S}_w \boldsymbol{w} = 1$$

$$\min -\omega^T S_b \omega + \lambda(\omega^T S_\omega \omega - 1)$$

对ω求导：

$$-2S_b\omega + 2\lambda S_\omega \omega = \vec{0}$$

$$S_b \omega = \lambda S_\omega \omega$$

- **运用拉格朗日乘子法**

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

$$\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x} \qquad (81)$$

p.11 matrixbook

二分类任务－ 线性判别分析

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

- **同向向量**

$$\underbrace{\mathrm{S}_b \boldsymbol{w}}_{\text{同向向量}} = \lambda \underbrace{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}_{\text{同向向量}}$$

- **结果**

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

- **求解**
  - **奇异值分解** $\mathrm{S}_w = \mathrm{U}\Sigma\mathrm{V}^T$ $\mathbf{S}_w^{-1} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$

- **LDA的贝叶斯决策论解释**
  - **两类数据同先验、满足高斯分布且协方差相等时，LDA达到最优分类**

LDA推广 – 多(N)分类任务

- **全局散度矩阵**

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

$$\mu = \frac{1}{m}\sum_{i=1}^{N} m_i \mu_i$$

- **类内散度矩阵**

$$\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}$$

其中

$$\mathbf{S}_{w_i} = \sum_{\boldsymbol{x}\in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$$

- **可求得**

$$\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

LDA推广－ 多分类任务

- **优化目标**

$$\max_{\mathbf{W}} \frac{\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}\right)}$$

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

$$\Updownarrow$$

$$\mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W}$$

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

[Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition (Second ed.). Academic Press.]

- **多分类LDA将样本投影到N-1维空间，N-1通常远小于数据原有的属性数，因此LDA也被视为一种监督降维技术**

多分类学习

- 多分类学习方法
  - 二分类学习方法推广到多类（如多分类LDA）
  - 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
    - 对于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）

多分类学习－ 一对一

- 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

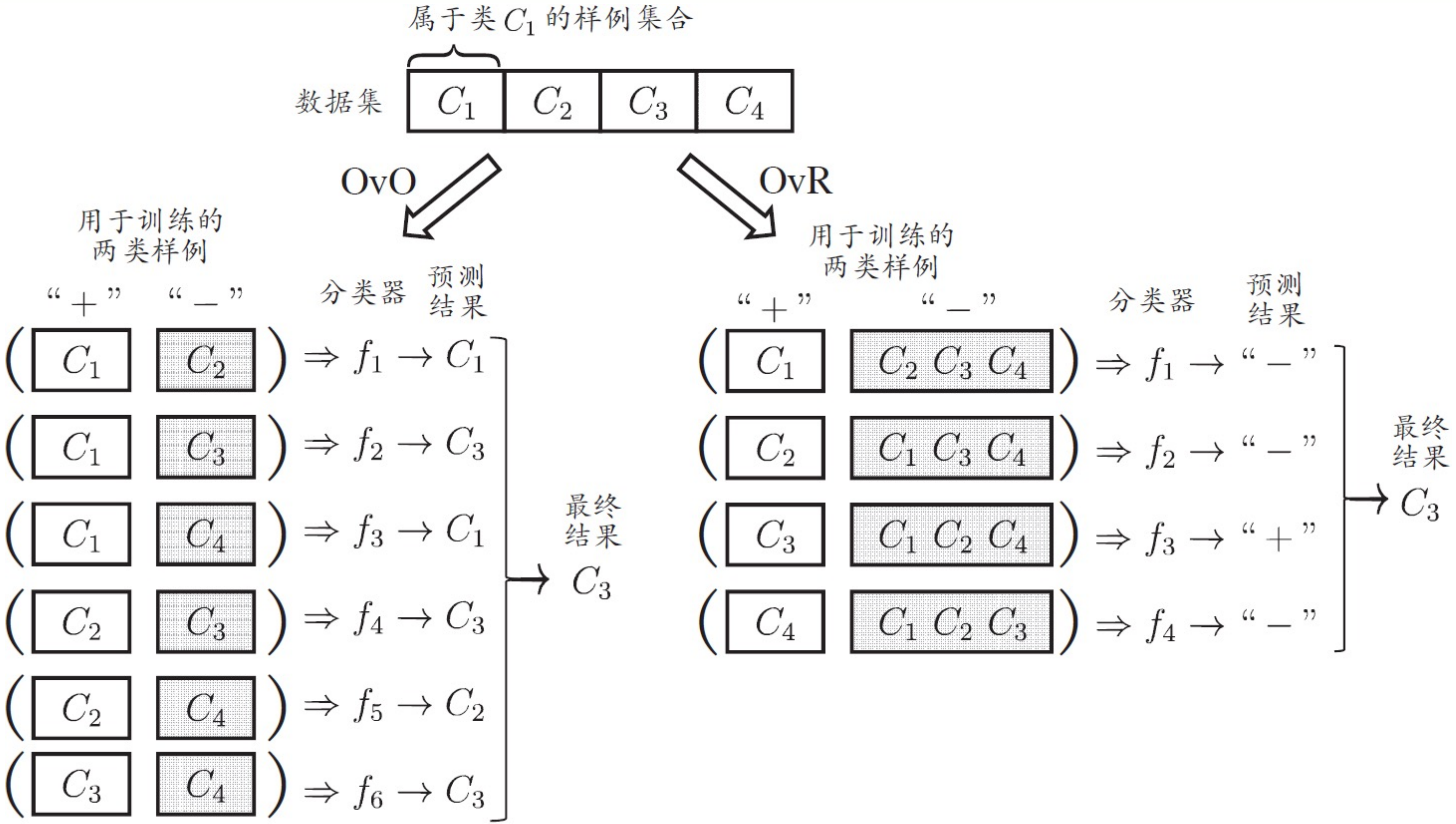
- 测试阶段
  - 新样本提交给所有分类器预测
    - N(N-1)/2 个分类结果
  - 投票产生最终分类结果
    - 被预测最多的类别为最终类别

多分类学习 – 一对其余

- 任务拆分
  - 某一类作为正例，其他反例
    - 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
  - 比较各分类器预测置信度
    - 置信度最大类别作为最终类别

多分类学习 – 两种策略比较



多分类学习– 两种策略比较

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

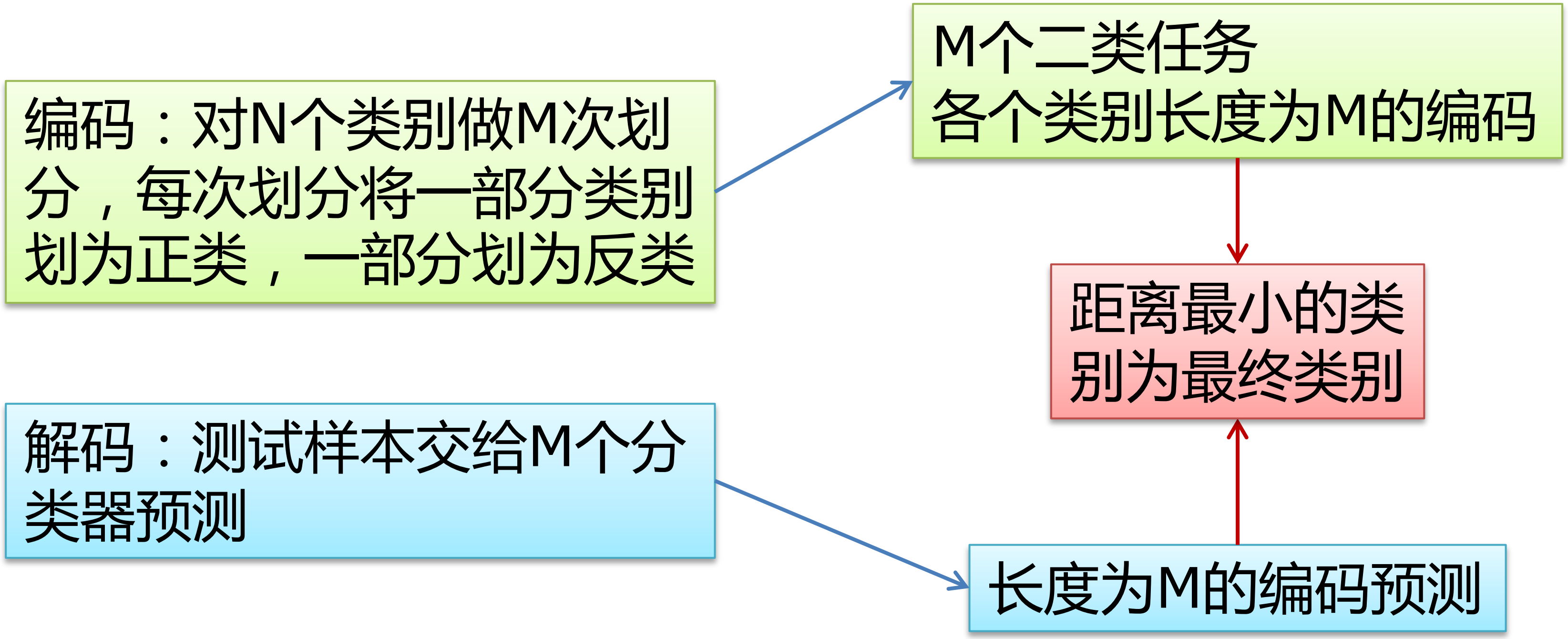
一对其余

- 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

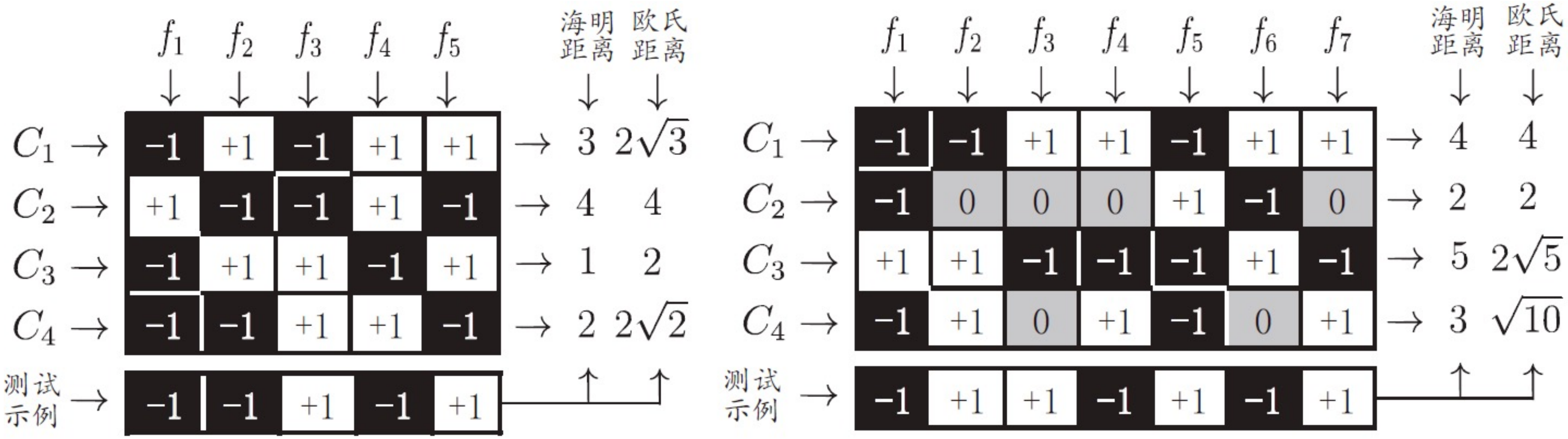
多分类学习 – 多对多

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



多分类学习 – 多对多

- **纠错输出码(Error Correcting Output Code, ECOC)**



(a) 二元 ECOC 码 [Dietterich and Bakiri,1995]

(b) 三元 ECOC 码 [Allwein et al. 2000]

- **ECOC编码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**

类别不平衡问题

- 类别不平衡（class imbalance）
  - 不同类别训练样例数相差很大情况（正类为小类）
- 线性模型 $y = w^T x + b$代表了样本为正例的可能性

类别平衡正例预测 $\frac{y}{1-y} > 1$ 　　 $\frac{y}{1-y} > \frac{m^+}{m^-}$ 　正负类比例

- 再缩放
  - 欠采样（undersampling）
    - 去除一些反例使正反例数目接近（EasyEnsemble [Liu et al.,2009]）
  - 过采样（oversampling）
    - 增加一些正例使正反例数目接近（SMOTE [Chawla et al.2002]）
  - 阈值移动（threshold-moving）

优化提要

- 各任务下（回归、分类）各个模型优化的目标
  - 最小二乘法：最小化均方误差
  - 对数几率回归：最大化样本分布似然
  - 线性判别分析：投影空间内最小（大）化类内（间）散度

- 参数的优化方法
  - 最小二乘法：线性代数
  - 对数几率回归：凸优化梯度下降、牛顿法
  - 线性判别分析：矩阵论、广义瑞利商

总结

- 线性回归
  - 最小二乘法（最小化均方误差）
- 二分类任务
  - 对数几率回归
    - 单位阶跃函数、对数几率函数、极大似然法
  - 线性判别分析
    - 最大化广义瑞利商
- 多分类学习
  - 一对一
  - 一对其余
  - 多对多
    - 纠错输出码
- 类别不平衡问题
  - 基本策略：再缩放

- **作业：3.3**

谢　　谢！