三、线性模型（续）

主讲教师：赵鹏

线性模型的变化

对于样例 $(\boldsymbol{x}, y)$, $y \in \mathbb{R}$, 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 $y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$

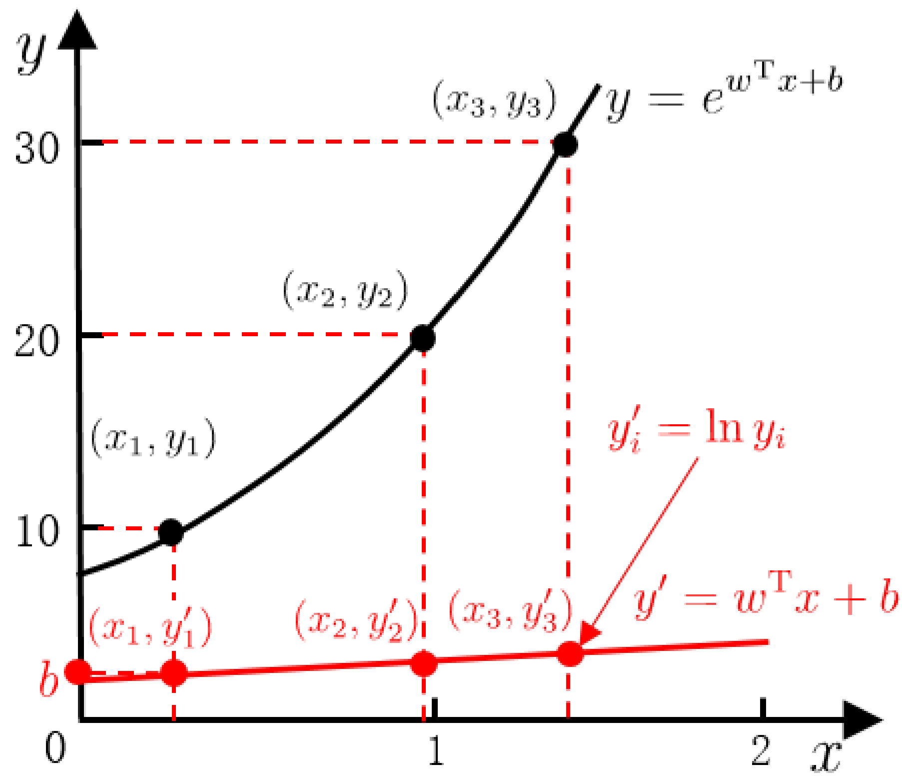
令预测值逼近 y 的衍生物？

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

则得到对数线性回归

(log-linear regression)

实际是在用 $e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}$ 逼近 y



广义(generalized)线性模型

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

单调可微的 联系函数 (link function)

令 $g(\cdot) = \ln(\cdot)$ 则得到 对数线性回归

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

… …

二分类任务

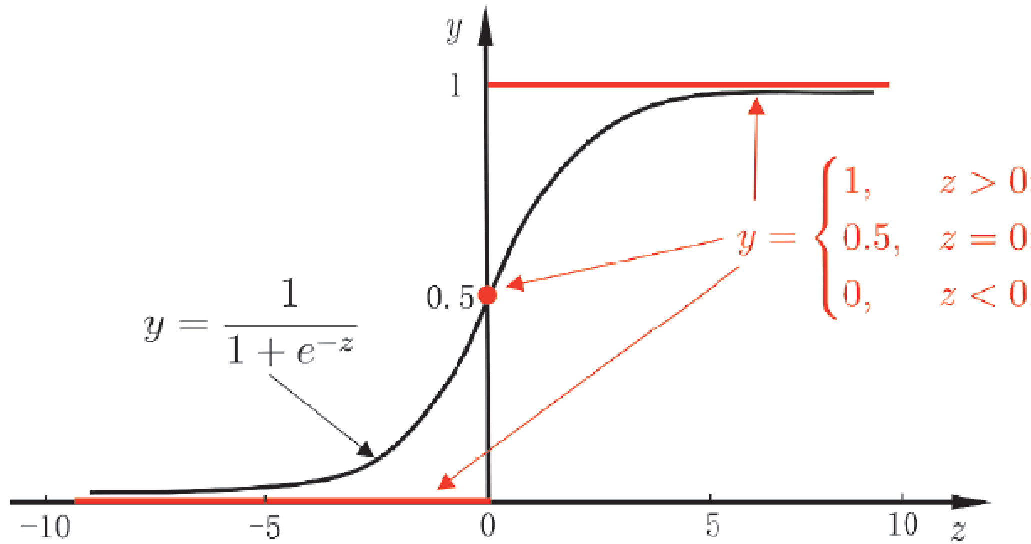
线性回归模型产生的实值输出 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$

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)

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



性质不好，
需找 “替代函数”
(surrogate function)

常用
单调可微、任意阶可导

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

对数几率函数
(logistic function)
简称 “对率函数”

注意：Logistic与 “逻辑” 没有半毛钱关系！
1. Logistic 源自 Logit，不是Logic； 2. 实数值，并非 “非0即1” 的逻辑值

对率回归

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:

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

几率(odds), 反映了 $\boldsymbol{x}$ 作为正例的相对可能性

“对数几率”
(log odds, 亦称 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

注意：它是
分类学习算法!

求解思路

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

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

于是，可使用“极大似然法” ⟶ 第7章
(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

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

求解思路

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

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

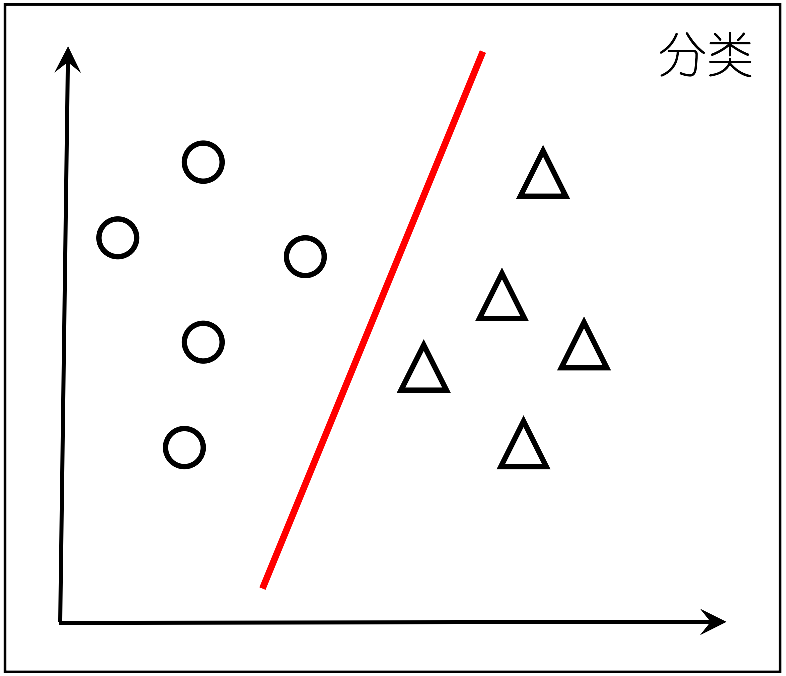
则似然项可重写为 $p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}_i, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})$

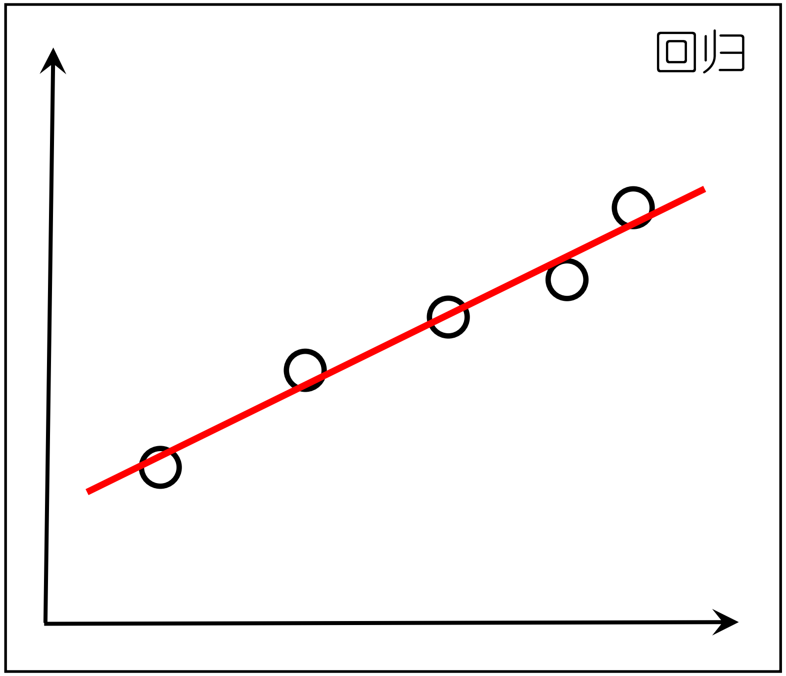
于是，最大化似然函数 $\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b)$

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

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]

线性模型做“分类”



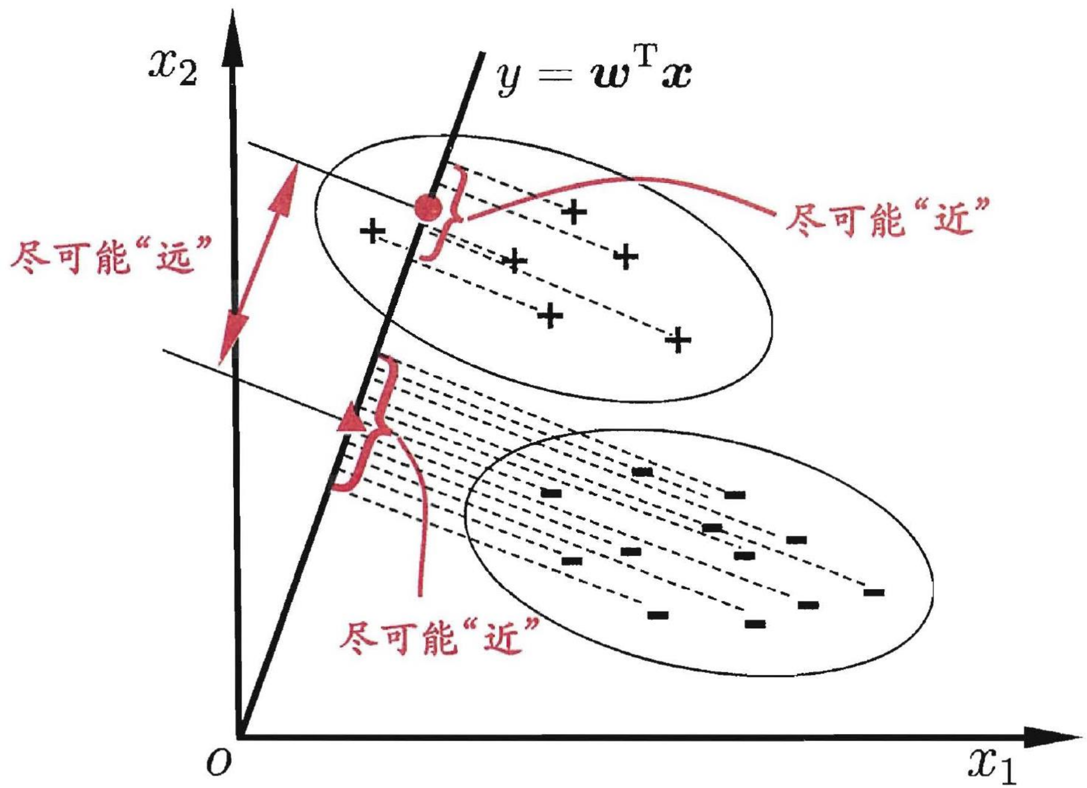


如何“直接”做分类？

广义线性模型；
通过“联系函数”

例如，对率回归

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 ⟶ 第10章

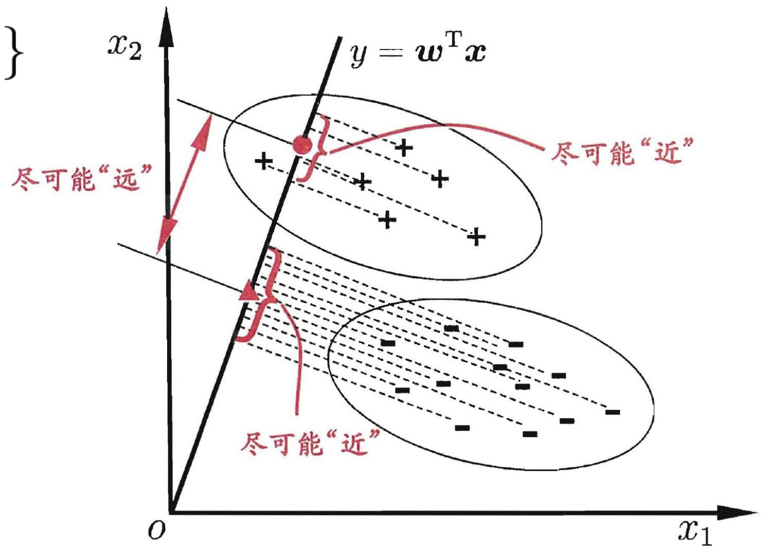
LDA的目标

给定数据集 $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{m}$，$y_k \in \{0,1\}$

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

LDA的目标

基本想法：

- 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

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

⇨ 综合两方面考虑，最大化

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

LDA的目标

类内散度矩阵 (within-class scatter matrix)

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

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)

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

$\boldsymbol{w}$ 成倍缩放不影响 J 值
仅需考虑方向

LDA求解

令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1$，最大化广义瑞利商等价求解如下优化表达式

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

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

使用拉格朗日乘子法求解：

定义拉格朗日函数为 $L(\boldsymbol{w},\nu)=-\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}+\nu\left(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}-1\right)$

对 $\boldsymbol{w}$求导并使之为0可得，$\dfrac{\partial L(\boldsymbol{w},\nu)}{\partial \boldsymbol{w}}=-2\mathbf{S}_b\boldsymbol{w}+2\nu\mathbf{S}_w\boldsymbol{w}=\mathbf{0}$

$$\Longrightarrow \mathbf{S}_b\boldsymbol{w}=\nu\mathbf{S}_w\boldsymbol{w} \quad \Longrightarrow (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\underline{(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}=\nu\mathbf{S}_w\boldsymbol{w}$$

此项为标量，记为 γ

LDA求解

令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1$，最大化广义瑞利商等价求解如下优化表达式

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

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

使用拉格朗日乘子法求解：

$$\Longrightarrow \quad \mathbf{S}_b\boldsymbol{w}=\nu\mathbf{S}_w\boldsymbol{w} \quad \Longrightarrow \quad \gamma(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)=\nu\mathbf{S}_w\boldsymbol{w}$$

如果只考虑方向，则只需求解如下的“简化问题”

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

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

考虑数值稳定性，可通过奇异值分解实现求逆，

$\mathbf{S}_w=\mathbf{U\Sigma V}^{\mathrm{T}}$ 然后 $\mathbf{S}_w^{-1}=\mathbf{V\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$

LDA：推广到多类

假定总计m个样本，有N个类，每个类有m_i个样本，$i \in [N]$

- 全局散度矩阵 $$\mathbf{S}_t = \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$$
- 类内散度矩阵 $$\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}, \quad \mathbf{S}_{w_i} = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top$$
- 类间散度矩阵 $$\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top$$

多分类LDA有多种实现方法：采用 $\mathbf{S}_b$, $\mathbf{S}_w$, $\mathbf{S}_t$ 中的任何两个

LDA：推广到多类

假定总计m个样本，有N个类，每个类有m_i个样本，$i \in [N]$

- 全局散度矩阵 $$\mathbf{S}_t = \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$$

- 类内散度矩阵 $$\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}, \quad \mathbf{S}_{w_i} = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top$$

- 类间散度矩阵 $$\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top$$

例如，$$\max_{\mathbf{W}} \frac{\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W})}{\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W})} \quad \Longrightarrow \quad \mathbf{S}_b\mathbf{W} = \lambda \mathbf{S}_w \mathbf{W}$$

$$\mathbf{W} \in \mathbb{R}^{d \times (N-1)}$$

$\mathbf{W}$的闭式解是 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的 d' ($\leq N$-1) 个最大非零广义特征值对应的特征向量组成的矩阵

多分类学习

直接法：一些方法可以直接推广（例如LDA）

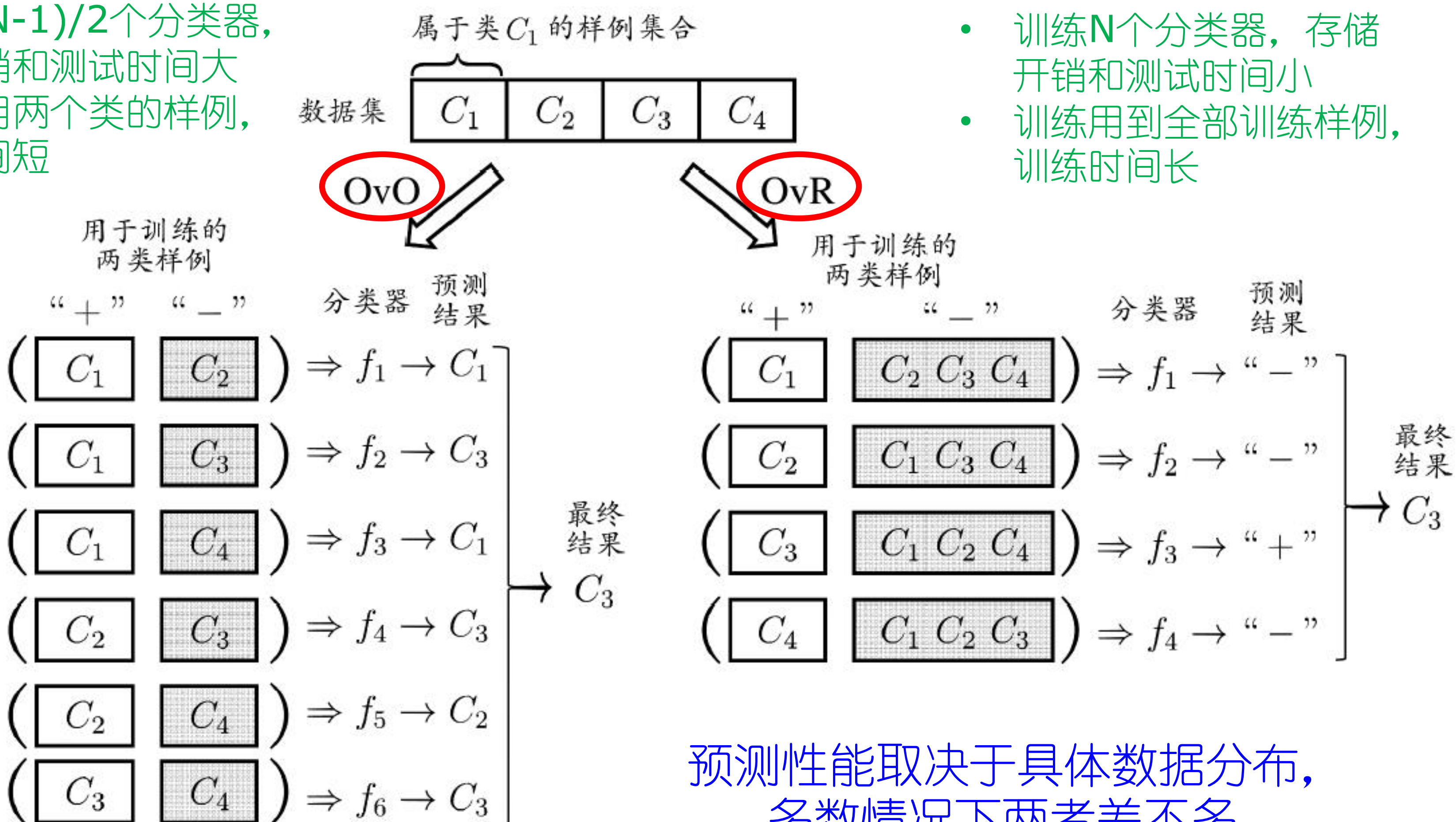
拆解法：将一个多分类任务拆分为若干个二分类任务求解

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

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，多数情况下两者差不多

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， ECOC)**

编码：对 N 个类别做 M 次划分，每次将一部分类别划为正类，一部分划为反类

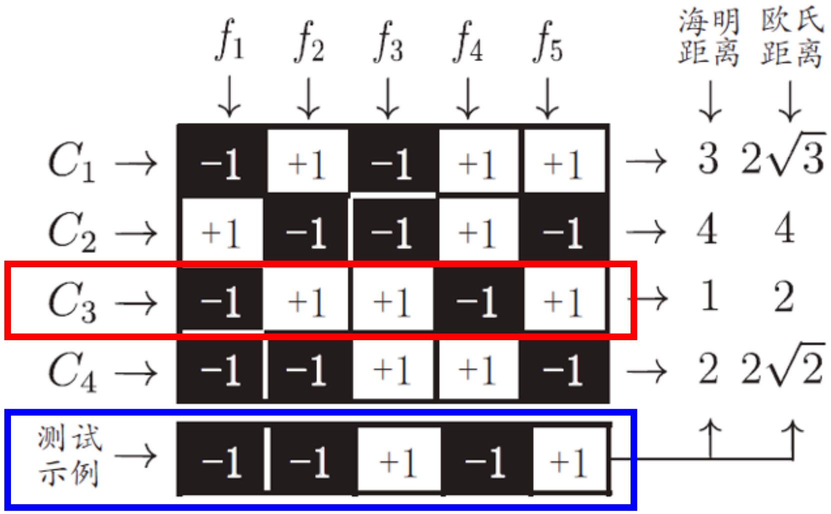
M 个二类任务；
(原)每类对应一个长为 M 的编码

距离最小的类为最终结果

解码：测试样本交给 M 个分类器预测

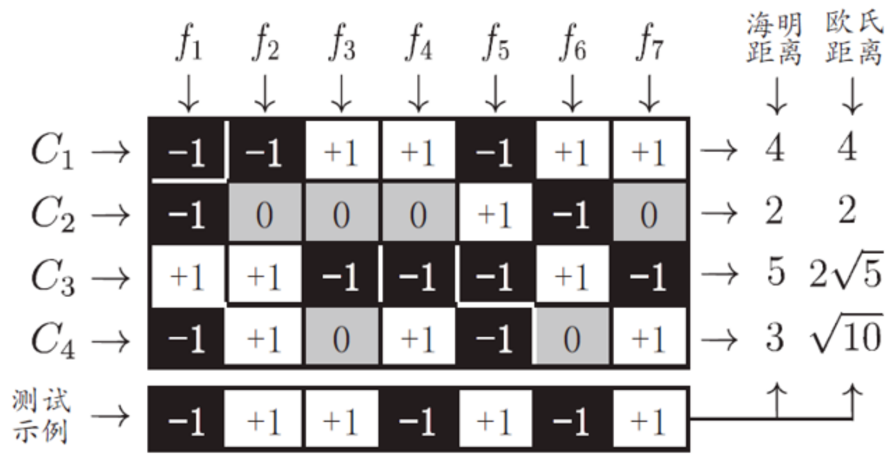
长为 M 的预测结果编码

纠错输出码 (ECOC)



(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 (class-imbalance)

不同类别的样本比例相差很大；“小类”往往更重要

基本思路：

若 $\frac{y}{1-y} > 1$ 则 预测为正例. ⇒ $\boxed{\text{若 } \frac{y}{1-y} > \frac{m^+}{m^-} \text{ 则 预测为正例.}}$

基本策略
—— “再缩放” (rescaling)：

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

然而，精确估计 m^-/m^+ 通常很困难！

常见类别不平衡学习方法：

- 过采样 (oversampling)
  例如：SMOTE
- 欠采样 (undersampling)
  例如：EasyEnsemble
- 阈值移动 (threshold-moving)

前往第四站……