课程2：
线性模型

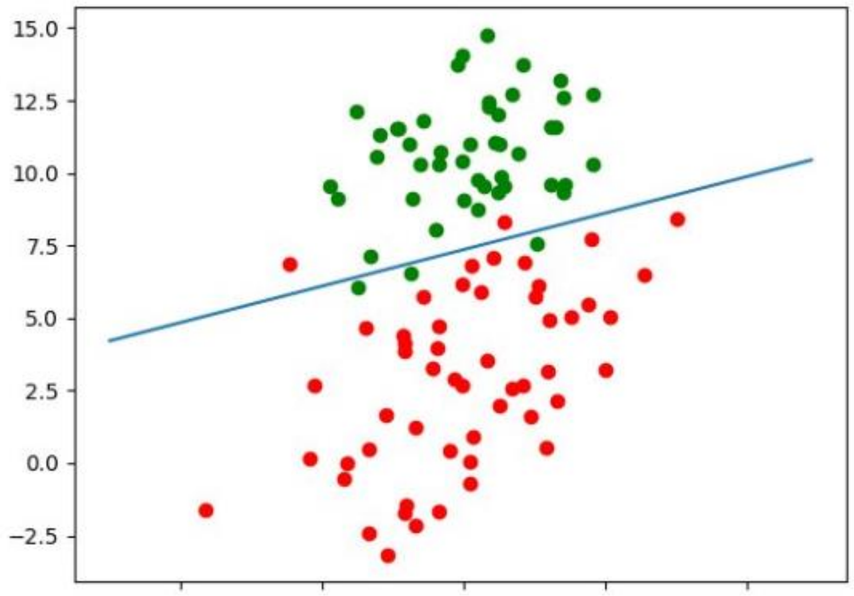
基本形式

问题描述：给定由d个属性描述的示例 $x=(x_1;x_2;...;x_d)$，其中 x_i是x在第i个属性上的取值，线性模型(linear model)试图学得一个通过属性的线性组合来进行预测的函数，即

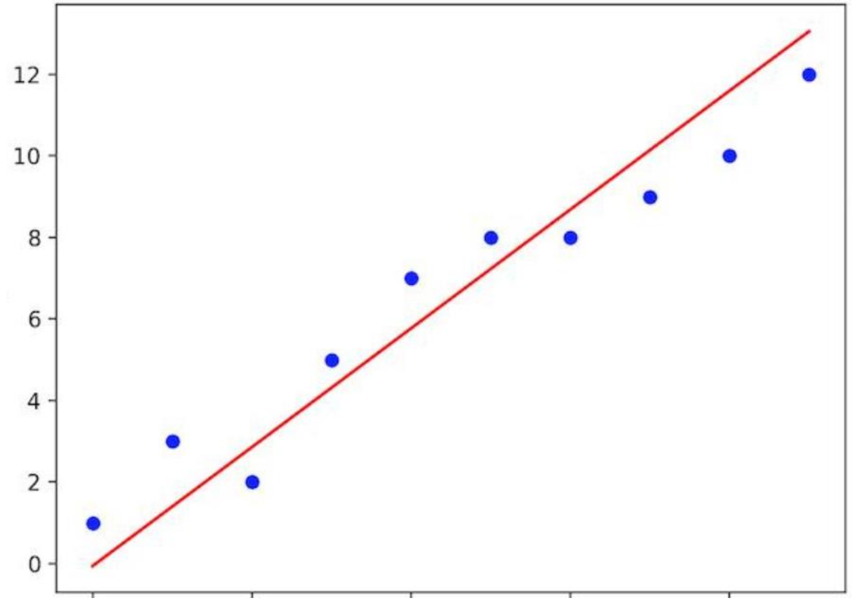
$$f(x)=\omega_1 x_1+\omega_2 x_2+...+\omega_d x_d+d$$

其向量形式为

$$f(x)=\omega^T x+b$$



分类



回归

线性回归

给定数据集其中 $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$，$x_i=(x_{i1};x_{i2};\ldots;x_{id})$， $y_i\in\mathbb{R}$

“线性回归”(linear regression)试图学得一个线性模型以尽可能准确地预测实值输出标记.

- 一元线性回归
- 多元线性回归
- 对数线性回归

线性回归-一元线性回归

问题描述： 只有一个属性，即d=1 ， w，b为单个的数

线性回归试图学得 $f(x_1)=\omega^T x_1+b$ ，使得 $f(x_1)\approx y_i$

要确定w和b，关键在于如何衡量f(x)与y之间的差别

目标函数

$$(\omega^*,b^*)=\underset{(\omega,b)}{\arg\min}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

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

求解w和b使 $E(\omega,b)=\sum_{i=1}^{m}(y_i-\omega x_i-b)^2$ 最小化的过程，称为线性回归模型的最小二乘“参数估计”

线性回归-一元线性回归

将 $E(\omega,b)=\sum_{i=1}^{m}(y_i-\omega x_i-b)^2$ 分别对w和b求导，并令其各自导数等于0可解得

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

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

线性回归-多元线性回归

更一般的情形是对于数据集D，样本由d个属性描述

此时我们试图学得 $f(x_1)=\omega^T x_1+b$ ，使得 $f_{\hat{\omega}}(x_i)\approx y_i$

$$\hat{\omega} = (\omega; b) = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \ldots \\ \omega_d \\ b \end{pmatrix} \qquad X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1d} & 1 \\ X_{21} & X_{22} & \cdots & X_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{md} & 1 \end{pmatrix}$$

线性回归-多元线性回归

$$X*\omega=\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} & 1 \\ x_{21} & x_{22} & \dots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{md} & 1 \end{pmatrix} * \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_d \\ b \end{pmatrix} = \begin{pmatrix} \omega_1 x_{11}+\omega_2 x_{12}+\dots+\omega_d x_{1d}+b \\ \omega_2 x_{21}+\omega_2 x_{22}+\dots+\omega_d x_{2d}+b \\ \vdots \\ \omega_1 x_{m1}+\omega_2 x_{m2}+\dots+\omega_d x_{md}+b \end{pmatrix}$$

将标记写为向量形式y=（$y_1, y_2, \cdots, y_m$），与一元线性回归求解w，b类似，有

$$\hat{\omega}^* = \arg\min_{\hat{\omega}} (y - X\hat{\omega})^T (y - X\hat{\omega})$$

令 $E_{\hat{\omega}} = (y - X\omega)^T (y - X\omega)$ ，对w求导得 $\dfrac{\partial E_{\hat{\omega}}}{\partial \hat{\omega}} = 2X^T(X\hat{\omega} - y)$

- 若X^TX满秩或正定，则 $\hat{\omega}^* = (X^TX)^{-1}X^Ty$
- 若X^TX不满秩，则 $\hat{\omega}$ 有多个解

线性回归-对数线性回归

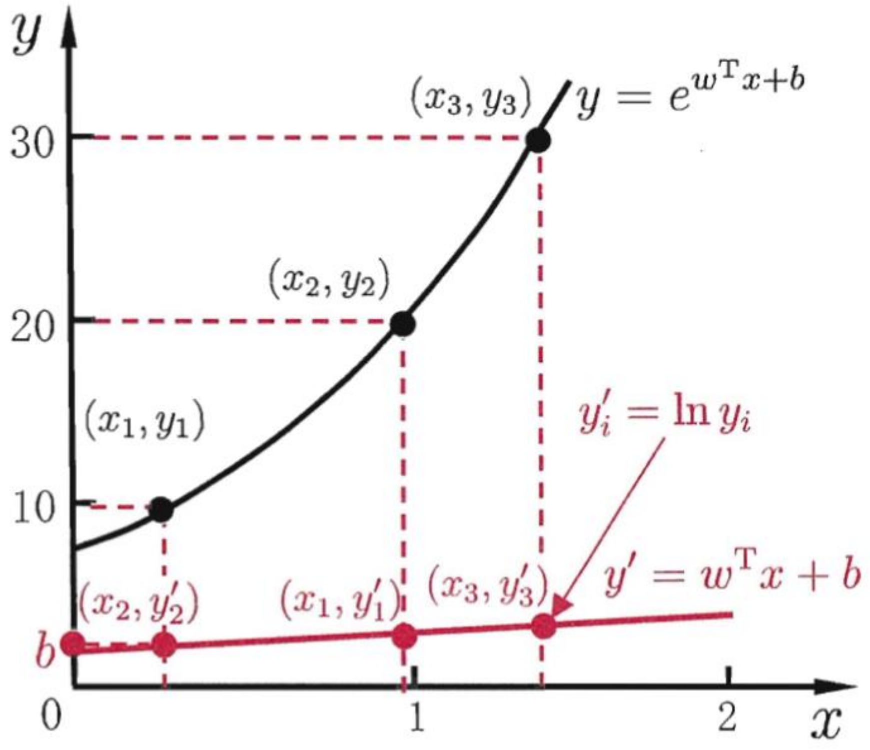
对于一般线性模型，若希望线性模型的预测值逼近真实标记，则得到线性回归模型：

$$y=\omega^T x+b$$

若令预测值逼近 y 的衍生物，例如 $\ln y=\omega^T x+b$

则得到对数线性回归(log-linear regression)

实际是在用 $e^{\omega^T x+b}$ 逼近 y



线性回归-广义线性模型

更一般地，考虑所有y的衍生物的情形，就得到了“广义的线性模型”（generalized linear model）

其中，g（*）称为联系函数

一般形式为 $y = \underbrace{g^{-1}}_{\text{单调可微的 联系函数}}(\omega^T x + b)$

单调可微的 ***联系函数***

对数线性回归是广义线性模型在 $g(\cdot) = \ln(\cdot)$ 时的特例，

即： $\ln y = \omega^T x + b$

对数几率回归

问题描述：分类问题，将单调可微函数将分类任务的真实标记y与线性回归模型的预测值联系起来

对于二分类任务，其输出标记$y \in \{0,1\}$，而线性回归模型产生的预测值$z= w^T x + b$是实值,于是,我们需将实值z转换为0/1值.最理想的是“单位阶跃函数”(unit-step function)

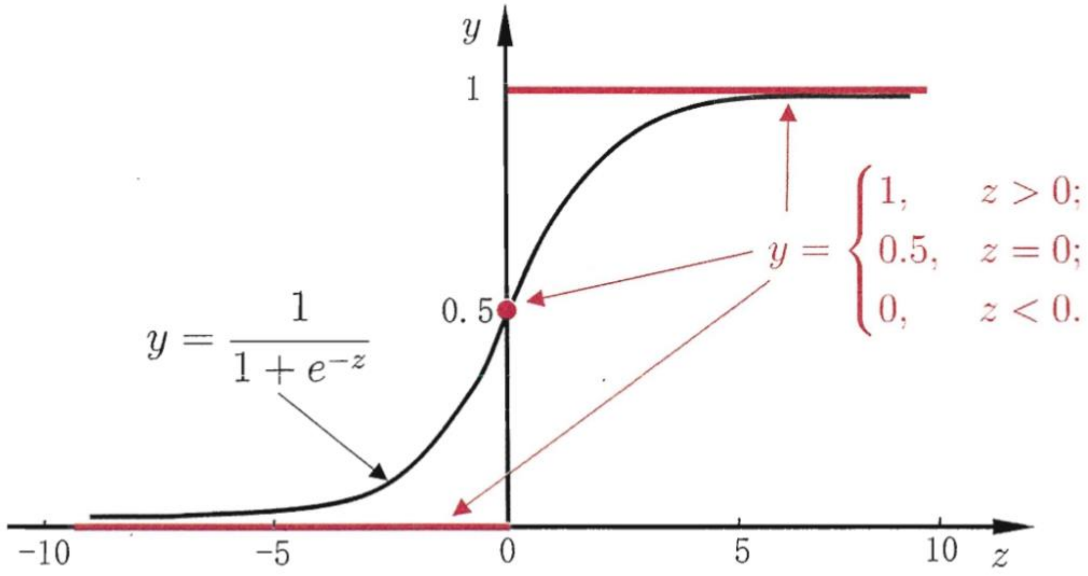
$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

若预测值z>0则为正例，否则为反例

对数几率回归

但是由于上述阶跃函数在零点处不可微，故使用对数几率函数来替代，即

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



单位阶跃函数与对数几率函数

对数几率回归

将 $f(x_i)=\omega^T x_i + b$　带入对数几率函数(即sigmoid函数) $y=\dfrac{1}{1+e^{-(\omega^T x+b)}}$

取对数可得

$$\ln\frac{y}{1-y}=\omega^T x+b$$

若将**y看做样本为正例的概率，（1-y）看做样本为反例的概率**，则上式实际上使用线性回归模型的预测结果器逼近真实标记的对数几率。因此这个模型称为“对数几率回归”（logistic regression）

- 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

求解方式

若将 y 看作类后验概率估计 $p(y=1|x)$ 则

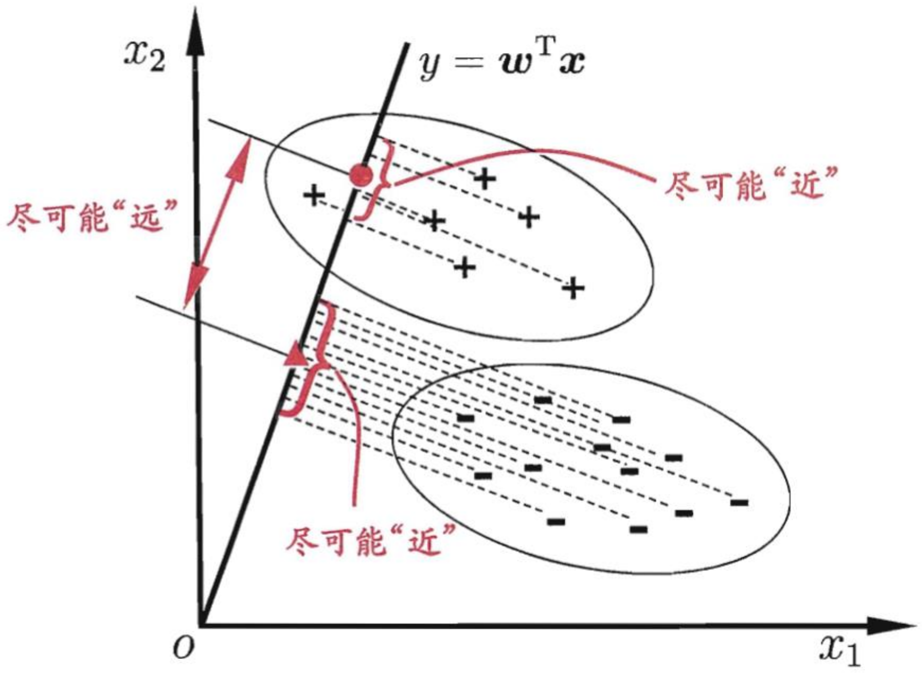
$$\ln\frac{y}{1-y}=\omega^T x+b \qquad \ln\frac{p(y=1|x)}{p(y=1|x)}=\omega^T x+b$$

通过极大似然估计法(maximum likelihood method)

给定数据集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ 最大化“对数似然”(log-likelihood)函数得

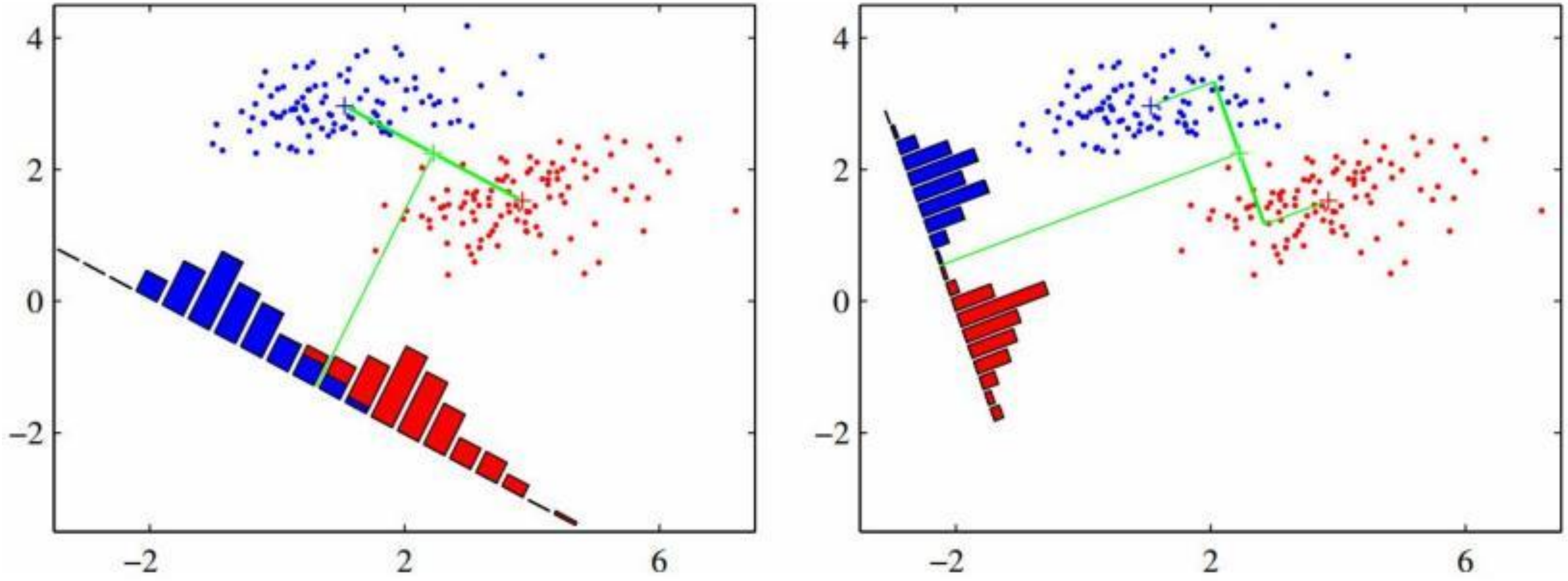
$$l(\omega,b)=\sum_{i=1}^{m}\ln p(y_i|x_i;\omega,b)$$

线性判别分析(Linear Discriminant Analysis)



由于将样例投影到一条直线（低维空间），因此也被视为一种“监督降维”技术

降维



LDA的目标

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

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 $\boldsymbol{\Sigma}_i$

两类样本的中心在直线上的投影：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$

两类样本的协方差：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$

同类样例的投影点尽可能接近 → $\omega^T \sum_0 \omega + \omega^T \sum_1 \omega$ 尽可能小

异类样例的投影点尽可能远离 → $\left\|\omega^T \mu_0 - \omega^T \mu_1\right\|_2^2$ 尽可能大

于是，变为最大化下式

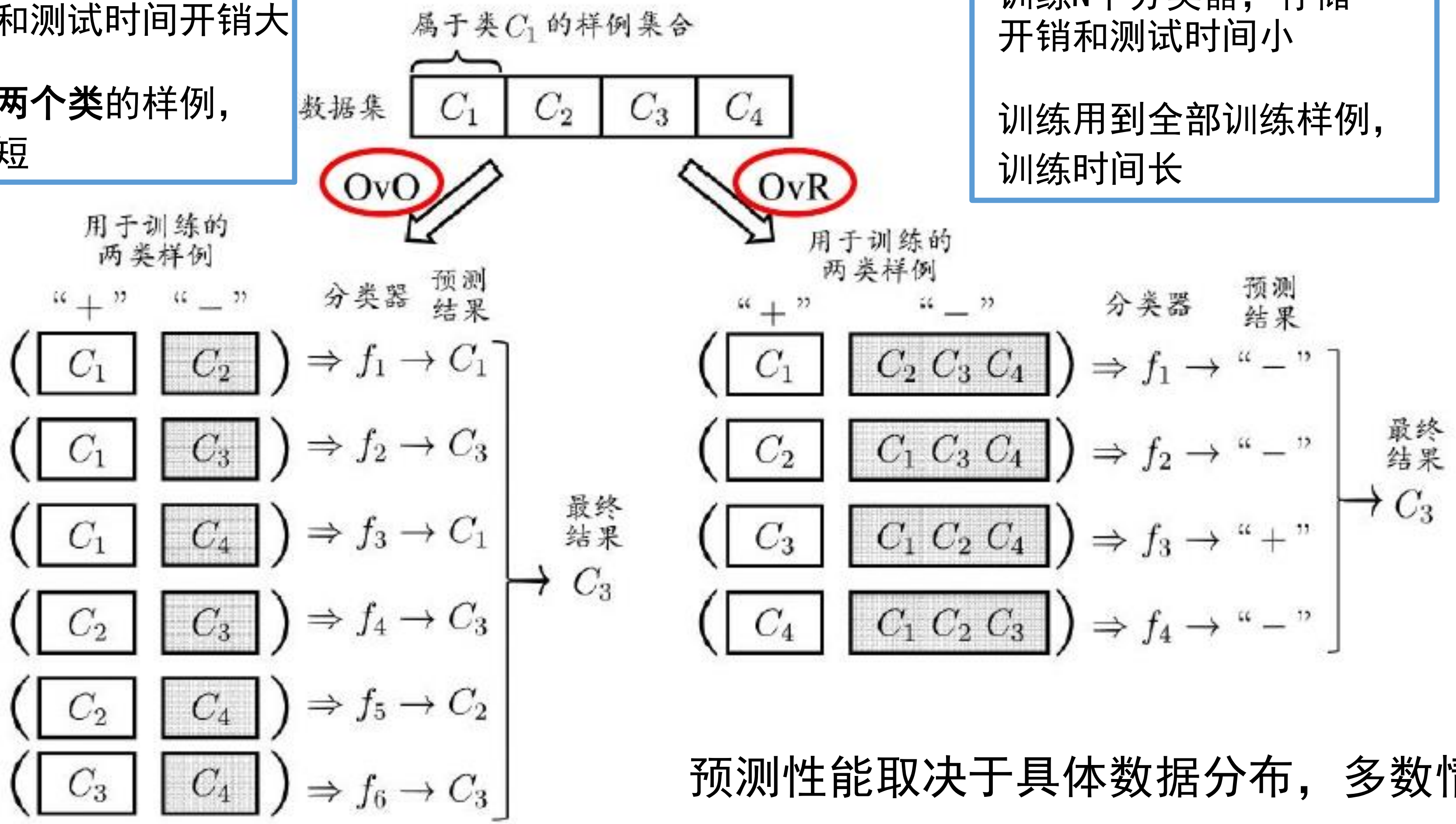
$$J = \frac{\left\|\omega^T \mu_0 - \omega^T \mu_1\right\|_2^2}{\omega^T \sum_0 \omega + \omega^T \sum_1 \omega} = \frac{\omega^T (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T \omega}{\omega^T (\sum_0 + \sum_1)\omega}$$

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

训练N(N-1)/2个分类器，
存储开销和测试时间开销大

训练只用**两个类**的样例，
训练时间短



训练N个分类器，存储
开销和测试时间小

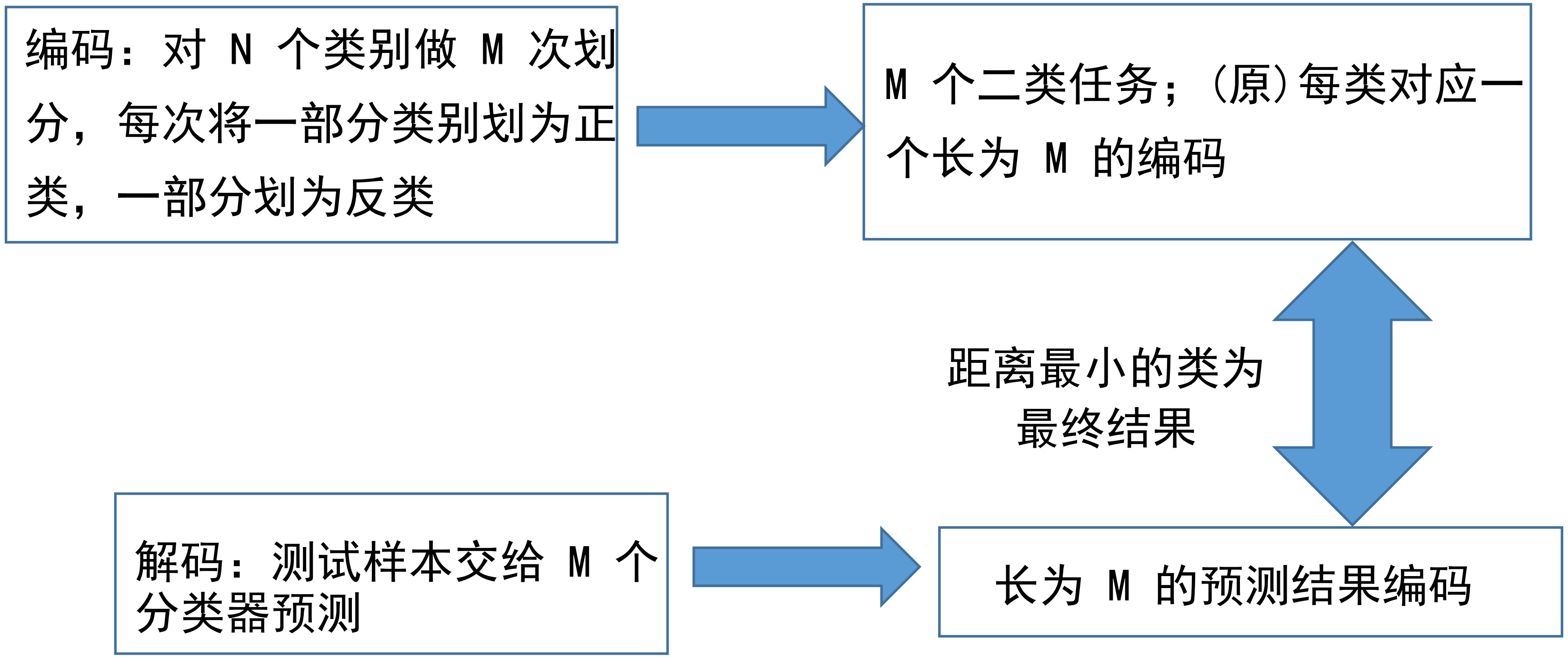
训练用到全部训练样例，
训练时间长

预测性能取决于具体数据分布，多数情况下两者相近

纠错输出码 (ECOC)

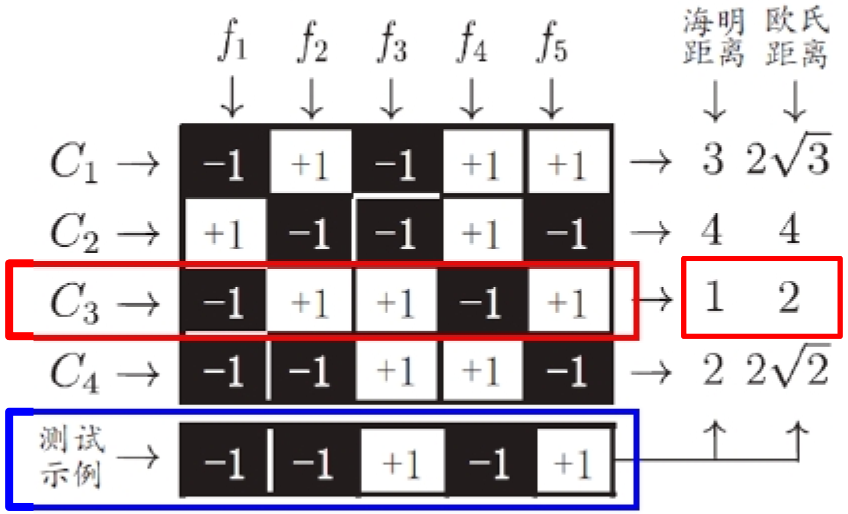
多对多(Many vs Many, MvM): 将若干类作为正类，若干类作为反类

一种常见方法：纠错输出码(Error Correcting Output Code)

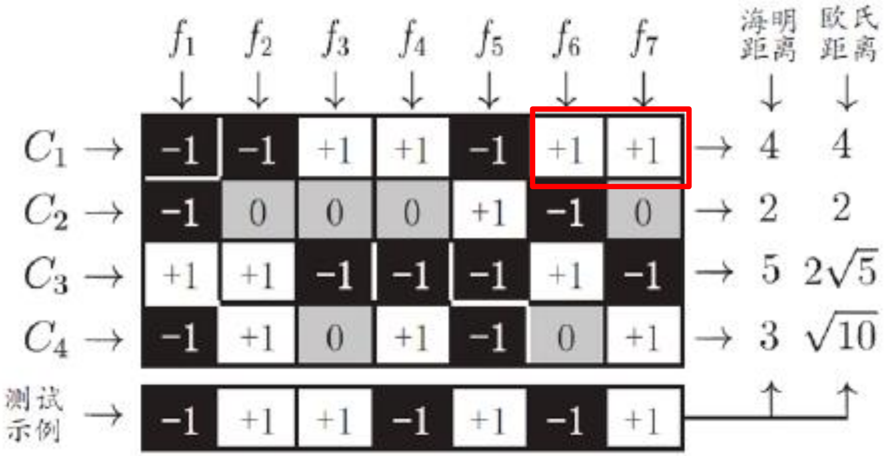


纠错输出码

计算海明距离的一种方法，就是对两个位串进行异或（xor）运算，并计算出异或运算结果中1的个数。例如110和011这两个位串，对它们进行异或运算，其结果是：110⊕011=101，异或结果中含有两个1，因此110和011之间的海明距离就等于2



(a) 二元 ECOC 码



(b) 三元 ECOC 码

- 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

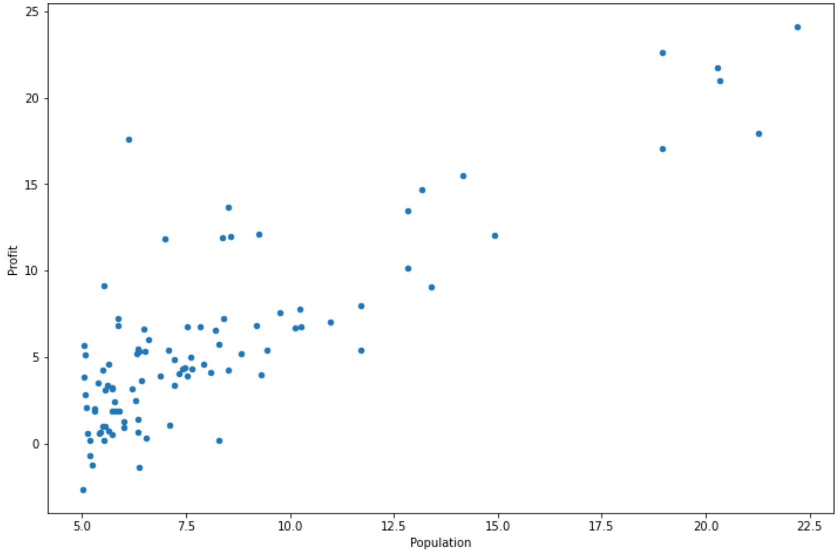
- 对同等长度的编码，理论上来说，任意两个类别之间的编码距离越远，则纠错能力越强

作业一:使用线性模型实现餐厅利润和房屋价格的预测

- 推荐编程环境:Anaconda+Jupyter notebook

  安装教程： 点这

- 卡车利润问题为一元线性回归，房屋价格预测为多元线性回归
- 实现数据以及预测函数的可视化，并展示预测结果



城市人口和利润关系图

房屋序号	房屋面积	卧室数量	成交价格
1	2104	3	399900
2	1600	3	329900
3	2400	3	369000
4	1416	2	232000

数据集二部分样本展示

参考文献

1. E. L. Allwein, R. E. Schapire and Y. Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers[J]. Journal of Machine Learning Research, 2000:113:141.
2. K. Crammer, Y. Singer. On the Learnability and Design of Output Codes for Multiclass Problem[J]. Machine Learning, 2002:201-233.
3. Z.H. Zhou, X. Y. Liu. On Multi-Class Cost-Sensitive Learning[C]. Proceedings of the 21th AAAI Conference on Artificial Intelligence (AAAI), 2006:567-572
4. Z.H. Zhou, X. Y. Liu. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem[J]. IEEE Transactions on Knowledge and Data Engineering,2006:63-77

相关论文会放到课程网页中，如有需要请自行下载。