第三章. 线性模型

3.1 基本形式

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_i在第i个属性上的取值，线性模型试图通过一个组合的线性组合来进行预测的函数，即

$$f(x)=w_1x_1+w_2x_2+...+w_dx_d+b$$

向量形式为

$$f(x)=w^Tx+b$$

其中$\omega=(w_1,w_2,.....w_d)$,w 和b学到之后，模型就能确定。

3.2 线性回归

一元线性回归:

给定一个属性描述的示例$x=(x_1;x_2;...;x_d)$

线性回归试图学得

$f(x_i)=wx_i+b$, 使得$f(x_i)\approx y_i$

确定w和b:关键在于衡量$f(x)$与y之间的差别，均方误差是回归任务中最常用的性能衡量，因此可以试图让均方误差最小化，即

$$(w^*,b^*)=argmin_(w,b)\sum_{i=1}^{m}(f(x_i)-y_i)^2=argmin_(w,b)\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

求解w和b最小化的过程，称为线性回归模型的最小二乘“参数估计”

分别对w和b求导:

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

分别令求导为零，得到最优解

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

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

多元线性回归

样本由d个属性描述，试图学得

$$f(x_i)=w^Tx_i+b,使得f(x_i)\approx y_i$$

这称为“多元线性回归”

同样利用最小二乘法来对w和b进行估计，把w和b吸收进入向量形式$\hat{w} = (w; b)$,相应地，把数据集D表示为一个$m * (d + 1)$大小的矩阵X，其中每行对应于一个示例，该行前d个元素对应d个属性值，最后一个元素置为1

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

再把标记也写成向量形式$y = (y_1; y_2; \ldots; y_m)$

$$\hat{w}^* = \hat{w}argmin_(\hat{w}) = (y - X\hat{w}^T)(y - \hat{w})$$

对$\hat{w}$求导

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

令导数为0，得最优解

$$\hat{w}^* = (X^T X)^{-1} X^T y$$

其中$X^T X$得为满秩矩阵或正定矩阵，否则不成立

要解决非满秩矩阵，需引入正则化。

对数线性回归

让输出标记在指数尺度上变化，那就可将输出标记的对数作为线性模型逼近的目标，即

$$lny = w^T x + b$$

这就是“对数线性回归”，试图让e^{w^T+b}逼近y,形式上仍是线性回归，但实质上已是在求取输入空间到输出空间的非线性函数映射，这里的对数函数起到了将线性回归模型的预测值与真实标记联系起来的作用。

更一般的，将y变成单调可微函数g(y),

$$y = g^{-1}(w^T x + b)$$

3.3 对数几率回归

二分类任务

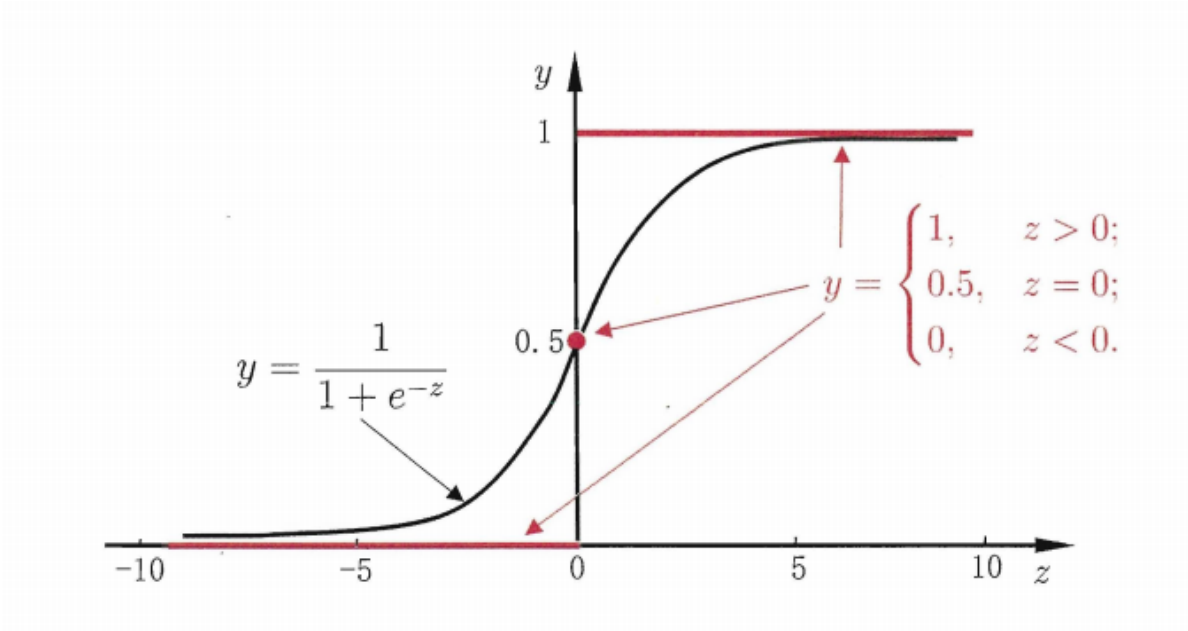
对于二分类任务，其输出标记为$y \in${0,1},但线性回归模型产生的预测值$z = w^T x + b$是实值，故需要找到激活函数使之转换

单位跃迁函数

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

但函数不连续，故又找到对数几率函数

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



再将对数几率函数代入广义线性模型

$$y=\frac{1}{1+e^{-(w^Tx+b)}}$$

变形

$$ln\frac{y}{1-y}=w^Tx+b$$

将y作为样本x的正例，则1-y为其反例的概率，两者比值

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

称为几率，反映了x作为正例的相对可能性，对几率取对数则得到“对数几率”

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

实际上是用线性回归模型去逼近真实标记的对数几率

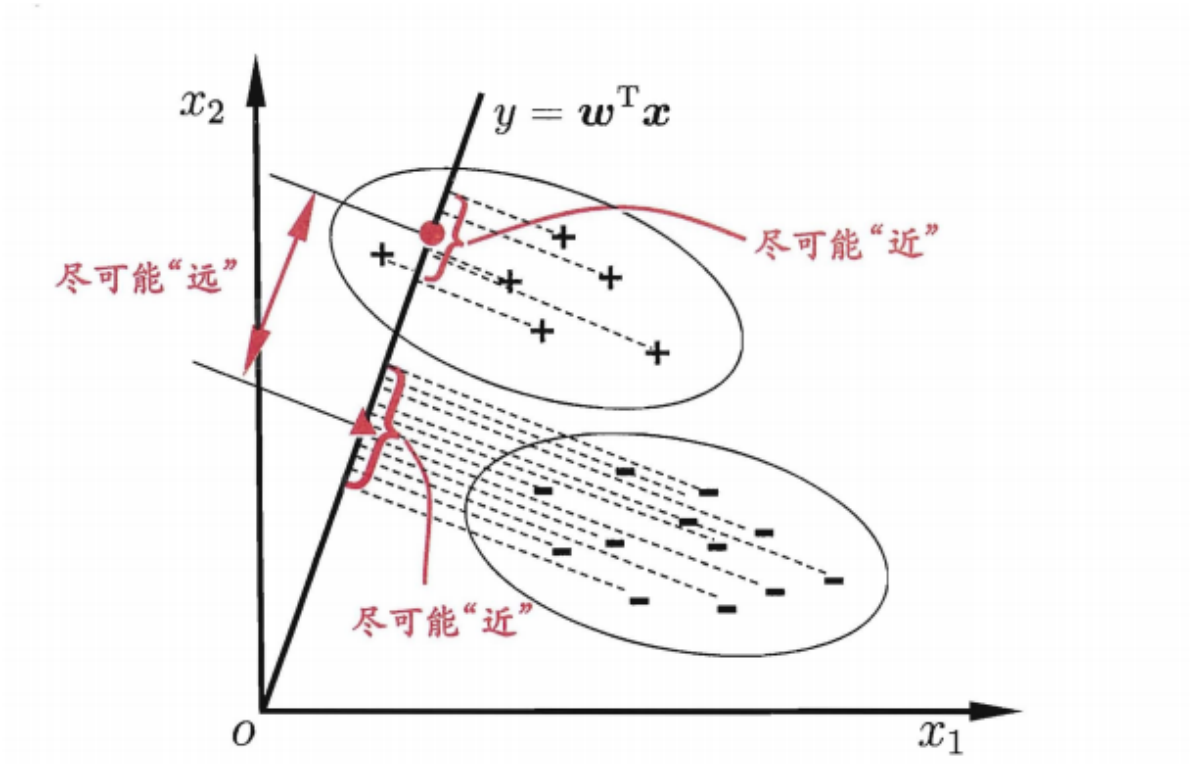
对数几率回归优点

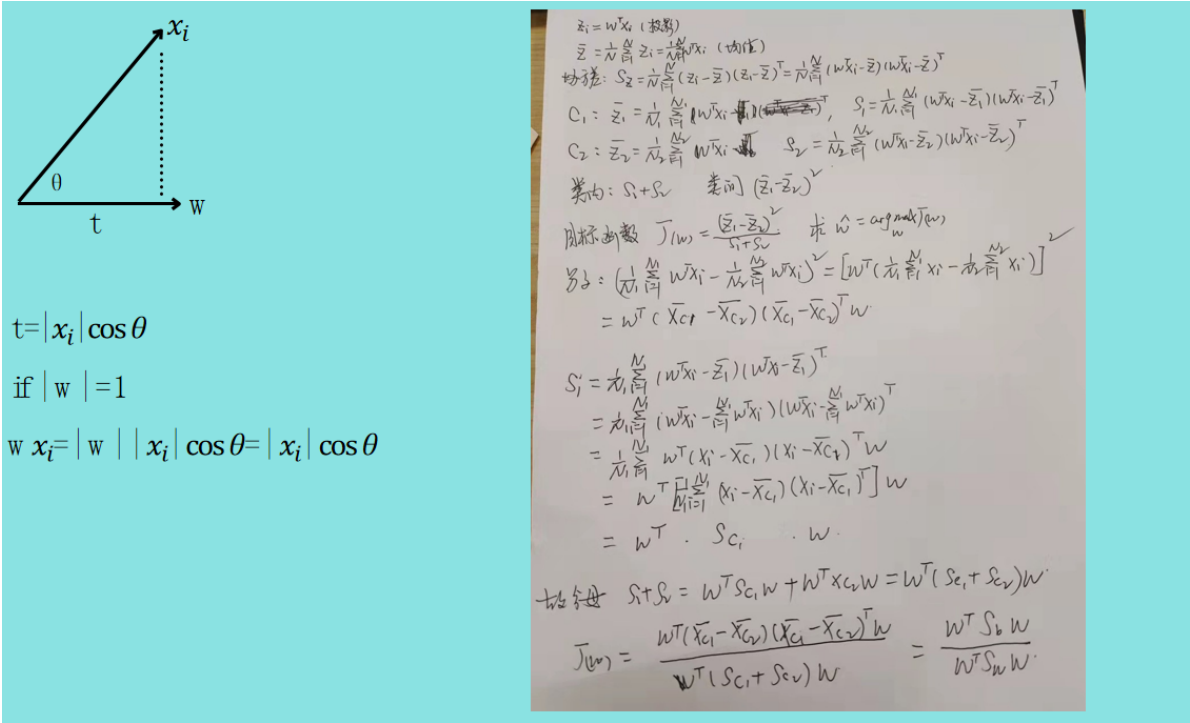
- 直接对分类可能性建模
- 无需事先假设数据分布，避免了假设分布不准确所带来的的问题
- 不仅是预测出"类别"，而是可达到近似概率，这对许多利用辅助决策的任务很有用
- 目标函数是任意阶可导的凸函数，有很好的数学性质，现有的许多数值优化算法都可用于求解最优解

3.4 线性判别分析

LDA思想

给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近，异类样例的投影点尽可能远离；在对新样本进行分类时，将其投影到同样的这样条直线上，再根基投影点的位置来确定新样本的类别。**(类内小，类间大)**（方差）





3.5 多分类学习

不失一般性，考虑N个类别$C_1, C_2, \ldots, C_N$,多分类学习的基本思路是“拆解法”，即将多分类任务拆分为若干个二分类任务求解。

拆分策略

OvO: 将N个类别两两配对，从而产生N(N-1)/2个二分类任务，C_i作为正例，C_j作为反例训练一个分类器，从而产生N(N-1)/2个分类结果，把被预测最多的类别作为最终分类结果。

OvR: 每次将一个类的样例作为正例，所有其他样例作为反例来训练N个分类器。在测试时若仅有一个分类器预测为正例，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例，则通常考虑分类器的预测置信度，选择置信度最大的类别标记作为分类结果。

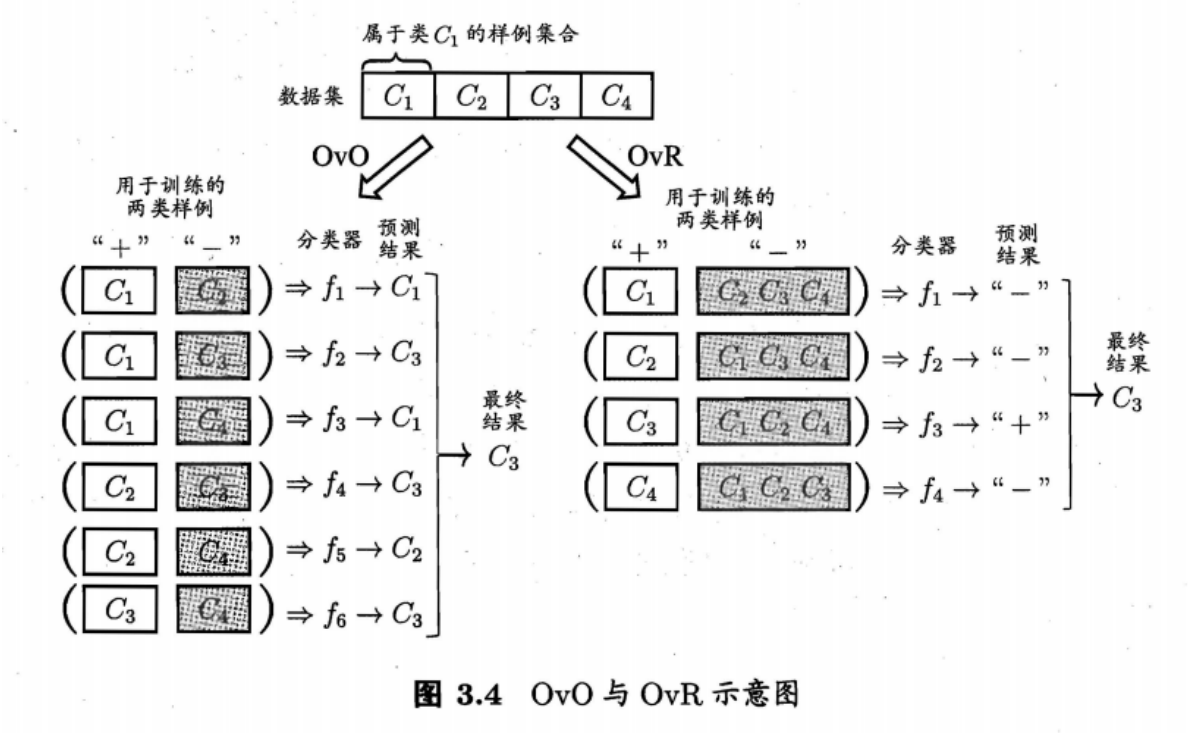


图 3.4 OvO 与 OvR 示意图

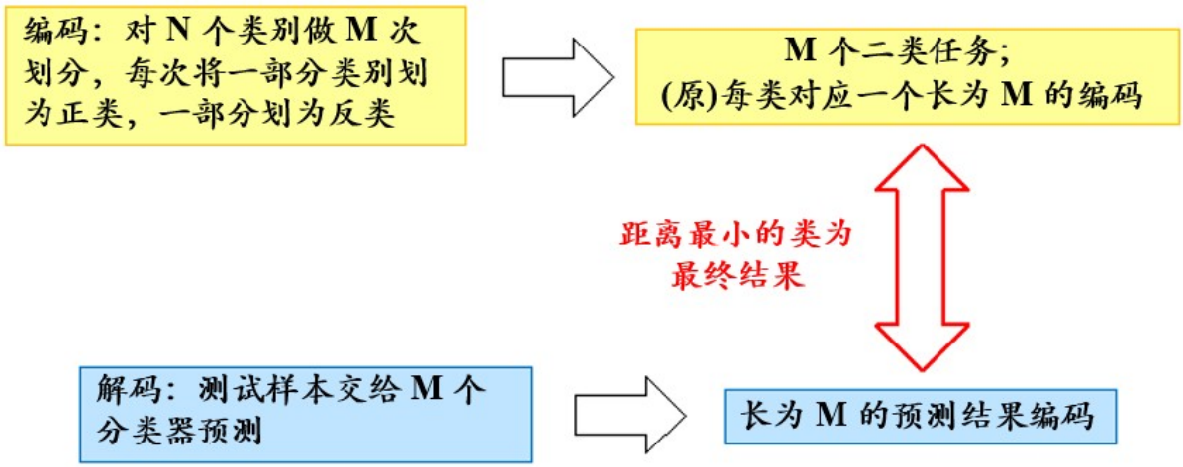
OvO vs OvR:

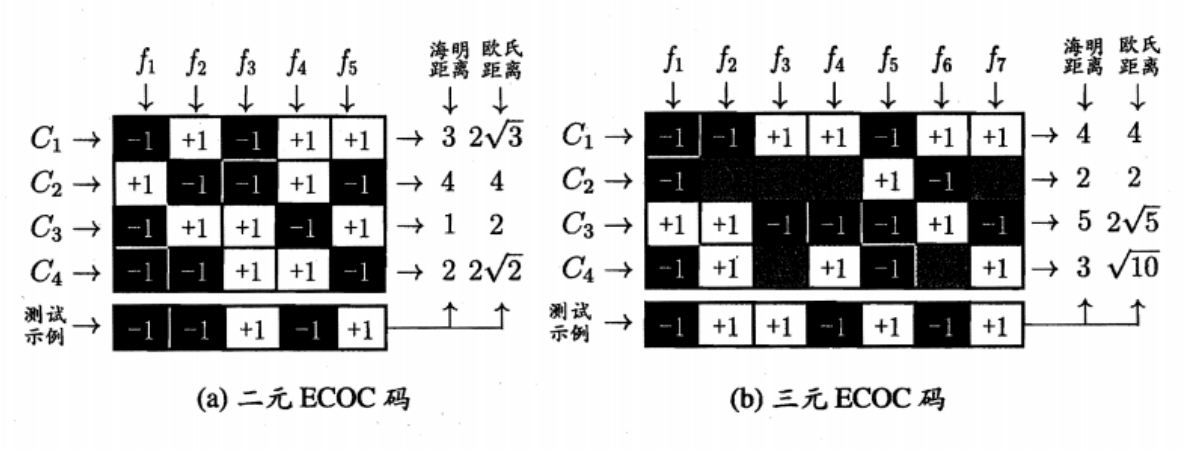
- OvR秩序训练N个分类器，而OvO需要训练N(N-1)/2分类器，因此OvO的存储开销和测试时间开销比通常OvR大。
- 训练时，OvR的每个分类器均使用全部训练样例，而OvO的每个分类器仅用到两个类的样例，因此，在类别很多时，OvO的训练时间开销通常比OvR更小
- 预测性能，取决于具体的分布数据，大多数情况都差不多。

MvM: 每次将若干个类作为正类，若干其他类作为反类，显然OvO和OvR是其特例

由于MvM正反例不能随意选取，MvM技术“纠错输出码”

ECOC





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

3.6 类别不平衡问题

类别不平衡是指分类任务中不同类别的训练样例数目差别很大的情况，在通过拆解法解决对分类问题时，即使原始问题中不同类别训练样例数 目相当，在使用OvR、MvM策略后 仍可能出现类别不平衡现象。

从线性分类器的角度理解，用$y = w^T x + b$对新样本x进行分类时，事实上在预测出的y与一个阈值进行比较，几率$\frac{y}{1-y}$则反映了正例可能性与反例可能性之比，阈值设为0.5表面真实正，反例可能性相同，则分类器规则为

$$若\ \frac{y}{1-y} > 1，则预测为正例$$

若正反例数目不同，令m^+表示正例数目，m^-表示反例数目，则观测几率是$\frac{m^+}{m^-}$,但通常假设训练集是真实样本总体的无偏采样，因此观测几率就代表了真实几率

$$若\ \frac{y}{1-y} > \frac{m^+}{m^-}，则预测为正例$$

基本策略----再缩放

$$\frac{y'}{1-y'} = \frac{y}{1-y} * \frac{m^-}{m^+}$$

由于“训练集是真实样本总体的无偏采样”这一假设往往不成立，即未必有效地基于训练集观测几率来推断真实几率。

解决办法

- 欠采样：去除一些样例多的样例，使正反例数目接近，然后在进行学习
- 过采样：增加一些样例少的样例，使正反例数目接近，然后在进行学习
- 阈值移动：基于原始训练集进行学习，但在用训练好的分类器预测时，将再缩放嵌入其决策过程中。

方法评价

方法	优点	缺点
随机过采样	不会造成信息缺失,表现优于欠采样	过采样会引起噪声数据的权重过大，也会加大过拟合的可能性
随机欠采样	减少运行时间,并且当数据集很大时,可以通过减少样本数量来解决存储问题	可能会丢失重要信息,采样后的数据不一定能代表全部数据，导致分类结果不精准
阈值移动	不会造成信息缺失，过拟合的可能性小	很难建立参数与不平衡分类精度间的定量关系,不能准确处理不平衡数据
调整代价或权重	不会造成信息缺失,过拟合的可能性小	很难建立参数与不平衡分类精度间的定量关系，不能准确处理不平衡数据