第三章：线性模型

主讲：连德富 特任教授 | 博士生导师
邮箱：liandefu@ustc.edu.cn
手机：13739227137
主页：http://staff.ustc.edu.cn/~liandefu

基本形式

- 线性模型一般形式

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

$\boldsymbol{x} = (x_1; x_2; \cdots, x_d)$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

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

$\boldsymbol{w} = (w_1; w_2; \cdots, w_d)$是属性的权重

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

其中$\boldsymbol{x}_i=(x_{i1};x_{i2};\cdots,x_{id}),y_i\in\mathbb{R}$

- 线性回归目标
  - 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
  - 无“序”关系
    - 有k个属性值，则转换为k维向量

线性回归

- 单一属性的线性回归目标

$$f(x_i) = w\,x_i + b \text{ 使得 } f(x_i) \approx y_i$$

- 参数/模型估计：最小二乘法（least square method）

$$(w^\star, b^\star) = \underset{w,b}{\arg\min} \sum_{i=1}^{m} \left(f(x_i) - y_i\right)^2$$

$$= \underset{w,b}{\arg\min} \sum_{i=1}^{m} \left(w x_i + b - y_i\right)^2$$

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

$$\frac{\partial E(w,b)}{\partial w}=2\left(w\sum_i x_i^2-\sum_i (y_i-b)x_i\right)$$

$$\frac{\partial E(w,b)}{\partial b}=2\left(mb-\sum_i (y_i-wx_i)\right)$$

线性回归 - 最小二乘法

- 令导数梯度等于0，得到闭形式解

$$\frac{\partial E(w,b)}{\partial w} = w\sum_i x_i^2 - \sum_i (y_i - b)x_i$$

$$= w\sum_i x_i^2 - \sum_i (y_i - \bar{y} + w\bar{x})x_i$$

$$= w\,(\sum_i x_i^2 - \bar{x}\sum_i x_i) - \sum_i (y_i - \bar{y})x_i$$

$$= w\,(\sum_i x_i^2 - \bar{x}\sum_i x_i) - (\sum_i y_i x_i - \sum_i \bar{y}x_i)$$

$$= w\,(\sum_i x_i^2 - \bar{x}\sum_i x_i) - (\sum_i y_i x_i - \sum_i y_i\bar{x})$$

$$= w\,(\sum_i x_i^2 - \bar{x}\sum_i x_i) - \sum_i y_i(x_i - \bar{x})$$

$$\boxed{b = \frac{1}{m}\sum_i (y - wx_i) = \bar{y} - w\bar{x}}$$

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

多元线性回归

- 给定数据集

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

$$\text{其中}\boldsymbol{x}_i=(x_{i1};x_{i2};\cdots,x_{id}),y_i\in\mathbb{R}$$

- 多元线性回归目标

$$f(\boldsymbol{x}_i)=\boldsymbol{w}^\top\boldsymbol{x}_i+b\ \text{使得}\ f(\boldsymbol{x}_i)\approx y_i$$

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

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

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

多元线性回归 - 最小二乘法

- 最小二乘法（least square method）

$$\hat{\boldsymbol{w}}^{\star} = \arg\min_{\hat{\boldsymbol{w}}} \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}}\|_2^2 = \underbrace{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})^{\top}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})}_{E(\hat{\boldsymbol{w}})}$$

- 求$E(\hat{\boldsymbol{w}})$关于变量$\hat{\boldsymbol{w}}$的导数得到

$$\nabla_{\hat{\boldsymbol{w}}} E(\hat{\boldsymbol{w}}) = 2\boldsymbol{X}^{\top}(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})$$

多元线性回归 – 满秩讨论

- $\boldsymbol{X}^\top \boldsymbol{X}$是满秩矩阵或正定矩阵，则

$$\nabla_{\hat{\boldsymbol{w}}} E(\hat{\boldsymbol{w}}) = 2\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) = 0 \Rightarrow \boxed{\hat{\boldsymbol{w}}^\star = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}}$$

- 把$\hat{\boldsymbol{w}}^\star$代回$f(\boldsymbol{x}_i)$，线性回归模型为

$$\boxed{f(\hat{\boldsymbol{x}}_i) = \boldsymbol{x}_i(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}}$$

- 如果$\boldsymbol{X}^\top \boldsymbol{X}$不是满秩矩阵
  - 根据归纳偏好选择解
  - 引入正则化

一元线性回归

- 重新考虑一个特征的情形

$$\boldsymbol{X} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix} \qquad \boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & m \end{pmatrix} \qquad \boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{pmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{1}{m \sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} m & -\sum_i x_i \\ -\sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{1}{m \sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} m \sum_i x_i y_i - \sum_i x_i \sum_j y_j \\ \sum_i x_i^2 \sum_j y_j - \sum_i x_i y_i \sum_j x_j \end{pmatrix}$$

一元线性回归

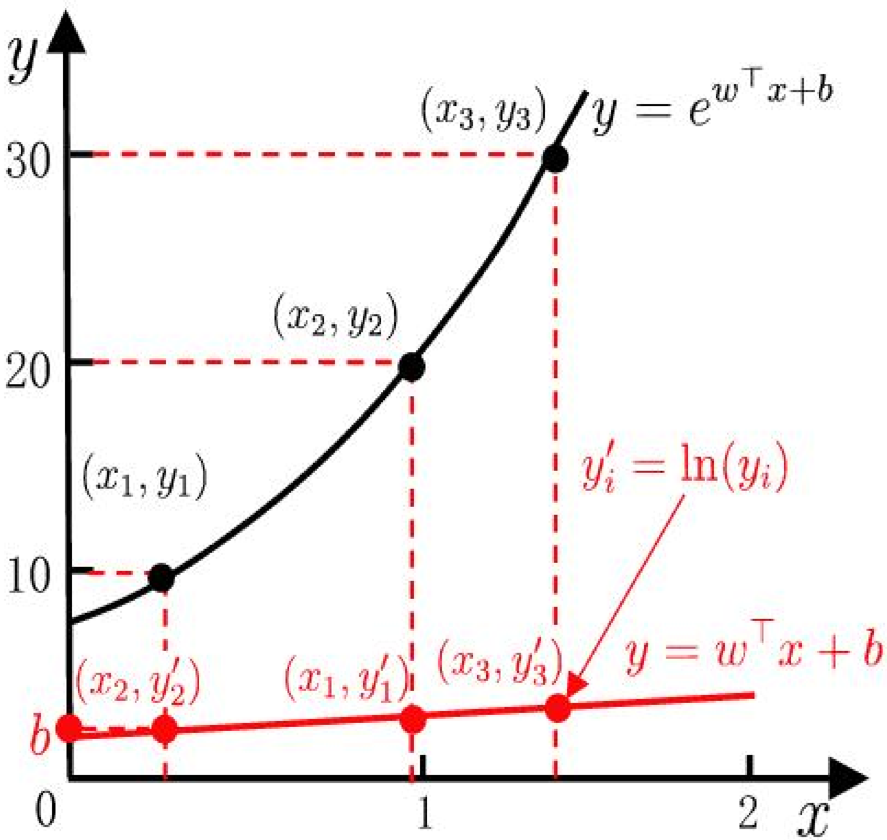
$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \frac{1}{m\sum_i x_i^2 - (\sum_i x_i)^2}\begin{pmatrix} m\sum\limits_i x_i y_i - \sum\limits_i x_i \sum\limits_j y_j \\ \sum\limits_i x_i^2 \sum\limits_j y_j - \sum\limits_i x_i y_i \sum\limits_j x_j \end{pmatrix}$$

$$w = \frac{m\sum_i x_i y_i - \sum_i x_i \sum_j y_j}{m\sum_i x_i^2 - (\sum_i x_i)^2} = \frac{\sum_i x_i y_i - \sum_j y_j\,\bar{x}}{\sum_i x_i^2 - \frac{1}{m}(\sum_i x_i)^2} = \boxed{\frac{\sum_i (x_i - \bar{x})\,y_i}{\sum_i x_i^2 - \frac{1}{m}(\sum_i x_i)^2}}$$

$$b = \frac{\sum_i x_i^2 \sum_j y_j - \sum_i x_i y_i \sum_j x_j}{m\sum_i x_i^2 - (\sum_i x_i)^2} = \frac{\overline{x^2}\sum_j y_j - \sum_i x_i y_i \bar{x}}{\sum_i x_i^2 - \frac{1}{m}(\sum_i x_i)^2} = \boxed{\frac{\sum_j y_j\,(\overline{x^2} - x_i\bar{x})}{\sum_i x_i^2 - \frac{1}{m}(\sum_i x_i)^2}}$$

对数线性回归

- 输出标记的对数为线性模型逼近的目标



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

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

线性回归 - 广义线性模型

- 一般形式

$$y = g^{-1}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

- $g(\cdot)$称为链接函数 (link function)
  - 单调可微

- 对数线性回归 $g(\cdot) = \ln(\cdot)$ 就是广义线性模型的特列

二分类任务

- 预测值与输出标记

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

- 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0就判别为正例，小于0判别为负例，预测值为临界值0时可以任意判别

二分类任务

- 单位阶跃函数缺点
  - 不连续，无法用在广义线性模型中

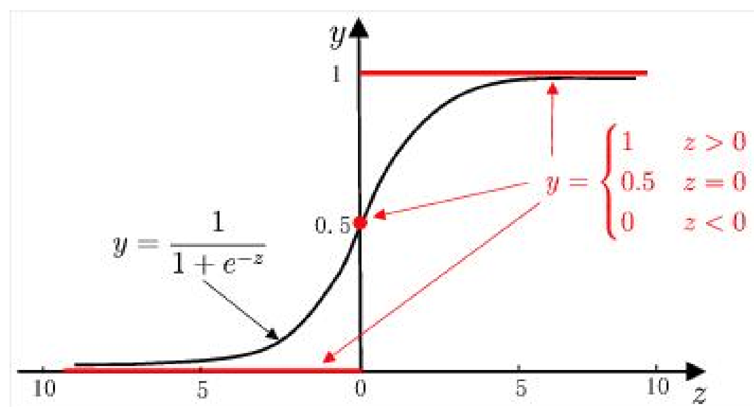
- 替代函数——对数几率函数（logistic function）

$$\sigma(z)=\frac{1}{1+e^{-z}}$$

单调可微、任意阶可导

$$\sigma^{-1}(y)=\ln\frac{y}{1-y}$$

单位阶跃函数与对数几率函数的比较



对数几率回归

- 运用对数几率函数

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \Rightarrow \quad \hat{y} = \frac{1}{1+e^{-(\boldsymbol{w}^\top \boldsymbol{x}+b)}}$$

$\hat{y}$视为样本$\boldsymbol{x}$ 作为正例的可能性

$$\Downarrow$$

$$\boldsymbol{w}^\top \boldsymbol{x} + b = \ln \frac{\hat{y}}{1-\hat{y}}$$

$\frac{\hat{y}}{1-\hat{y}}$称为几率，反映了$\boldsymbol{x}$作为正例的相对可能性

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

如何优化参数$(\boldsymbol{w}, b)$？

$$\hat{y} = \frac{1}{1 + e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)}} \triangleq P(y = 1 | \boldsymbol{x}; \boldsymbol{w}, b)$$

$$P(y = 0 | \boldsymbol{x}; \boldsymbol{w}, b) = \frac{e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)}}{1 + e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)}} = \frac{1}{1 + e^{\boldsymbol{w}^\top \boldsymbol{x} + b}}$$

给定数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}, \quad y_i \in \{0,1\}$

- 极大似然法 最大化 对数似然

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} y_i \log P(y = 1 | \boldsymbol{x}_i; \boldsymbol{w}, b) + (1 - y_i) \log P(y = 0 | \boldsymbol{x}_i; \boldsymbol{w}, b)$$

$$= \sum_{i=1}^{m} \log P(y = y_i | \boldsymbol{x}_i; \boldsymbol{w}, b)$$

对数几率回归 - 极大似然法

- 极大似然法 最小化 负对数似然

$$\ell(\hat{\boldsymbol{w}}) = -\sum_{i=1}^{m} \log P(y = y_i | \hat{\boldsymbol{x}}_i; \hat{\boldsymbol{w}})$$

$$P(y = 1 | \boldsymbol{x}; \boldsymbol{w}, b) = \frac{e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}} = \frac{e^{1 \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}} = \frac{e^{y \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}$$

$$P(y = 0 | \boldsymbol{x}; \boldsymbol{w}, b) = \frac{1}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}} = \frac{e^{0 \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}} = \frac{e^{y \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}}}$$

$$\ell(\hat{\boldsymbol{w}}) = -\sum_{i=1}^{m} \log \frac{e^{y_i \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}_i}}{1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}_i}} = \sum_{i=1}^{m} - y_i \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}_i + \log\left(1 + e^{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}_i}\right)$$

对数几率回归 - 极大似然法

- 考察函数$f(x) = -ax + \ln(1 + \exp(x))$

一阶导数 $$f'(x) = -a + \frac{1}{1 + e^{-x}}$$

二阶导数 $$f''(x) = \left(\frac{1}{1 + e^{-x}}\right)' = \frac{e^{-x}}{(1 + e^{-x})^2} = \sigma(x)(1 - \sigma(x))$$

因为$\sigma(x) \in (0,1)$，所以$f''(x) > 0$，因此$f(x)$是凸函数。

因为$\ell(\hat{\boldsymbol{w}})$是$f(x)$ 和$g(\boldsymbol{w}) = \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}_i$的复合函数，即$\ell(\hat{\boldsymbol{w}}) = f(g(\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{x}}_i))$，
所以$\ell(\hat{\boldsymbol{w}})$是关于$\hat{\boldsymbol{w}}$的凸函数

对数几率回归 - 极大似然法

- 负对数似然 $\ell(\hat{\boldsymbol{w}}) = \sum_{i=1}^{m}(-y_i\hat{\boldsymbol{w}}^{\top}\hat{\boldsymbol{x}}_i + \log(1+e^{\hat{\boldsymbol{w}}^{\top}\hat{\boldsymbol{x}}_i}))$

一阶导数

$$\nabla\ell(\hat{\boldsymbol{w}}) = \sum_i f'(z_i)\frac{\partial z_i}{\partial \hat{\boldsymbol{w}}} = \sum_i(-y_i + \frac{1}{1+e^{-z_i}})\,\boldsymbol{x}_i$$

$$= -\sum_i (y_i - P(y=1|\hat{\boldsymbol{x}}_i;\hat{\boldsymbol{w}}))\hat{\boldsymbol{x}}_i$$

p_1

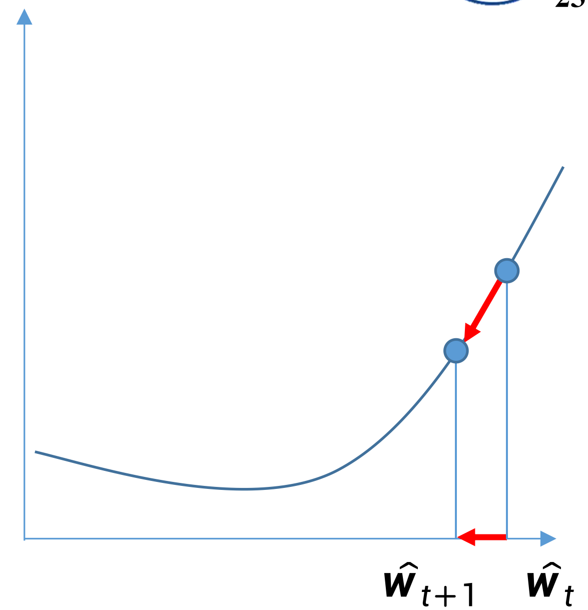
二阶导数

$$\nabla^2\ell(\hat{\boldsymbol{w}}) = \frac{\partial\nabla\ell(\hat{\boldsymbol{w}})}{\partial\hat{\boldsymbol{w}}^{\top}} = \sum_i p_1(1-p_1)\boldsymbol{x}_i\boldsymbol{x}_i^{\top}$$

对数几率回归 - 极大似然法

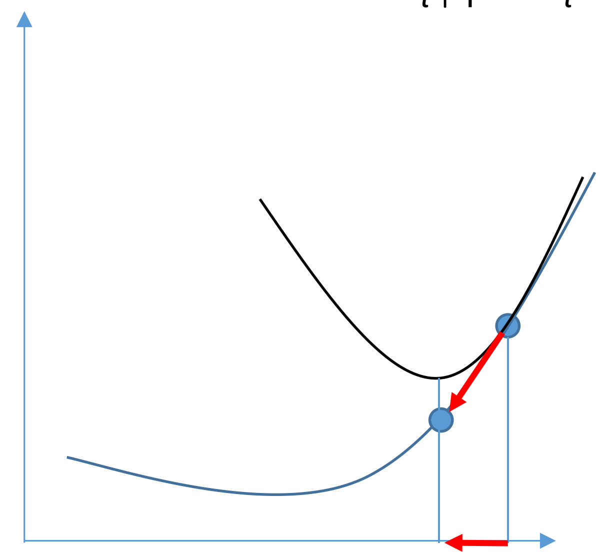
- 梯度下降法

$$\text{while } \|\nabla \ell(\hat{\boldsymbol{w}})\| > \delta \text{ do}$$
$$\quad \hat{\boldsymbol{w}}_{t+1} \leftarrow \hat{\boldsymbol{w}}_t - a\nabla \ell(\hat{\boldsymbol{w}})$$
$$\text{end while}$$



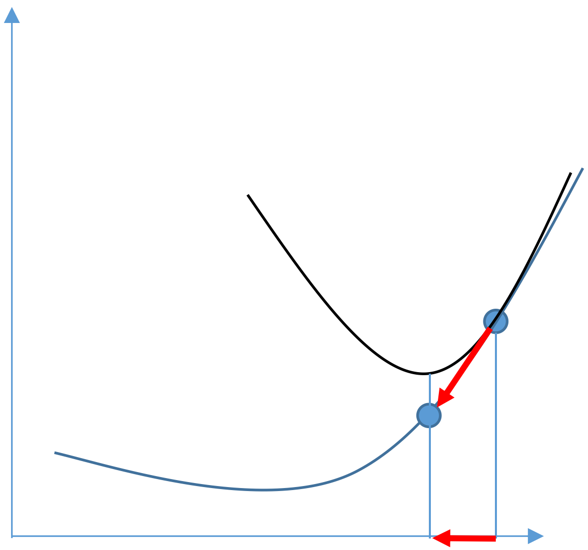
- 牛顿法

$$\text{while } \|\nabla \ell(\hat{\boldsymbol{w}})\| > \delta \text{ do}$$
$$\quad \hat{\boldsymbol{w}}_{t+1} \leftarrow \hat{\boldsymbol{w}}_t - (\nabla^2 \ell(\hat{\boldsymbol{w}}))^{-1}\nabla \ell(\hat{\boldsymbol{w}})$$
$$\text{end while}$$



对数几率回归 - 极大似然法

牛顿法



while $\|\nabla \ell(\hat{\boldsymbol{w}})\| > \delta$ do

$\quad \hat{\boldsymbol{w}}_{t+1} \leftarrow \hat{\boldsymbol{w}}_t - (\nabla^2 \ell(\hat{\boldsymbol{w}}))^{-1} \nabla \ell(\hat{\boldsymbol{w}})$

end while

- 考虑$\ell(\hat{\boldsymbol{w}})$在$\hat{\boldsymbol{w}}_t$处的二阶泰勒展开

$$\ell(\hat{\boldsymbol{w}}) \approx \ell(\hat{\boldsymbol{w}}_t) + \nabla \ell(\hat{\boldsymbol{w}}_t)^\top (\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_t) + \frac{1}{2} (\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_t)^\top \nabla^2 \ell(\hat{\boldsymbol{w}}) (\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_t)$$

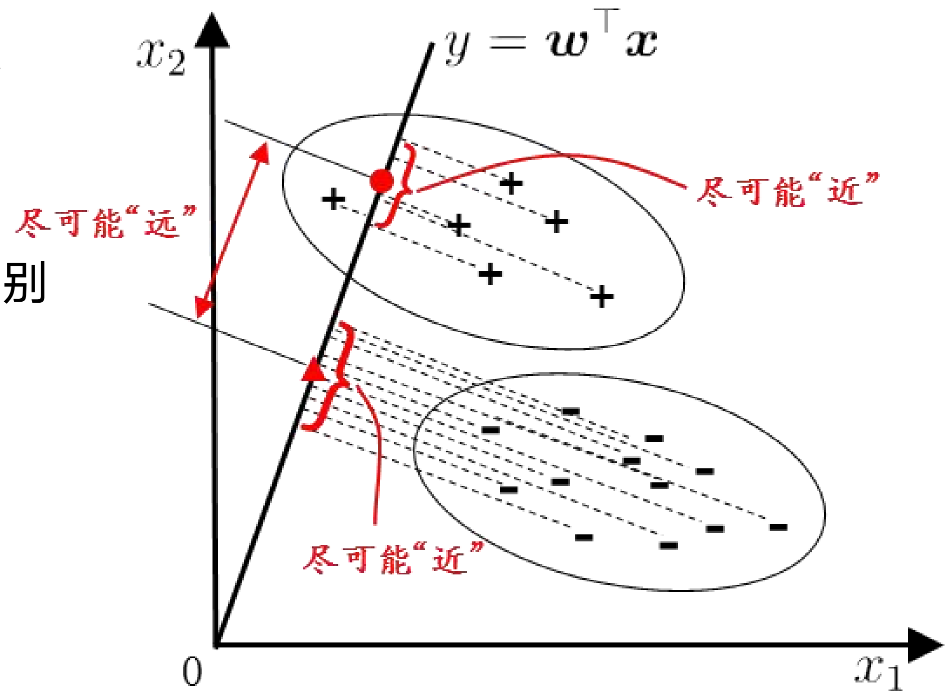
- 对$\hat{\boldsymbol{w}}$求导数并令其等于0，得到$\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}_t - (\nabla^2 \ell(\hat{\boldsymbol{w}}))^{-1} \nabla \ell(\hat{\boldsymbol{w}})$

二分类任务 – 线性判别分析

- 线性判别分析（Linear Discriminant Analysis）[Fisher, 1936]

投影到低纬空间，使得

- 欲使同类样例的投影点尽可能接近
- 欲使异类样例的投影点尽可能远离
- 新样本投影后根据投影位置进行判别



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

- 给定数据集$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}, y_i \in \{0,1\}$

令

  - 第c类示例的集合 X_c
  - 第c类示例的均值向量 $\boldsymbol{\mu}_c = \frac{1}{|X_c|}\sum_{\boldsymbol{x}\in X_c}\boldsymbol{x}$
  - 第c类示例的协方差矩阵 $\boldsymbol{\Sigma}_c = \sum_{x\in X_c}(\boldsymbol{x}-\boldsymbol{\mu}_c)(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top$

若将数据投影到方向$\boldsymbol{w}$确定的直线上

- 两类样本的中心在直线上的投影 $\frac{1}{|X_c|}\sum_{\boldsymbol{x}\in X_c}\boldsymbol{w}^\top\boldsymbol{x} = \boldsymbol{w}^\top\boldsymbol{\mu}_c$
- 两类样本的协方差 $\sum_{\boldsymbol{x}\in X_c}(\boldsymbol{w}^\top\boldsymbol{x} - \boldsymbol{w}^\top\boldsymbol{\mu}_c)^2 = \boldsymbol{w}^\top\boldsymbol{\Sigma}_c\boldsymbol{w}$

二分类任务 – 线性判别分析

- 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

- 同时考虑两者，则可得到最大化目标

定义类内散度

$$S_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top$$

$$J = \frac{\|\boldsymbol{w}^\top \boldsymbol{\mu}_0 - \boldsymbol{w}^\top \boldsymbol{\mu}_1\|_2^2}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_0 \boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w}} = \frac{\boldsymbol{w}^\top (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}}{\boldsymbol{w}^\top (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \boldsymbol{w}} = \frac{\boldsymbol{w}^\top \mathbf{S}_b \boldsymbol{w}}{\boldsymbol{w}^\top \mathbf{S}_w \boldsymbol{w}}$$

$$S_w = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$$

定义类间散度

二分类任务－ 线性判别分析

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \mathbf{S}_b \boldsymbol{w}}{\boldsymbol{w}^\top \mathbf{S}_w \boldsymbol{w}}$$

广义瑞利商
(generalized Rayleigh quotient)

- 若$\boldsymbol{w}$是一个解，则对于任意常数α, $\alpha\boldsymbol{w}$也是解

- 不失一般性，令$\boldsymbol{w}^\top \mathbf{S}_w \boldsymbol{w} = 1$，则等价于

$$\max_{\boldsymbol{w}} \boldsymbol{w}^\top \mathbf{S}_b \boldsymbol{w} \qquad \text{s.t } \boldsymbol{w}^\top \mathbf{S}_w \boldsymbol{w} = 1$$

- 引入拉格朗日乘子λ，并令朗格拉日函数梯度等于0，可以得到

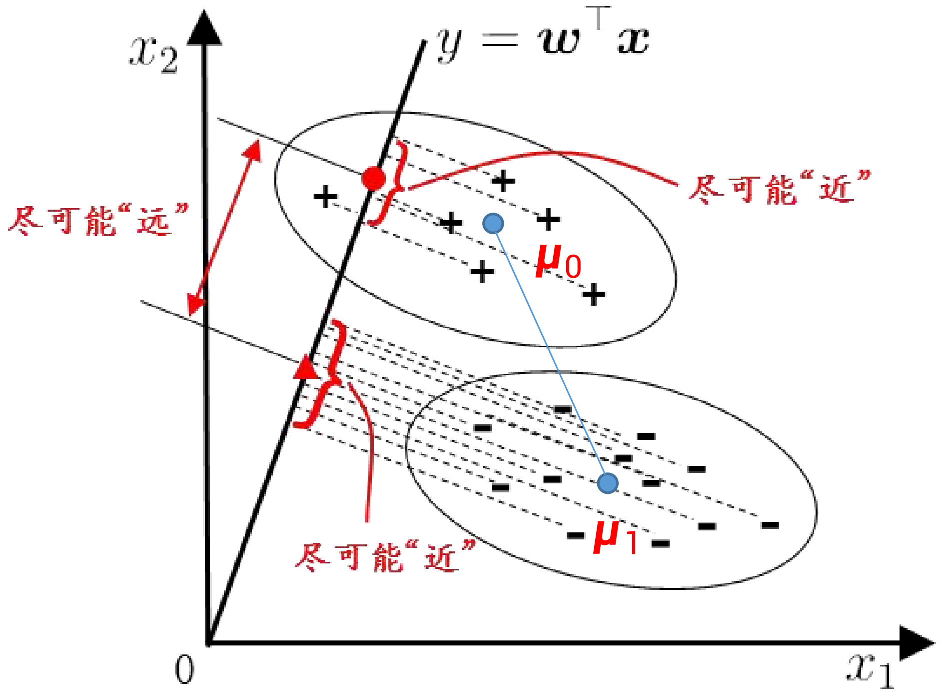
$$\mathbf{S}_b \boldsymbol{w} = \lambda \mathbf{S}_w \boldsymbol{w}$$

二分类任务－ 线性判别分析

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

- $\mathbf{S}_b\boldsymbol{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top\boldsymbol{w} \quad \propto (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

- 由此可得 $\boldsymbol{w} \propto \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$



LDA推广 – 多分类任务

- 全局散度矩阵 $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$

- 类内散度矩阵 $\mathbf{S}_w = \sum_c \mathbf{S}_{w_c}$ $\quad \mathbf{S}_{w_c} = \sum_{\boldsymbol{x} \in X_c} (\boldsymbol{x} - \boldsymbol{\mu}_c)(\boldsymbol{x} - \boldsymbol{\mu}_c)^\top$

- 类间散度矩阵 $\mathbf{S}_{w_c} = \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}\,\boldsymbol{x}^\top - m_c \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top$ $\quad \mathbf{S}_t = \sum_i \boldsymbol{x}_i\,\boldsymbol{x}_i^\top - m\boldsymbol{\mu}\,\boldsymbol{\mu}^\top$

$$\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_c m_c \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top - m\boldsymbol{\mu}\,\boldsymbol{\mu}^\top$$

$$m\boldsymbol{\mu} = \sum_c m_c \boldsymbol{\mu}_c \qquad m = \sum_c m_c$$

$$= \sum_c m_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top$$

LDA推广－ 多分类任务

- 优化目标

$$\max_{\boldsymbol{W}} \frac{\operatorname{tr}(\boldsymbol{W}^\top \boldsymbol{S}_b \boldsymbol{W})}{\operatorname{tr}(\boldsymbol{W}^\top \boldsymbol{S}_w \boldsymbol{W})} = \frac{\sum_i \boldsymbol{w}_i^\top \boldsymbol{S}_b \boldsymbol{w}_i}{\sum_i \boldsymbol{w}_i^\top \boldsymbol{S}_w \boldsymbol{w}_i}$$

- 若$\boldsymbol{W}$是一个解，则对于任意常数a, $a\boldsymbol{W}$也是解

- 等价于 $\max_{\boldsymbol{W}} tr(\boldsymbol{W}^\top \boldsymbol{S}_b \boldsymbol{W}) \quad \text{s.t. } tr(\boldsymbol{W}^\top \boldsymbol{S}_w \boldsymbol{W}) = 1$

- 引入拉格朗日乘子λ，并令朗格拉日函数梯度等于0，可以得到广义特征值问题

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

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

多分类学习

- 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

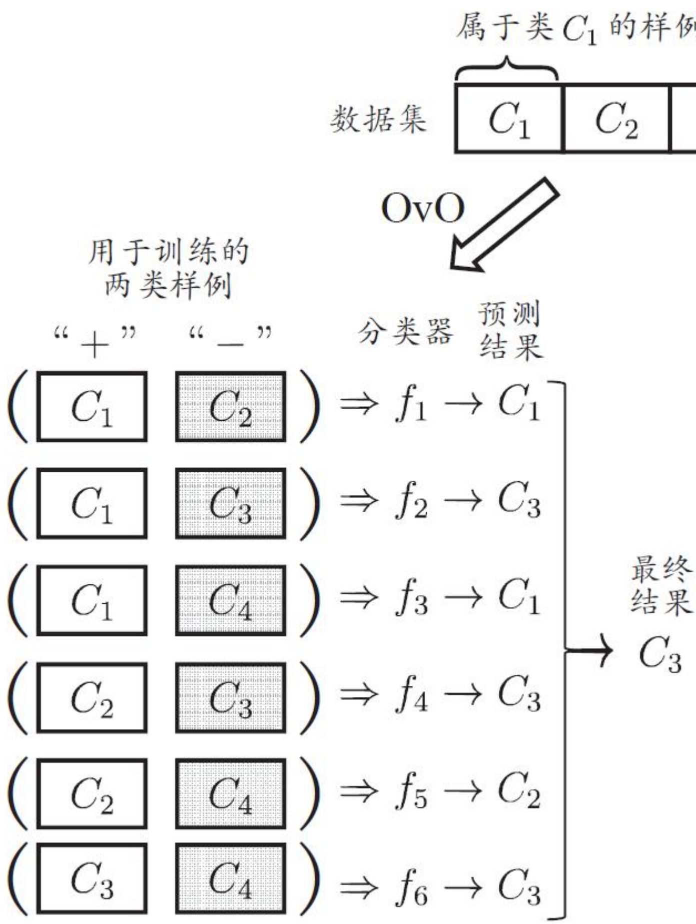
✓对问题进行拆分，为拆出的每个二分类任务训练一个分类器

✓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



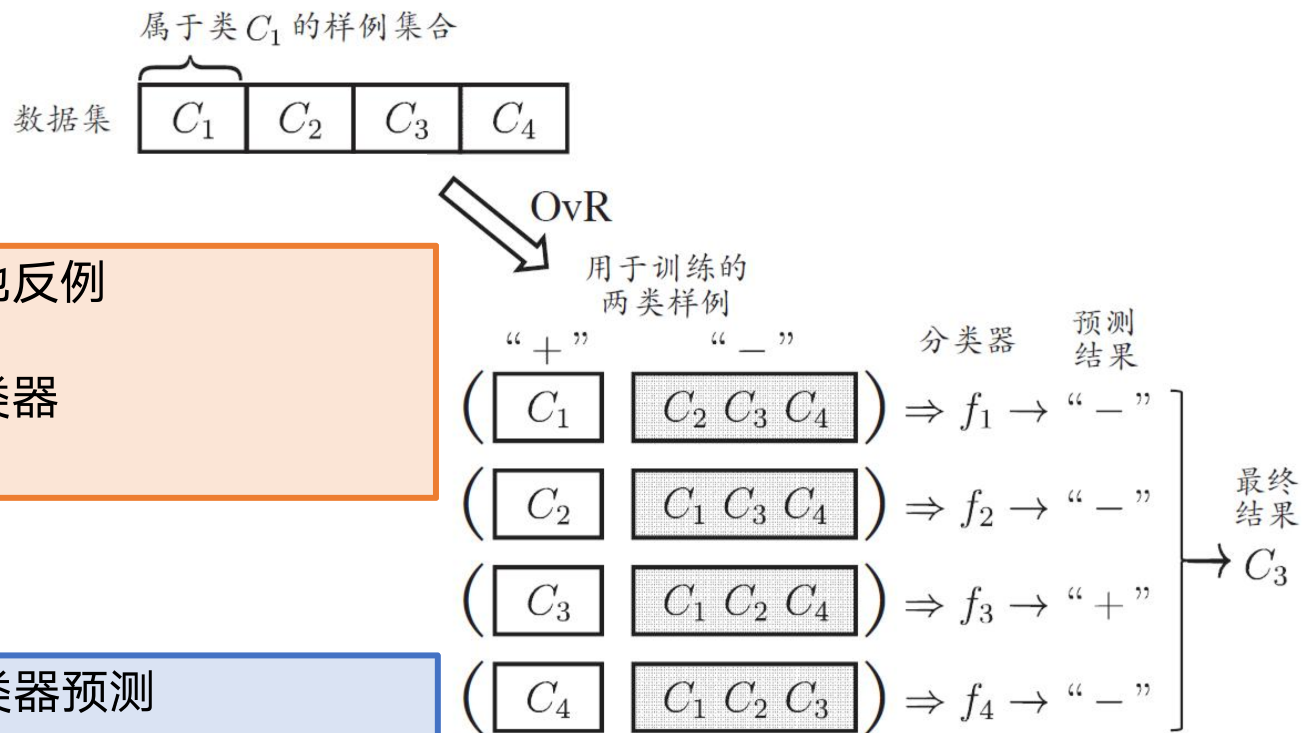
拆分阶段

- 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



拆分阶段

- 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

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

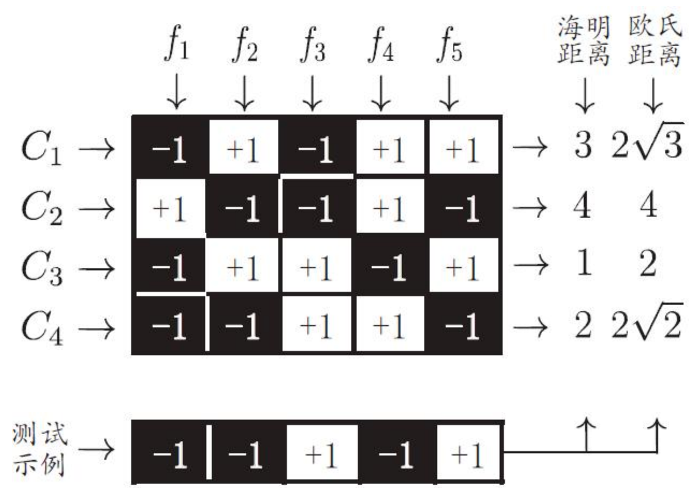
- 纠错输出码（Error Correcting Output Code, ECOC）

编码

对N个类别做M次划分，每次划分将一部分类别划为正类，一部分划为反类，形成二分类训练集

解码

M个分类器分别对测试样本进行预测，预测标记组成一个编码。将距离最小的类别为最终类别



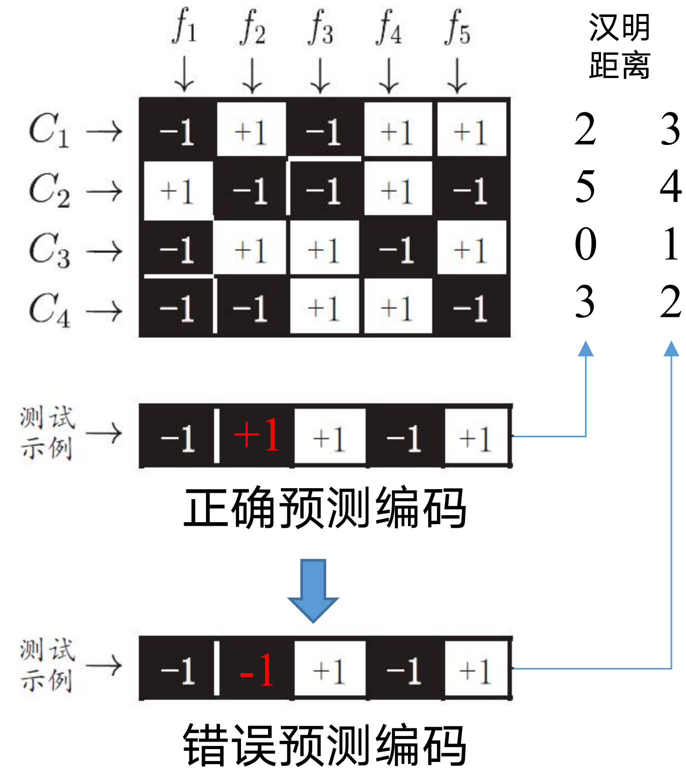
多分类学习 – 多对多

编码

对N个类别做M次划分，每次划分将一部分类别划为正类，一部分划为反类，形成二分类训练集

解码

M个分类器分别对测试样本进行预测，预测标记组成一个编码。将距离最小的类别为最终类别



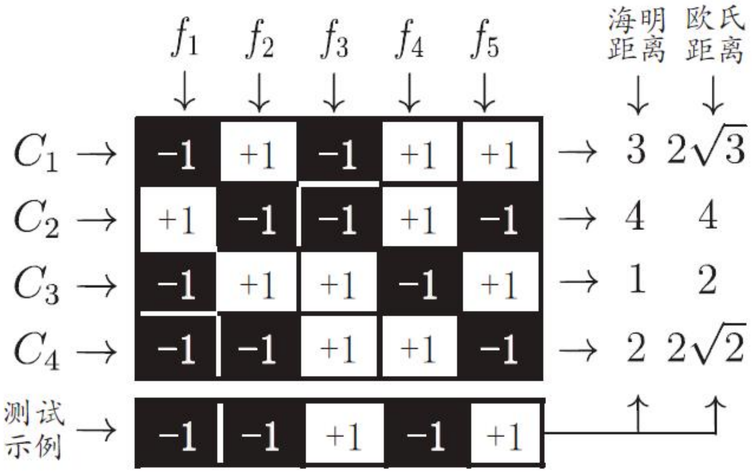
- 纠错能力
  - 若 f_2 预测错误
  - 仍然能产生正确的最终分类

✓ECOC编码越长、对分类器错误纠错能力越强

✓对同等长度编码，理论上任意两个类别之间的编码距离越远，则纠错能力越强

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]

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）

类别平衡正例预测

分类决策规则：$\frac{y}{1-y} > 1$，则为正例

类别不平衡正例预测

分类决策规则：$\frac{y}{1-y} > \frac{m^+}{m^-}$，则正例

y为分类器预测值，表达了正例可能性，几率 $\frac{y}{1-y}$ 反映相对可能性

- 分类器都基于类别平衡分类决策规则决策的，只能对预测值进行缩放

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

类别不平衡问题

"训练集是真实样本总体的无偏采样"假设往往不成立，未必能基于训练集观测几率来推断真实几率

解决办法

- 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）

作业

- 3.2
- 3.7
- 在LDA多分类情形下，试计算类间散度矩阵S_b的秩并证明
- 给出公式3.45的推导过程
- 证明$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$是投影矩阵，并对线性回归模型从投影角度解释