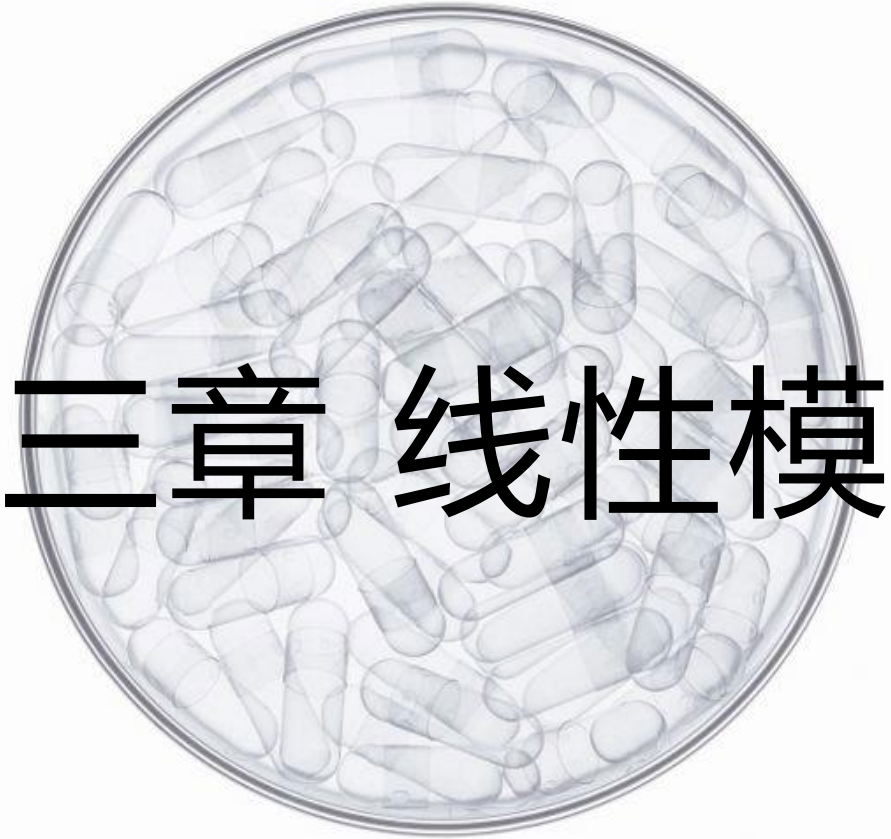
机器学习



第三章 线性模型

汇报人：杜昭武

目录

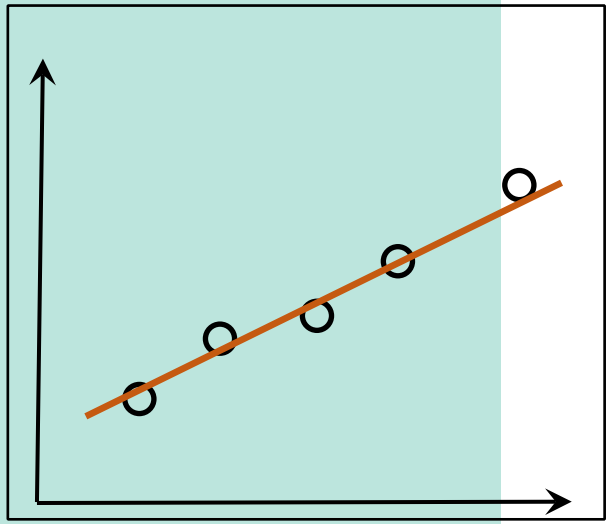
1. 基本形式
2. 线性回归
3. 对数几率回归
4. 线性判别分析
5. 多分类学习
6. 类别不平衡问题

一、基本形式

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，即向量形式为

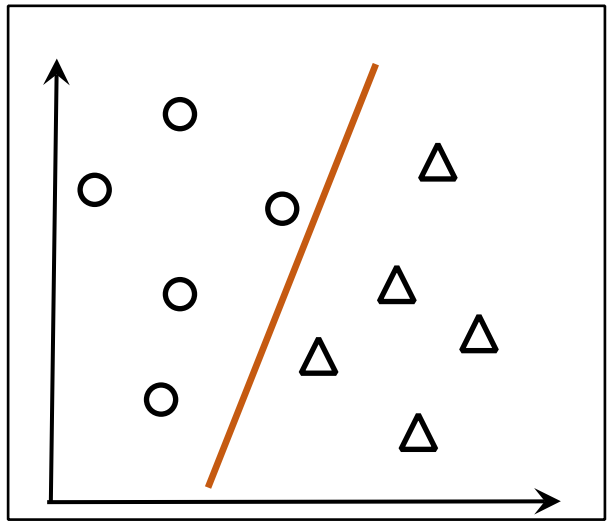
$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

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



回归

- 线性回归



分类

- 对数几率
- 线性判别分析
- 多分类学习

二、线性回归

1. **假设属性数目只有一个**：线性回归试图学得

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

非数值属性的变换问题：离散属性的处理：若有“序”(order)，则连续化；否则，转化为kk维向量

- 若属性值之间存在“序”的关系，可通过连续化将其转换为连续值，例如二值属性“身高”的取值“高”“矮”可转化为{1.0，0.0}
- 若属性之间不存在序的关系。假定有k个属性，则通常转换为k维向量，例如属性“瓜类”的取值，”西瓜”“黄瓜”“南瓜”可转化为（0,0,1） （0,1,0） （1,0,0）

令均方误差最小化，有

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

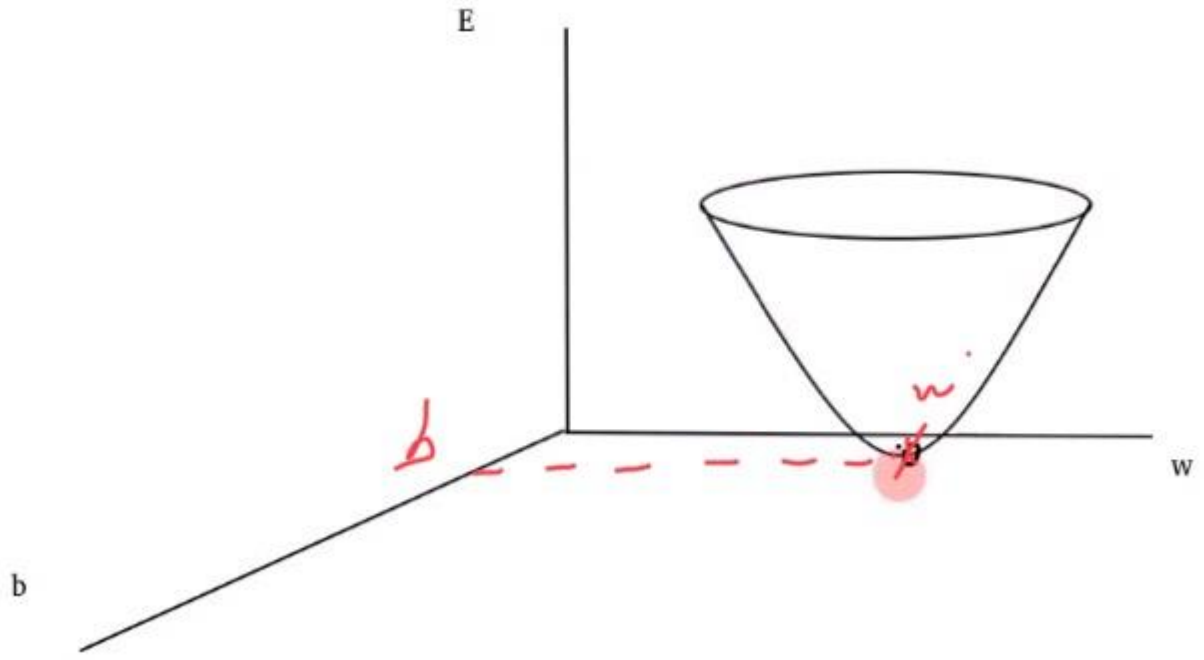
二、线性回归

$$E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$$ 用最小二乘进行估计

这里需要注意的是，w和b是未知数；x和y 是常数。选择合适的w，b使得均方误差最小，将式子展开得到

$$E(w,b)=\Sigma_{i=1}^{m}(y_i-wx_i-b)^2=\Sigma_{i=1}^{m}(x_i^2w^2+y_i^2+b^2-2y_ix_iw-2y_ib-2wbx_i)$$

$$E(w,b)=(\Sigma x_i)w^2+mb^2+(-2\Sigma(x_iy_i))w+(-2\Sigma y_i)b+(-2\Sigma x_i)wb+\Sigma y_i^2$$



二、线性回归

分别对 w 和 b 求导:

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

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

令导数为 0, 得到闭式(closed-form)解:

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

二、线性回归

2. **一般的情形数据集D，样本由d个属性描述**：此时线性回归试图学得

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

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

把 w和 b吸收入向量形式， 数据集表示为

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} * \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{pmatrix} = \begin{pmatrix} f(\vec{x}_1) \\ f(\vec{x}_2) \\ \vdots \\ f(\vec{x}_m) \end{pmatrix}$$

二、线性回归

同样用最小二乘法求解

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

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

附录A

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

令其为零可得 $\hat{\boldsymbol{w}}$

$$2X^T(X\hat{w}-y)=0$$

$$2X^TX\hat{w}=2X^Ty$$

$$\hat{w}=(X^TX)^{-1}X^Ty$$

- 若 满秩或正定，令上式等于0则可解出 $\hat{\boldsymbol{w}}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$
- 若 不满秩，则可解出多个 $\hat{\boldsymbol{w}}$

此时需要求助于归纳偏好，或引入 正则化 (regularization) 第6、11章

三 、对数几率回归

1. 二分类任务：

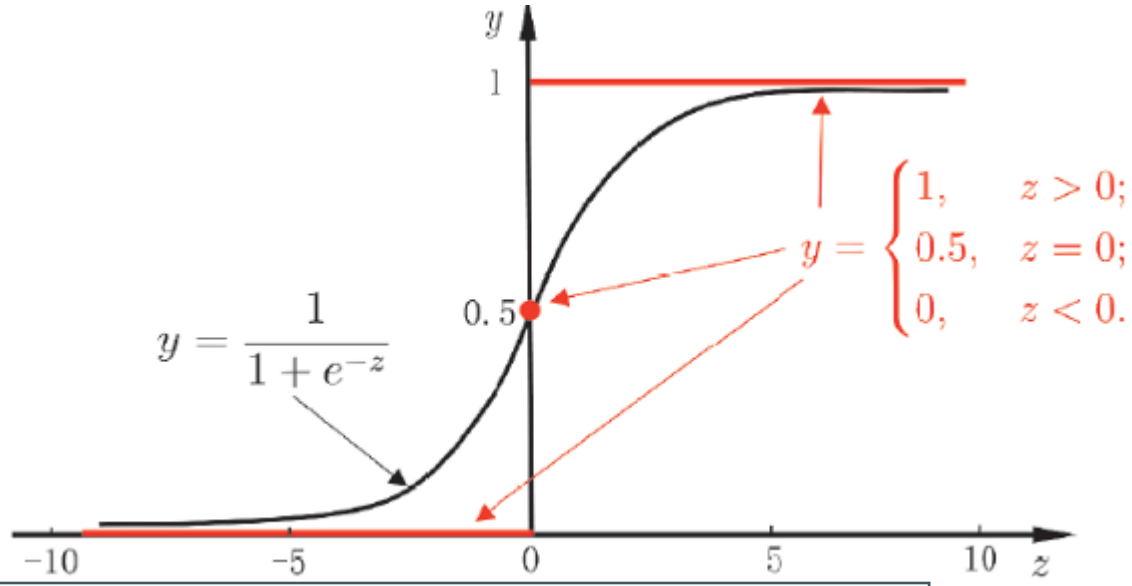
线性回归模型产生的实值输出 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$

期望输出 $y \in \{0, 1\}$

找 z 和 y 的联系函数

理想的“单位阶跃函数” （unit-step function）

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



- 阶跃函数性质性质不好，需要找“替代函数”（surrogate function ）
- 常用单调可微且任意阶可导的函数为

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

对数几率函数（logistic function）简称“对率函数”

三 、对数几率回归

以对率函数为联系函数 ：

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

即：$$\ln\frac{y}{1-y} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$$

将y视为样本x作为正例的可能性，则1-y 是其反例可能性，例如Y 为看作“好瓜”的概率，1-y则是看作“坏瓜”的概率。Y与1-y 的比值称为”几率”（odds），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”（log odd,亦称为logit)

“对数几率回归”(logistic regression) 简称“对率回归”

- 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

三、对数几率回归

求解思路：

若将 y 看作类后验概率估计 $p(y=1 \mid \boldsymbol{x})$ 则

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

解的

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

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

于是，使用"极大似然法" (maximum likelihood method)

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

$P(y_1|\vec{x}_1)*P(y_2|\vec{x}_2)*...*P(y_m|\vec{x}_m)$ |

对率回归模型最大化"对数似然" (log-likelihood)函数

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

三、对数几率回归

求解思路：

令 $\boldsymbol{\beta}=(\boldsymbol{w};b)$ $\hat{\boldsymbol{x}}=(\boldsymbol{x};1)$ $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ 可简写为 $\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$

再令 $p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)=p\left(y=1\mid\hat{\boldsymbol{x}};\boldsymbol{\beta}\right)=\dfrac{e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}$

$$p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)=p\left(y=0\mid\hat{\boldsymbol{x}};\boldsymbol{\beta}\right)=1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}$$

则似然项可重写为 $p\left(y_i\mid\boldsymbol{x}_i;\boldsymbol{w}_i,b\right)=y_ip_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)+\left(1-y_i\right)p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)$

于是，最大化似然函数 $\ell(\boldsymbol{w},b)=\sum_{i=1}^{m}\ln p(y_i\mid\boldsymbol{x}_i;\boldsymbol{w},b)$

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

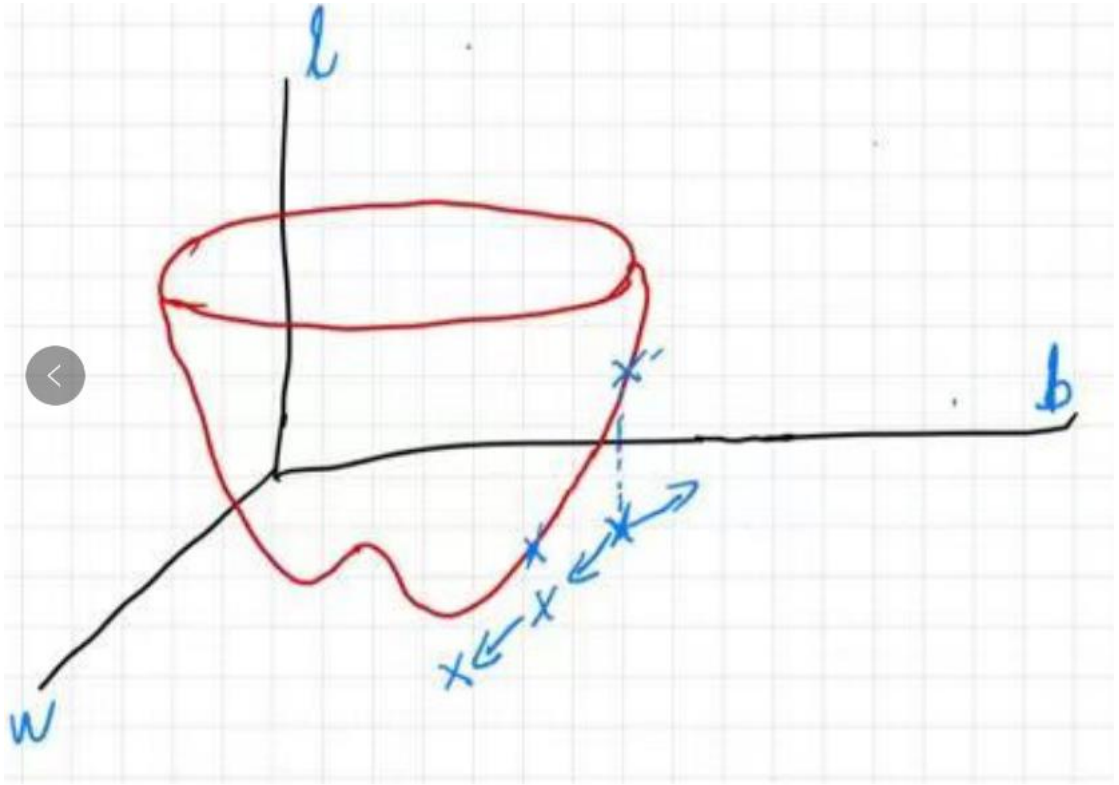
三 、对数几率回归

求解思路：

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 [Boyd and Vandenberghe, 2004]

过程：假设只有两个参数的情形下，数值化寻找极小点

1. 取曲面上随机点，计算梯度
2. 朝着梯度下降的方向走一步，
3. 在新的点上计算梯度
4. 最终有一定概率找到极小点



四 、线性判别分析

线性判别分析（LDA）的思想非常朴素：给定训练集例集，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。
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

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$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$

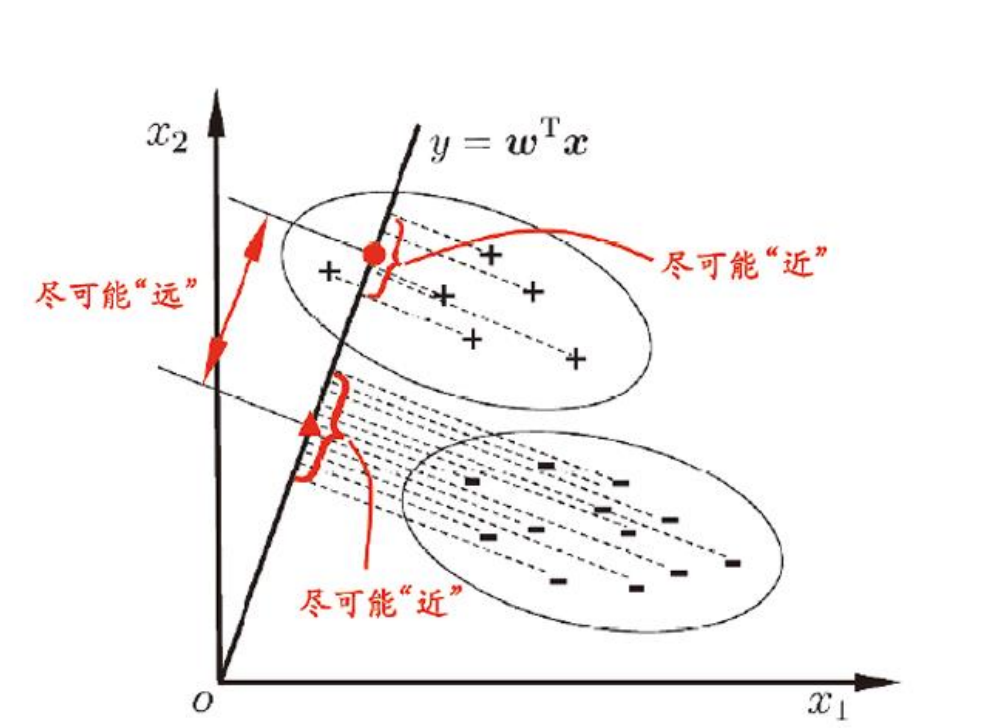
同类样例的投影点尽可能接近 → $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小

异类样例的投影点尽可能远离 → $\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2$ 尽可能大

于是，最大化

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

$$\Sigma_0=\sum(\mathbf{x}-\mu_0)(\mathbf{x}-\mu_0)^T$$
$$w^T\Sigma_0 w=\sum w^T(\mathbf{x}-\mu_0)(\mathbf{x}-\mu_0)^T w$$



四、线性判别分析

求解思路：

类内散度矩阵 (within-class scatter matrix)：

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

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

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

LDA的目标：最大化广义瑞利商 (generalized Rayleigh quotient)：

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

四、线性判别分析

求解思路：

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

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

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

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

运用拉格朗日乘子法，有 $\mathbf{S}_b\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$

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

注意到 $(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}$ 标量，令其等于 λ

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

拉格朗日乘子法：

$$\min(-w^T S_b w+\lambda(w^T S_w w-1))$$

$$\partial[-w^T S_b w+\lambda(w^T S_w w-1)]/\partial w$$

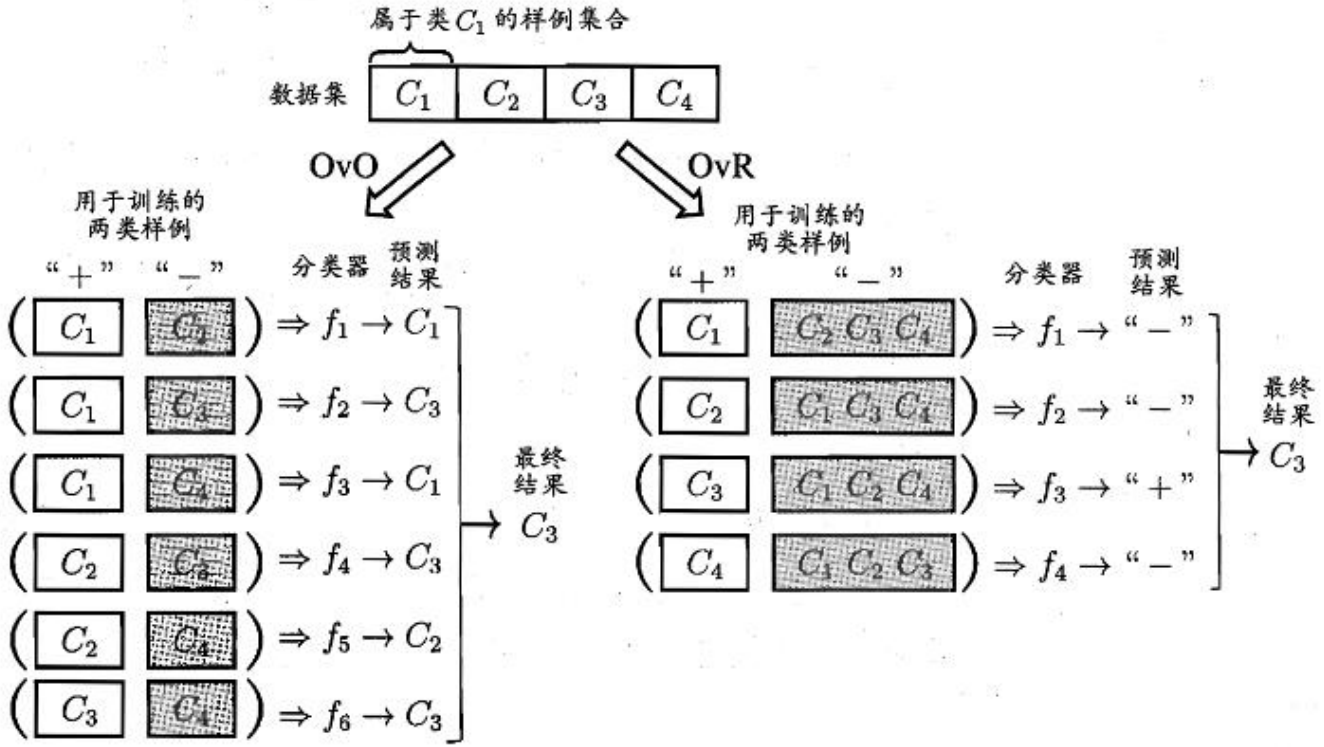
$$=S_b w-\lambda S_w w=0$$

五、多分类学习

基本思路:“拆解法”，将多分类学习任务拆为若干个二分类任务求解，最经典的三种拆分策略:“一对一”(OvO)、“一对其余”(OvR)和“多对多”(MvM)

OVO:
给定数据集
D={(x1,y1),(x2,y2),…(xm,ym)},yi∈(C1,C2,…CN),OVO将这N个类别两两配对，从而产生N(N-1)/2个分类任务，最终结果可以通过投票产生:
把被预测得最多的类别作为最终分类结果。

OVR:
则是每次将一个类的样例作为正例，所有其他的样例作为反例来训练N个分类器。在测试时若只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。



五、多分类学习

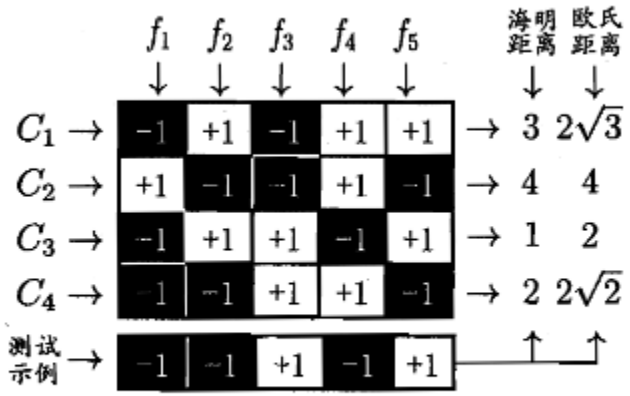
MVM技术

纠错输出码（Error Correcting Output Codes，ECOC），ECOC是将编码的思想引入类别拆分，并尽可能在编码过程中具有容错性。ECOC工作过程主要分为两步：

编码：对N个类别做M次划分，每次划分将一部分类别化为正类，另一部分分为反类，产生M个训练集——M个分类器。

解码：M个分类器分别预测，这些预测标记组成一个编码，将其与每个类别的编码比较，区别最小的就是最终结果。

类别划分通过“编码矩阵”指定，即如下二元阵：将类别指定为正类反类



(a) 二元 ECOC 码

欧式距离：在a图中分类器f2将C1和C3类的样例作为正例；若基于欧式距离，预测结果是C3。以测试集和第一行为例，sqrt[（1-（-1））2+（-1-1））2+（1-（-1））2]=sqrt(12)

海明距离：两个合法代码对应位上编码不同的位数为码距，又称海明码，例如测试示例：101010，分类示例：100101，此时海明距离为4

六、类别不均衡问题

定义：类别不平衡，就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

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

用上式对新样本分类，实际上在用预测的y值与一个阈值比较。若 $\frac{y}{1-y} > 1$，则预测为正例

然而，正反例数目不同时，观测几率m+/m−，由于我们假设无偏采样，则观测几率就代表真实几率，于是若 $\frac{y}{1-y} > \frac{m^+}{m^-}$，则预测为正例

原来的分类器基于上式进行决策，所以需要进行缩放操作，即 $\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^+}{m^-}$

但是很难保证“训练集是真实样本总体的无偏采样”。现有技术有三种做法：

- “欠采样”：直接除去一些反例
- “过采样”：增加一些正例
- “阈值移动“：不增不减，但是预测时采取 $\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^+}{m^-}$

七、课后习题

3.1 试析在什么情形下式(3.2)中不必考虑偏置项 b.

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

偏置项在数值上代表了当自变量取0时，因变量的取值，也就是拟合直线的截距，加偏置项是为了更好的拟合数据。
1.当学习到的模型恰好经过原点时，可以不考虑偏置项b
2.如果对数据进行了归一化处理，即对目标变量减去均值向量，此时就不需要考虑偏置项了

3.2 试证明, 对于参数 $\boldsymbol{w}$, 对率回归的目标函数(3.18)是非凸的, 但其对数似然函数(3.27)是凸的.

$$y = \frac{1}{1 + e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}. \tag{3.18}$$

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

3.18 $y = \frac{1}{1+e^{-(w^Tx+b)}}$ 证明是非凸即.

求二阶导 $y''<0$. 则是非凸即

$$\frac{dy}{dw} = \frac{e^{-(w^Tx+b)}\cdot x}{[1+e^{-(w^Tx+b)}]^2}$$

$$\frac{d^2y}{dw^2} = \frac{-x^2\cdot e^{-(w^Tx+b)}[1+e^{-(w^Tx+b)}]^2 - 2x\cdot e^{-(w^Tx+b)}[1+e^{-(w^Tx+b)}]}{[1+e^{-(w^Tx+b)}]^4} < 0$$

y是非凸的

3.27 $l(\beta) = \sum_{i=1}^{m}(-y_i\beta^T\hat{x}_i + \ln(1+e^{\beta^T\hat{x}_i}))$ $\beta=(w;b)$

$$\frac{\partial^2 l}{\partial\beta\partial\beta^T} = \sum_{i=1}^{m}\hat{x}_i\hat{x}_i^T p(\hat{x}_i;\beta)(1-p_1(\hat{x}_i;\beta)) > 0$$

(3.31式)

y是凸即.

七、课后习题

3.6 线性判别分析仅在线性可分数据上能获得理想结果，试设计一个改进方法，使其能较好地用于非线性可分数据

对于非线性可分的数据，要想使用判别分析，一般思想是将其映射到更高维的空间上，使它在高维空间上线性可分进一步使用判别分析。

3.7 令码长为 9，类别数为 4，试给出海明距离意义下理论最优的 ECOC 二元码并证明之.

类别数为4，因此1V3有四种分法，2V2有六种分法，3V1同样有四种分法。按照书上的话，理论上任意两个类别之间的距离越远，则纠错能力越强。那么可以等同于让各个类别之间的累积距离最大。对于1个2 2分类器，4个类别的海明距离累积为4；对于3V1与1V3分类器，海明距离均为3，因此认为2V2的效果更好。因此码长为9，类别数为4的最优EOOC二元码由6个2V2分类器和3个3v1或1v3分类器构成。

七、课后习题

3.8* ECOC 编码能起到理想纠错作用的重要条件是: 在每一位编码上出错的概率相当且独立. 试析多分类任务经 ECOC 编码后产生的二类分类器满足该条件的可能性及由此产生的影响.

重要条件为：①出错概率相当；②出错概率相互独立
①书上p66说，将多个类解分为2个“类别子集”，所形成的两个类别子集区分难度往往不同，即其导致的二分类问题难度不同。所以每个编码拆解后类别之间差异越相同，区分难度越相当，则满足条件①的可能性越大。但其实很难实现。
②一个好的纠错输出码应该满足条件“各个位上的分类器相互独立”。当类别越多时，满足条件②的可能性越大。
影响：一个理论纠错性质很好但二分类问题较难的编码，与一个理论纠错性质差一些但导致的二分类问题较简单的编码，难分孰强孰弱。

3.9 使用 OvR 和 MvM 将多分类任务分解为二分类任务求解时, 试述为何无需专门针对类别不平衡性进行处理.

对OvR和MvM来说，由于每个类进行了相同的处理，其拆解出来的二分类任务中类别不平衡的影响会相互抵消，因此通常不需专门处理

谢谢！