第三章：Linear Models

目 录

- **Linear Regression**
  - 最小二乘法
- **Binary Classification**
  - 对数几率回归
  - 线性判别分析
- **Multi-Class Classification**
  - 一对一
  - 一对其余
  - 多对多
- 类别不平衡问题

基本形式

- **Linear regression model 一般形式**

给定由 d 个属性描述的示例 $\boldsymbol{x}=(x_1;x_2;\ldots;x_d)$;

线性模型试图学到一个通过属性的线性组合来进行预测的函数

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

其中 x_i 是 $\boldsymbol{x}$ 在第 $\boldsymbol{i}$ 个属性上的取值

- **一般写成向量形式:**

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

其中 $\boldsymbol{w}=(w_1;w_2;\ldots;w_d)$. $\boldsymbol{w}$ 和 b 学得之后, 模型就得以确定.

线性模型优点

- 形式简单、易于建模

  蕴含着机器学习的重要思想

- 非线性模型的基础

  许多功能强大的非线性模型可在线性模型基础上获得，例如

  - 引入层级结构或高维映射

- 可解释性

  w直观的表达了各个属性在预测中的重要性，例如：

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

  - 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

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

- 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

- **考虑最简单的情况：输入属性只有一个(忽略下标)**

$$D = \{(x_i, y_i)\}_{i=1}^{m},\ 其中\ x_i \in \mathbb{R}.$$

线性回归

$D = \{(x_i, y_i)\}_{i=1}^m$, 其中 $x_i \in \mathbb{R}$.

离散属性：

若属性间存在：

有“序”关系

可通过连续化转化为连续值，例二值属性：

“身高”的取值“高”“矮”可转化为{1.0，0.0}

若无“序”关系

有k个属性值，则转换为k维向量，如属性的“瓜类”的取值：

西瓜、南瓜、黄瓜，可以转化为:

(0,0,1), (0,1,0), (1,0,0)

线性回归

- 单一属性的线性回归目标 $f(x) = wx_i + b$

使得 $f(x_i) \simeq y_i$

如何确定w,b ???

关键：**如何衡量f(x)与y的差别**

均方误差是回归任务中最常用的性能度量，

所以最小化均方误差

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

线性回归

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

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

用**均方误差最小化**来进行模型求解的方法**称为“最小二乘法”**

线性回归中，最小二乘法就是：

试图找到一条直线， 使得所有样本到直线上的欧式距离之和最小。

求解 w 和 b 使 $E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$ 最小化的过程

----称为线性回归模型的最小二乘参数估计（parameter estimation）

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

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

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

然后令式(3.5)和(3.6)为零可得到 w 和 b 最优解的闭式(closed-form)解

线性回归 - 最小二乘法

- 得到闭式（closed-form）解

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

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

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

多元线性回归

更一般的，给定数据集D, 样本由d个属性描述

- 给定数据集

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

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

- 多元线性回归目标

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

这称为“多元线性回归”(multivariate linear regression)

多元线性回归

类似的，可以用最小二乘法估计w,b

- 把$\boldsymbol{w}$和b 吸收入向量形式$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$

数据集D表示为m*(d+1)大小的矩阵X

每行为一个示例

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

该行前d个元素对于d个属性值

最后一个元素恒置为1

再把标记也写成向量形式:

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

多元线性回归 - 最小二乘法

类似的，

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

- 最小二乘法（least square method）

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

令 $E_{\hat{w}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$ ，对$\hat{\boldsymbol{w}}$求导得到

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

令上式为零可得$\hat{\boldsymbol{w}}$最优解的闭式解

下面做简单的讨论

多元线性回归 - 最小二乘法

- 最小二乘法（least square method）

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

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

矩阵求导推导过程：

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

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

对 $\hat{\boldsymbol{w}}$ 求导可得

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

由矩阵微分公式 $\dfrac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \dfrac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}, \dfrac{\partial \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\boldsymbol{x}$ 可得

$$\frac{\partial E_{\hat{w}}}{\partial \hat{\boldsymbol{w}}} = 0 - \mathbf{X}^{\mathrm{T}}\boldsymbol{y} - \mathbf{X}^{\mathrm{T}}\boldsymbol{y} + (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{w}}$$

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

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，令 $\dfrac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}\right) = 0$

可逆

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， 令 $\hat{\boldsymbol{x}}_i = (\boldsymbol{x}_i, 1)$

学得线性回归模型为 $f\left(\hat{\boldsymbol{x}}_i\right) = \hat{\boldsymbol{x}}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$

多元线性回归 - 满秩讨论

然而,现实任务中 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 往往不是满秩矩阵. 例如:

在许多任务中我们会遇到大量的变量，其数目甚至超过样例数，

导致X的列数多于行数， $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 显然不满秩.

此时: 可解出多个w，它们都能使均方误差最小化.

选择哪一个解作为输出，将由学习算法的归纳偏好决定，

常见的做法是: **引入正则化(regularization)项.**

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 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（参加6.4节，11.4节）

对数线性回归

线性模型虽简单，却有丰富的变化。例如对于样例（$\boldsymbol{x}$，y），$y \in \mathbb{R}$,

希望线性模型的预测值逼近真实标记y时候，就得到了线性回归模型；

为了便于观察，我们把线性回归模型简写为：

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

可否逼近**y的衍生物？**

例如：　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

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

这就是“对数线性回归”(log-linear regression)，

它实际上是在试图让 $e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}$ 逼近y

在形式上仍然是线性回归，但实质上是在求取输入空间到输出空间的非线性函数映射

对数线性回归

如图3.1所示，这里的**对数函数**起到了**将线性模型的预测值与真实标记联系起来的作用**

- **输出标记的对数**为线性模型逼近的**目标**

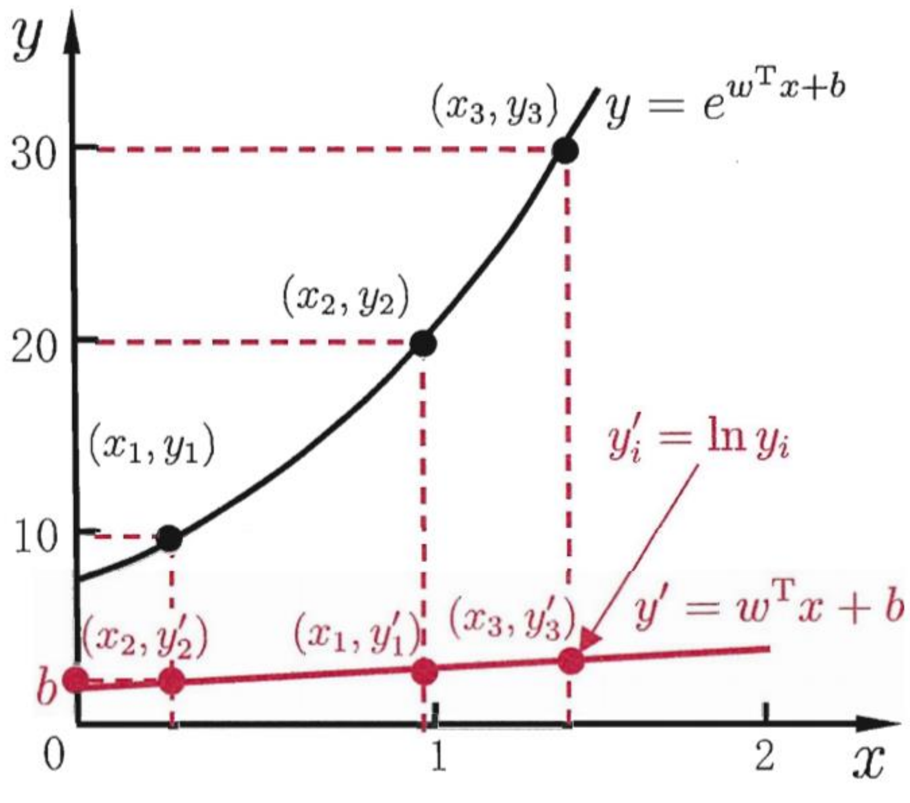


图 3.1　对数线性回归示意图

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

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

线性回归 - 广义线性模型

更一般的，考虑单调可微函数g(.):

- 一般形式，令 $y = g^{-1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right)$

这样得到的模型称为“广义线性模型”（Generalized linear model)

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

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

对数几率回归

讨论了如何使用线性模型进行回归学习，但若要做的是分类任务，

广义线性模型， $$y = g^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)$$

只需要找一个单调可微函数，

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

二分类任务

考虑二分类任务,其输出标记$y \in \{0,1\}$,

而线性回归模型产生的预测值

$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \text{ 是实值}$$

于是，我们需将实值z转换为0/1值

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

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

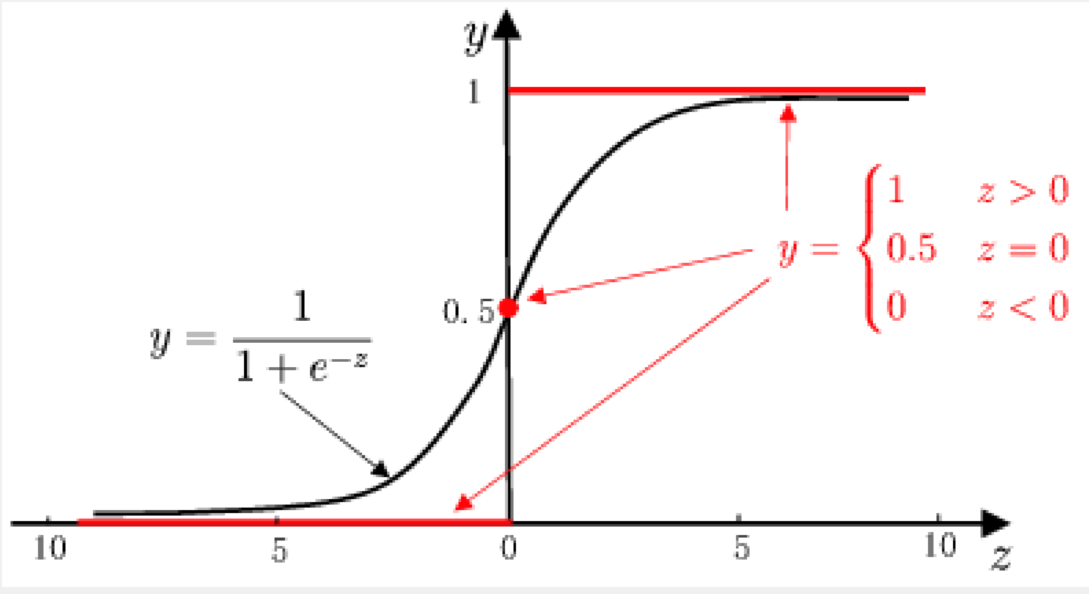
预测值大于零就判为正例，
小于零就判为反例，
预测值为临界值零则可任意判别

二分类任务

单位阶跃函数与对数几率函数的比较

- 如图

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



- 单位阶跃函数缺点
  - 不连续

但从图 3.2可看出：单位阶跃函数不连续,因此不能直接用作式(3.15)中的 $g^{-}(\cdot)$

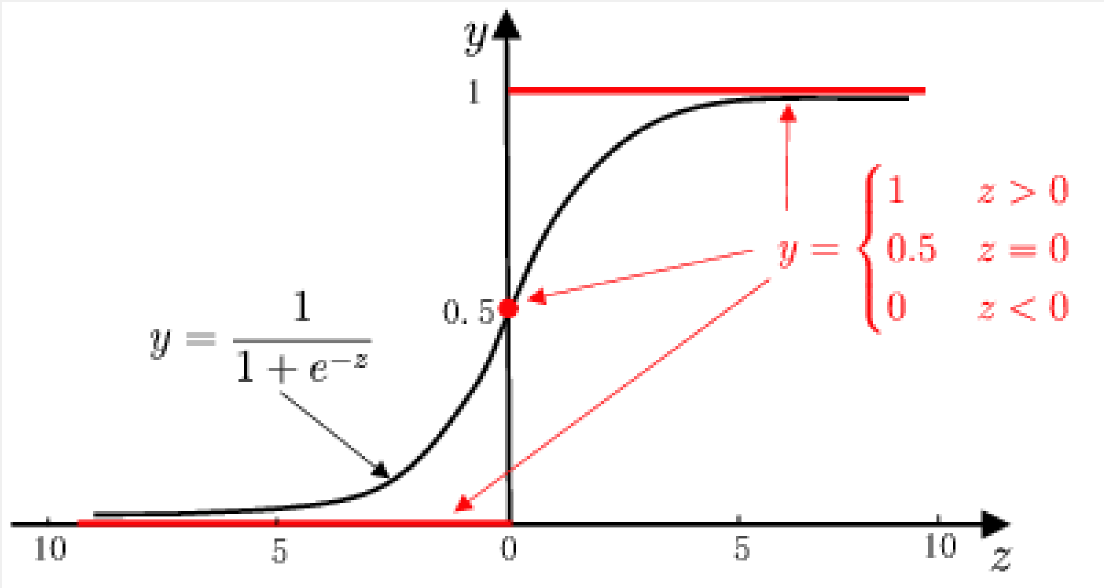
于是我们希望: 找到能在一定程度上**近似单位阶跃函数的“替代函数”**(surrogate function)，**并希望它单调可微**.

对数几率函数(logistic function)正是这样一个常用的替代函数:

二分类任务

单位阶跃函数与对数几率函数的比较

- 如图



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

- 单位阶跃函数缺点
  - 不连续
- **替代函数——对数几率函数（logistic function）**
  - **单调可微、任意阶可导**
  - **它将z值转化为一个接近0或1的y值，并且输出在z=0附近变化很陡**

对数几率回归

广义线性模型

- 运用对数几率函数代入 → $y = g^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)$

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

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

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

对数几率回归

虽然名字是回归，
但是是一种**分类算法**

■ 对数几率（log odds） $$\ln \frac{y}{1-y} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

若将y视为样本x作为正例的可能性,则1-y是其反例可能性，

两者的**比值** $$\frac{y}{1-y}$$

称为“几率”(odds)，反映了x作为正例的相对可能性.

取对数，则得到**对数几率:** $$\ln \frac{y}{1-y}$$

■ 对数几率回归优点

- 无需事先假设数据分布，直接对分类可行性建模，避免假设分布不准确带来的问题
- 不仅预测类别，可得到“类别”的近似概率预测，对利用概率辅助决策的任务有效
- **对率函数**是**任意阶可导的凸函数**，有很好的的数学性质，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

$$y=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}\ .$$ 如何确定式(3.18)中的 $\boldsymbol{w}$ 和 b.

将y视为后验概率$p(y=1\,|\,x)$则，
$$\ln\frac{y}{1-y}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$$

- 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

可通过“极大似然法”(maximum likelihood method)来估计 W,b

使联合概率最大化

- **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

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

即令每个样本属于其真实标记的概率越大越好

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

$$p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=p(y=1\mid\hat{\boldsymbol{x}};\boldsymbol{\beta})$$

$$p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=p(y=0\mid\hat{\boldsymbol{x}};\boldsymbol{\beta})=1-p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})$$

$$\ell(\boldsymbol{w},b)=\sum_{i=1}^{m}\ln p(y_i\mid\boldsymbol{x}_i;\boldsymbol{w},b)\ , \tag{3.25}$$

则3.25似然项可重写为

$$p(y_i\mid\boldsymbol{x}_i;\boldsymbol{w}_i,b)=y_ip_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})+(1-y_i)p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})$$

  - 上式代入3.25, 故等价形式为要最小化

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

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

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) \ , \tag{3.25}$$

$$p\left(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}_i, b\right) = y_i p_1\left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}\right) + \left(1 - y_i\right) p_0\left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}\right)$$

- 上式代入3.25, 可得

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \ln\left(y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})\right)$$

其中 $p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) = \dfrac{e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}}{1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}}, p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) = \dfrac{1}{1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}},$ 代入上式可得

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

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

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

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

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) \ , \qquad (3.25)$$

由于 y_i=0 或 1，则

$$\ell(\boldsymbol{\beta}) = \begin{cases} \sum_{i=1}^{m}(-\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})), & y_i = 0 \\ \sum_{i=1}^{m}(\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i - \ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})), & y_i = 1 \end{cases}$$

两式综合可得

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

此式为极大似然估计的似然函数，所以最大化似然函数等价于最小化似然函数的相反数，

故等价形式为要最小化

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

对数几率回归

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

式(3.27)是关于 $\boldsymbol{\beta}$ 的高阶可导连续凸函数,

经典数值优化算法梯度下降法，牛顿法均可以求解

- 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

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

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LDA**（Linear Discriminant Analysis）[Fisher, 1936]

  LDA也可被视为一种监督降维技术

LDA思想非常朴素：

给定训练样本集，设法将样本投影到一条直线，**使得：**

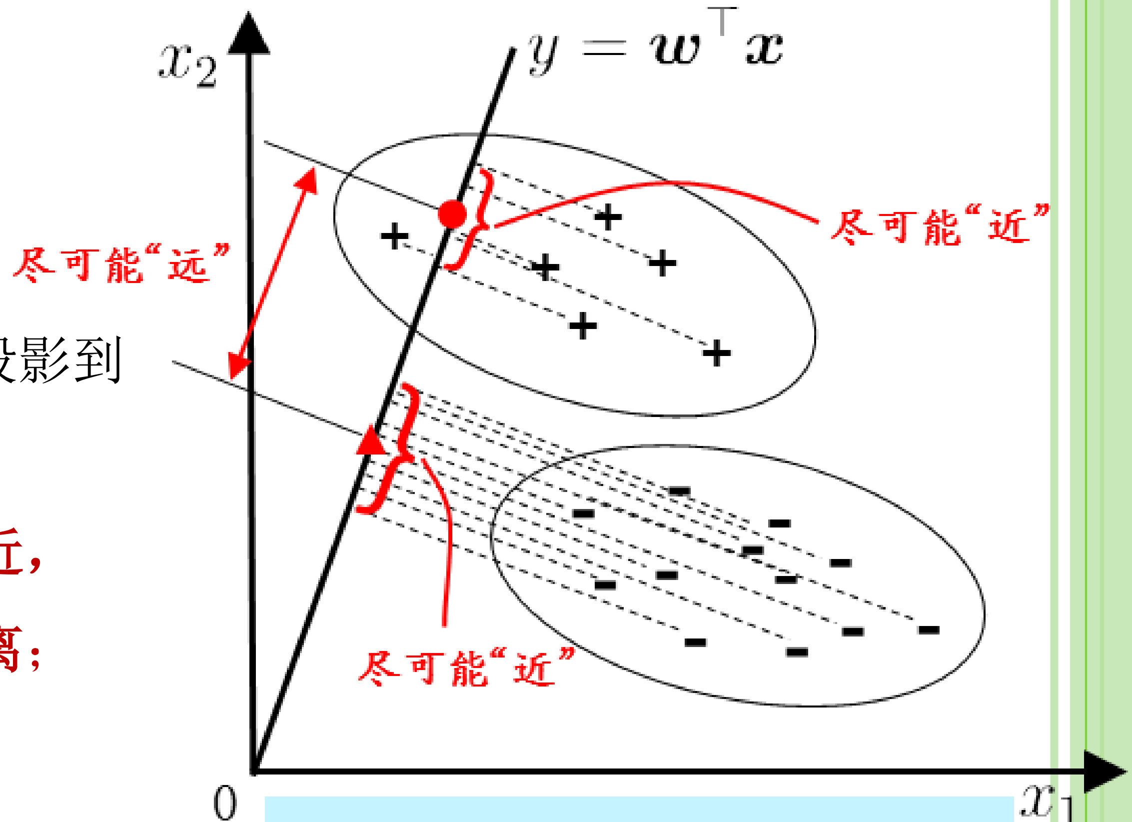
同类样例的投影点尽可能的接近，

异类样例的投影点尽可能的远离；

在对新样本进行分类时，

将其投影到同样的这条直线上，

再根据投影点的位置来确定新样本的类别，如图



"+"、"_"分别代表正例和反例，椭圆表示数据簇外轮廓,虚线表示投影，红色实心圆和实心三角形分别表示：两类样本投影后的中心点.

二分类任务－ 线性判别分析

- **线性判别分析LDA**（Linear Discriminant Analysis） [Fisher, 1936]
- **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 $\boldsymbol{\Sigma}_i$
  - 两类样本的中心在直线上的投影： $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$
  - 若将所有样本都投影到直线上，两类样本的协方差：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} \qquad \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$$

因为直线为一维空间，
所有均为实数

二分类任务－ 线性判别分析

要让同类样本投影点尽可能接近，可以让同类样本投影点协方差尽可能小

即 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小

要让异类样本投影点尽可能远，可以类中心距离尽可能的大

即 $\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2$ 尽可能大

同时考虑，则得到

- 最大化目标

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

二分类任务－ 线性判别分析

最大化目标

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

要让同类样本投影点尽可能接近，可以让同类样本投影点协方差尽可能小

即 $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Sigma}_0 \boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Sigma}_1 \boldsymbol{w}$ 尽可能小

要让异类样本投影点尽可能远，可以类中心距离尽可能的大

即 $\| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_1 \|_2^2$ 尽可能大

推导：

$$\begin{aligned}
J &= \frac{\| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_1 \|_2^2}{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \boldsymbol{w}} \\
&= \frac{\| (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_1)^{\mathrm{T}} \|_2^2}{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \boldsymbol{w}} \\
&= \frac{\| (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{w} \|_2^2}{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \boldsymbol{w}} \\
&= \frac{\left[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{w} \right]^{\mathrm{T}} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \boldsymbol{w}} \\
&= \frac{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \boldsymbol{w}}
\end{aligned}$$

二分类任务－ 线性判别分析

定义

- 类内散度矩阵 (within-class scatter matrix)

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

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

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

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

可重写为

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

这就是 LDA 欲最大化的目标，即 $\mathbf{S}_b$ 与 $\mathbf{S}_w$ 的“广义瑞利商”

二分类任务－ 线性判别分析

- 广义瑞利商（generalized Rayleigh quotient）

$$J = \frac{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}} \tag{3.35}$$

如何确定w呢？注意到式(3.35)的分子和分母都是关于w的二次项，式(3.35)的解与w的长度无关,只与其方向有关

不失一般性

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

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

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

二分类任务－ 线性判别分析

- 运用拉格朗日乘子法

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

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

等价于

推导

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w} \quad (3.37)$$

拉格朗日函数为

$$L(\boldsymbol{w},\lambda) = -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w} + \lambda(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} - 1)$$

对 $\boldsymbol{w}$ 求偏导可得

$$\begin{aligned}\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} &= -\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w})}{\partial \boldsymbol{w}} + \lambda\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} - 1)}{\partial \boldsymbol{w}} \\ &= -(\mathbf{S}_b + \mathbf{S}_b^{\mathrm{T}})\boldsymbol{w} + \lambda(\mathbf{S}_w + \mathbf{S}_w^{\mathrm{T}})\boldsymbol{w}\end{aligned}$$

由于 $\mathbf{S}_b = \mathbf{S}_b^{\mathrm{T}}, \mathbf{S}_w = \mathbf{S}_w^{\mathrm{T}}$，所以

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} = -2\mathbf{S}_b\boldsymbol{w} + 2\lambda\mathbf{S}_w\boldsymbol{w}$$

二分类任务－ 线性判别分析

- 运用拉格朗日乘子法

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

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

等价于

推导

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w} \quad (3.37)$$

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} = -2\mathbf{S}_b\boldsymbol{w} + 2\lambda\mathbf{S}_w\boldsymbol{w}$$

令导数等于 0 即可得：

$$-2\mathbf{S}_b\boldsymbol{w} + 2\lambda\mathbf{S}_w\boldsymbol{w} = 0$$

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

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

若令 $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w} = \gamma$，则

$$\gamma(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) = \lambda\mathbf{S}_w\boldsymbol{w}$$

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

由于最终要求解的 $\boldsymbol{w}$ 不关心其大小，只关心其方向，所以 γ/λ 这个常数项可以任意取值

“不妨令 $\mathbf{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$” 就等价于令 $\frac{\gamma}{\lambda} = 1$

二分类任务 - 线性判别分析

- 运用拉格朗日乘子法

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

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

等价于

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w} \quad (3.37)$$

λ 是拉格朗日乘子. 注意到 $\mathbf{S}_b\boldsymbol{w}$ 的方向恒为 $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, 不妨令

$$\mathbf{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \ ,$$

代入式(3.37)即得

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

二分类任务－ 线性判别分析

- 同向向量

$$\mathbf{S}_b \boldsymbol{w} = \lambda (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

- 结果

同向向量

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

- 求解
  - 考虑数值解的稳定性，通常进行 奇异值分解 $\mathbf{S}_w = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$

$\boldsymbol{\Sigma}$ 是一个实对角矩阵，其对角线上的元素是 $\mathbf{S}_w$ 的奇异值

再由 $\mathbf{S}_w^{-1} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$ 得到 $\mathbf{S}_w^{-1}$.

- LDA可从贝叶斯决策论解释
  - **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－ 多分类任务

可以将 LDA 推广到多分类任务中．假定存在 N 个类，

且第 i 类示例数为 m_i.

- 定义 **全局散度矩阵**

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

$\boldsymbol{\mu}$ 是所有示例的均值向量．将类内散度矩阵 $\mathbf{S}_w$ 重定义为每个类别的散度矩阵之和，即

- 类内散度矩阵

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

其中

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

LDA推广 - 多分类任务

- 定义 **全局散度矩阵**

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

- 类内散度矩阵

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

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

- 求解得

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

显然, 多分类 LDA 可以有多种实现方法: 使用 $\mathbf{S}_b$, $\mathbf{S}_w$, $\mathbf{S}_t$ 三者中的任何两个即可. 常见的一种实现是采用优化目标

LDA推广 – 多分类任务

- 定义 **全局散度矩阵** $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$

$$= \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

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

- 类内散度矩阵

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

求解得 $\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w$

$$= \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

推导：

$$\begin{aligned}
\mathbf{S}_b &= \mathbf{S}_t - \mathbf{S}_w \\
&= \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}} - \sum_{i=1}^{N} \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} - (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \right) \right) \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{T}}) - (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_i^{\mathrm{T}}) \right) \right) \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \right) \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left(-\boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \right)
\end{aligned}$$

LDA推广 - 多分类任务

推导：

$$
\begin{aligned}
\mathbf{S}_b &= \mathbf{S}_t - \mathbf{S}_w \\
&= \sum_{i=1}^{m}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}} - \sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \\
&= \sum_{i=1}^{N}\left(\sum_{\boldsymbol{x}\in X_i}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}-(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\right)\right) \\
&= \sum_{i=1}^{N}\left(\sum_{\boldsymbol{x}\in X_i}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}^{\mathrm{T}})-(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}_i^{\mathrm{T}})\right)\right) \\
&= \sum_{i=1}^{N}\left(\sum_{\boldsymbol{x}\in X_i}\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}}-\boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}-\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}+\boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}}+\boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right)\right) \\
&= \sum_{i=1}^{N}\left(\sum_{\boldsymbol{x}\in X_i}\left(-\boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}}-\boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+\boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}}+\boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{N}\left(-\sum_{\boldsymbol{x}\in X_i}\boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}}-\sum_{\boldsymbol{x}\in X_i}\boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}}+\sum_{\boldsymbol{x}\in X_i}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+\sum_{\boldsymbol{x}\in X_i}\boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}}+\sum_{\boldsymbol{x}\in X_i}\boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}}-\sum_{\boldsymbol{x}\in X_i}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right) \\
&= \sum_{i=1}^{N}\left(-m_i\boldsymbol{\mu}_i\boldsymbol{\mu}^{\mathrm{T}}-m_i\boldsymbol{\mu}\boldsymbol{\mu}_i^{\mathrm{T}}+m_i\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}+m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}-m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right) \\
&= \sum_{i=1}^{N}\left(-m_i\boldsymbol{\mu}_i\boldsymbol{\mu}^{\mathrm{T}}-m_i\boldsymbol{\mu}\boldsymbol{\mu}_i^{\mathrm{T}}+m_i\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right) \\
&= \sum_{i=1}^{N}m_i\left(-\boldsymbol{\mu}_i\boldsymbol{\mu}^{\mathrm{T}}-\boldsymbol{\mu}\boldsymbol{\mu}_i^{\mathrm{T}}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right) \\
&= \sum_{i=1}^{N}m_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^{\mathrm{T}}
\end{aligned}
$$

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 $\mathbf{W} \in \mathbb{R}^{d\times(N-1)}$, $\operatorname{tr}(\cdot)$ 表示矩阵的迹(trace)

上式可通过如下
广义特征值问题求解

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

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

- 若将W视为一个投影矩阵，则多分类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，通过投影减少了维数，并且使用了类别信息，**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
- 利用二分类学习器解决多分类问题（常用）
- **基本思路：拆解法，即将多分类任务拆为若干个二分类任务求解**
  - 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

关键：如何拆分？如何集成？

本节主要介绍拆分策略

多分类学习

- 经典拆分策略
  - 一对一（One vs. One, OvO）
  - 一对其余（One vs. Rest, OvR）
  - 多对多（Many vs. Many, MvM）

不失一般性，考虑 N 个类别 $C_1, C_2, \ldots, C_N$

给定数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, $y_i \in \{C_1, C_2, \ldots, C_N\}$

OvO 将这N个类别两两配对，从而产生N(N-1)/2个二分类任务:

例: 为区分类别 Ci和Cj 训练一个分类器,

该分类器把D中的Ci类样例作为正例，Cj类样例作为反例;

在测试阶段，新样本将同时提交给所有分类器将得到N(N-1)/2个分类结果,

最终结果可通过投票产生:

即把被预测得最多的类别作为最终分类结果

多分类学习 - 一对一

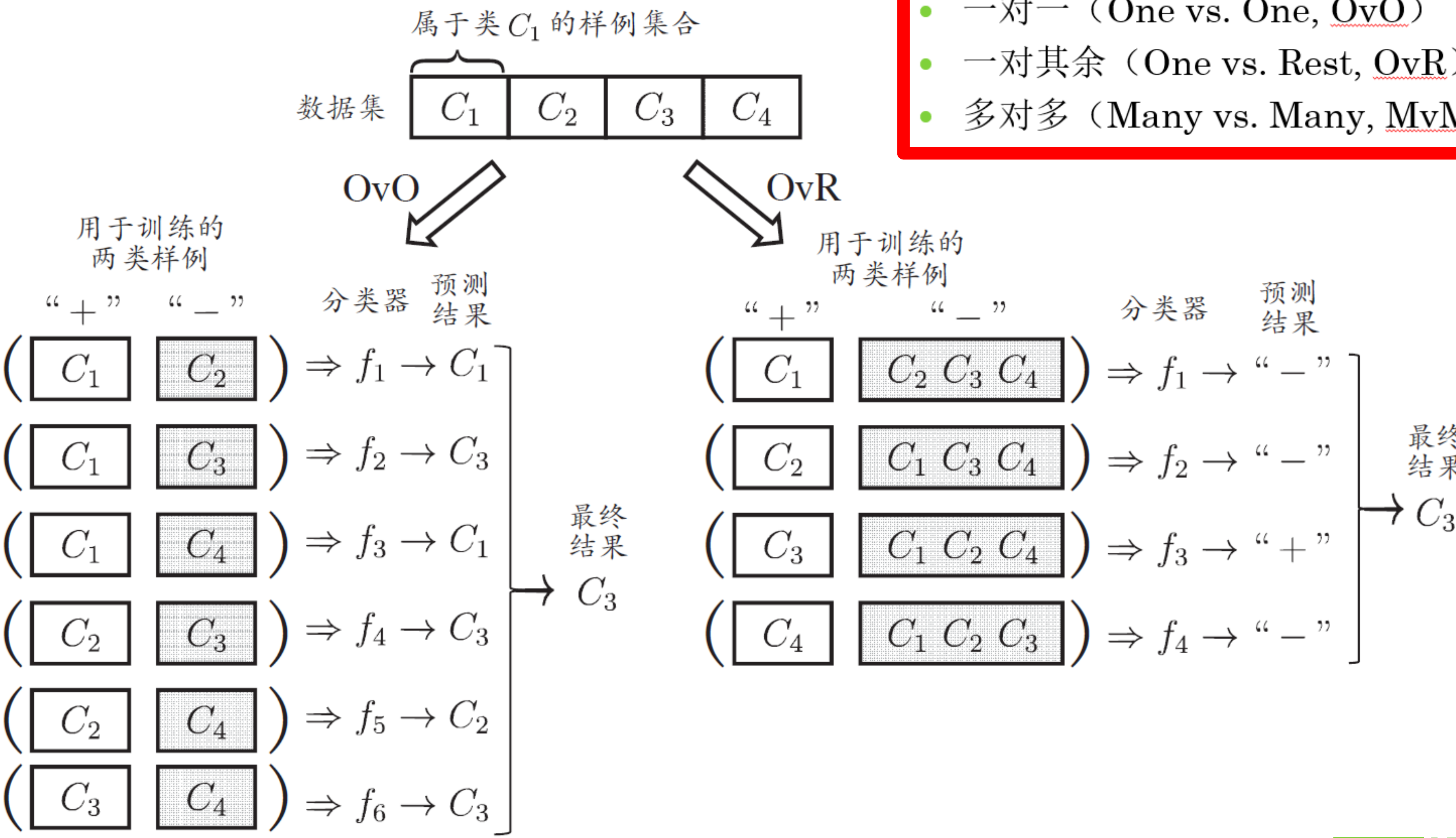
- **拆分阶段**
  - N个类别两两配对
    - N(N-1)/2 个二类任务
  - 各个二类任务学习分类器
    - N(N-1)/2 个二类分类器
- **测试阶段**
  - 新样本提交给所有分类器预测
    - N(N-1)/2 个分类结果
  - 投票产生最终分类结果
    - 被预测最多的类别为最终类别

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

- 一对一（One vs. One, OvO）
- 一对其余（One vs. Rest, OvR）
- 多对多（Many vs. Many, MvM）



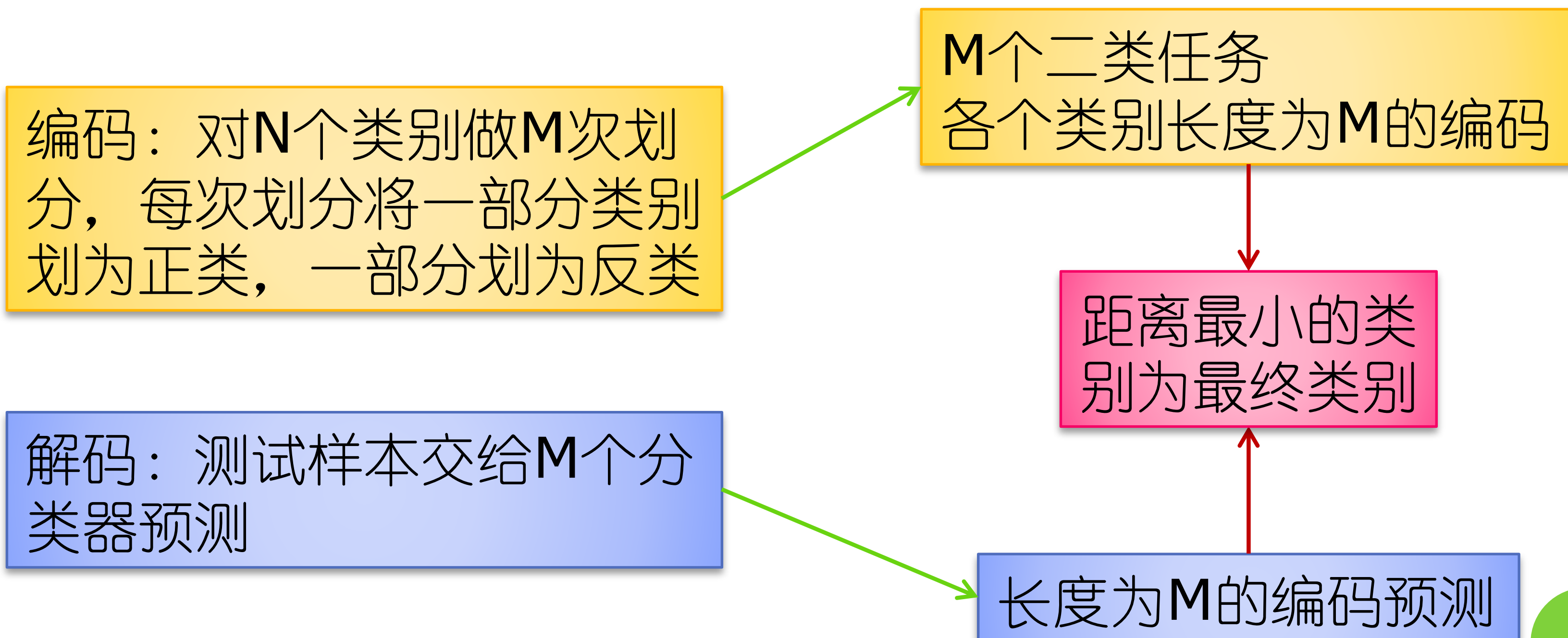
多分类学习 - 两种策略比较

一对一	一对其余
训练N(N-1)/2个分类器，存储开销和测试时间大	训练N个分类器，存储开销和测试时间小
训练只用两个类的样例，训练时间短	训练用到全部训练样例，训练时间长

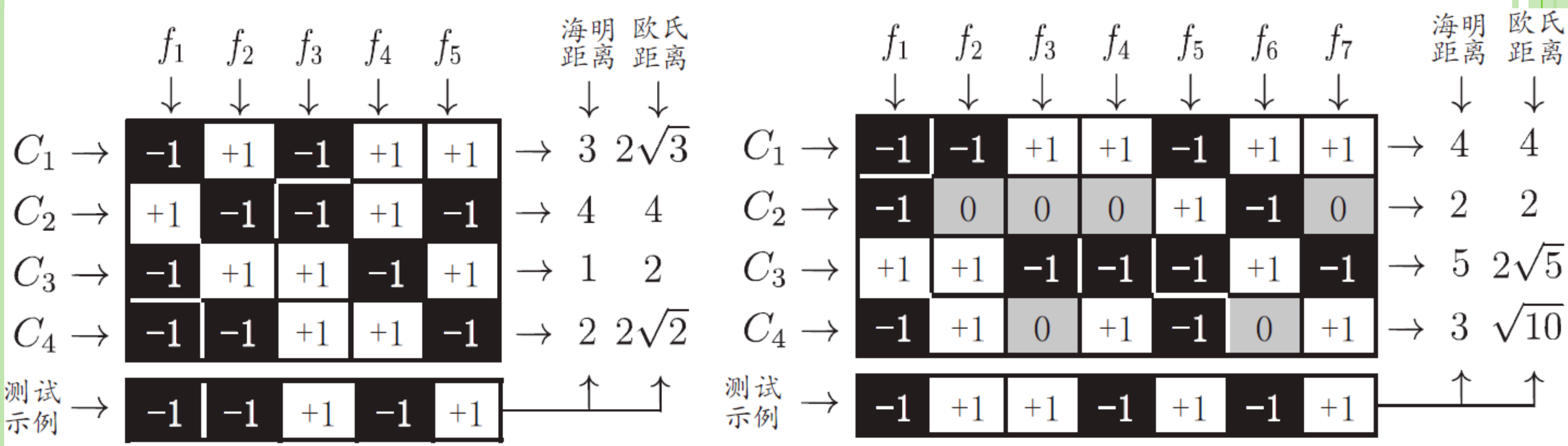
预测性能取决于具体数据分布，多数情况下两者差不多

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



(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）

类别平衡正例预测 $\frac{y}{1-y} > 1$ $\Rightarrow$ $\frac{y}{1-y} > \frac{m^+}{m^-}$ 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