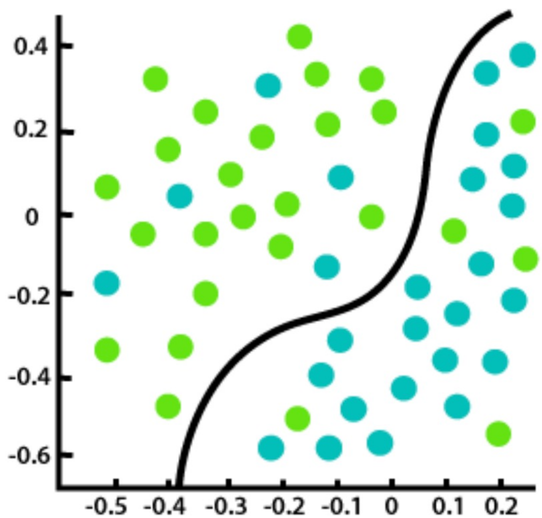
# 机器学习2：线性模型

Tieliang Gong 龚铁梁

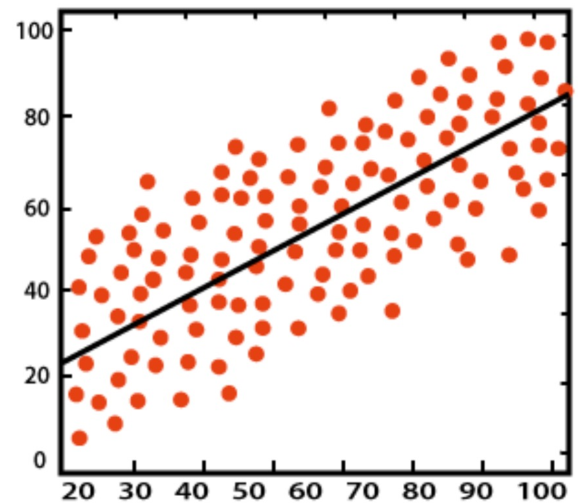
School of Computer Science & Technology,

Xi’an Jiaotong University

# 线性模型




分类




回归

- 线性模型(Linear model)试图学得通过属性的线性组合来进行预测的函数

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

- 向量形式

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

简单、基本、可解释性好

# 线性回归

线性回归试图学得

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

均方最小化：

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

求解$w$和$b$使得$E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$最小化的过程，称为**线性回归**

# 线性回归

分别对 $w$ 和 $b$ 求导可得：

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

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

令导数为0，可得闭式解（closed-form solution）

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

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

# 多元线性回归

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

把$\boldsymbol{w}$和$\boldsymbol{b}$表示成向量形式$\widehat{\boldsymbol{w}} = (\boldsymbol{w}; \boldsymbol{b})$,数据集表示为

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

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

# 多元线性回归

采用最小二乘法求解，有

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

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

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

令其为0，可得最优闭式解，但涉及矩阵求逆

- 若$X^TX$正定，则 $\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$ Gaussian-Markov定理
- 若$X^TX$非正定，则存在多个解，此时，需借助归纳偏好，或引入正则化

# 线性回归估计的统计性质

**Gauss Markov定理：**在线性回归模型中，如果误差满足零均值，同质方差且互不相关，则回归系数的最优线性无偏估计（BLUE, Best Linear Unbiased Estimator）是普通最小二乘估计。

**1.零均值** $E(e_i) = 0,$

**2.同质方差** $Var(e_i) = \sigma^2 < +\infty$

**3.互不相关** $Cov(e_i, e_j) = 0, \forall i \neq j$

$\Longrightarrow$ $E\hat{\boldsymbol{w}}^* = \boldsymbol{w}$

**证明：反证法**

**令** $\bar{\boldsymbol{w}} = \boldsymbol{M}\boldsymbol{y}$ $\Longrightarrow$ **考察** $Var(\bar{\boldsymbol{w}})$ $\Longrightarrow$ **比较** $Var(\bar{\boldsymbol{w}})$

$Var(\hat{\boldsymbol{w}}^*)$

# 线性模型的变化

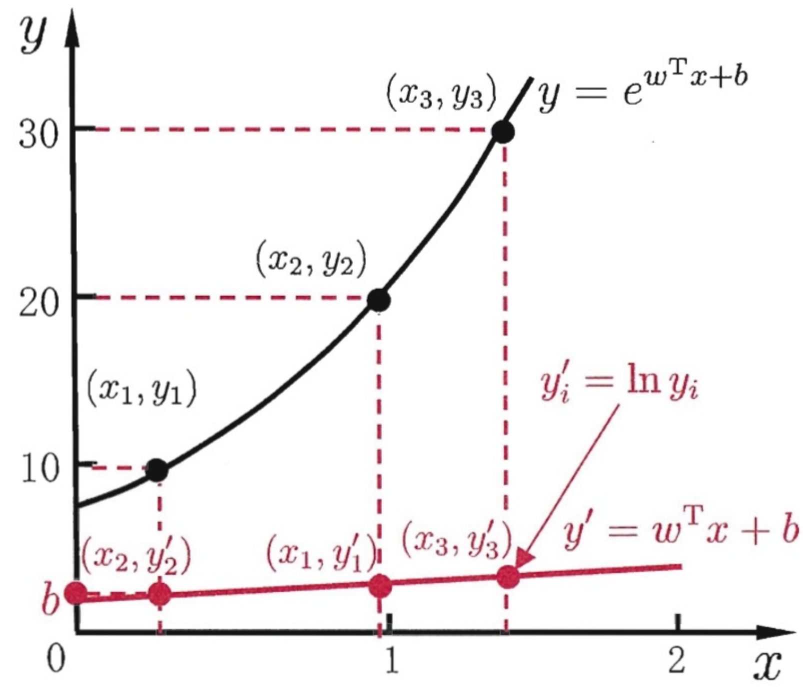
对于样例(x,y), 若希望线性模型的预测值逼近真实标记，则得到线性回归模型 $\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}$

令预测值逼近y的衍生物？

若令$\ln \boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}$, 则得到

对数线性回归(Log-linear regression)

实际上用$\boldsymbol{e}^{\boldsymbol{w}^T\boldsymbol{x}+\boldsymbol{b}}$逼近$\boldsymbol{y}$



# 广义线性模型

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

单调可微的联系函数(linking function)

令 $g(\cdot) = \ln(\cdot)$ ，则得到对数线性回归：

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

# 二分类任务

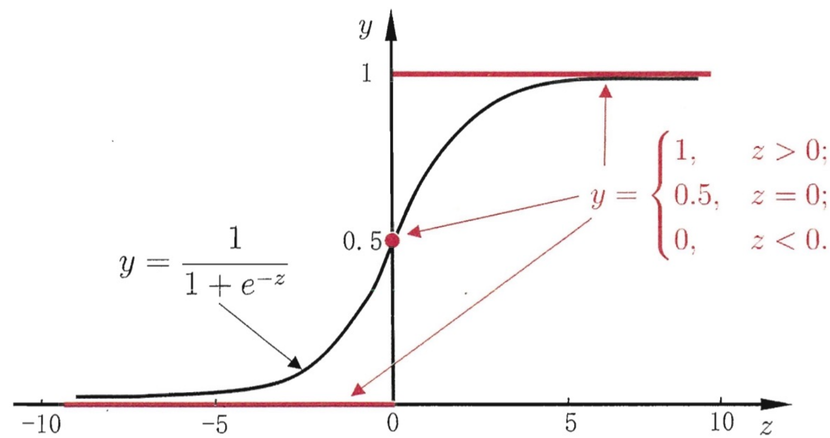
线性回归模型产生的实值输出：$z = \boldsymbol{w}^T \boldsymbol{x} + b$

期望输出：$y \in \{0, 1\}$

找出二者的 **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)**

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

常用，单调可微、任意阶可导

对数几率函数，简称**对率函数**

# 对率回归

以对率函数为联系函数：

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

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

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

几率(odds),反映了x作为正例的相对可能性

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
- 可直接应用于数值优化算法求最优解

注意：它是
分类学习算法

# 求解思路

若将y看作后验概率估计$P(y=1|x)$，则

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

可用“极大似然法”(maximum likelihood method)

对于给定数据集$\{(x_i, y_i)\}_{i=1}^{m}$，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 $p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) = p(y = 1 \mid \hat{\boldsymbol{x}}; \boldsymbol{\beta})$

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

似然项可重写为

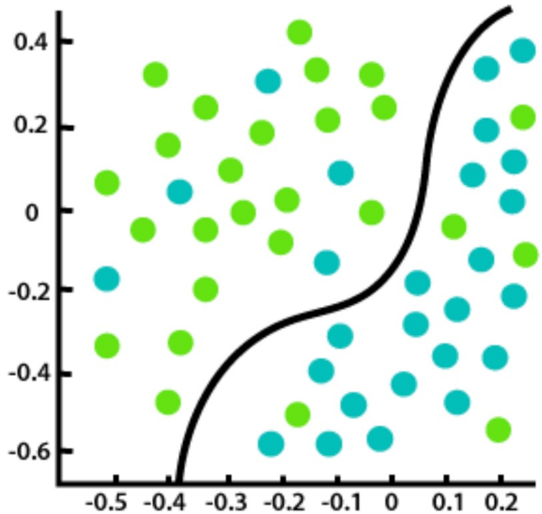
$$\boxed{p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})}$$

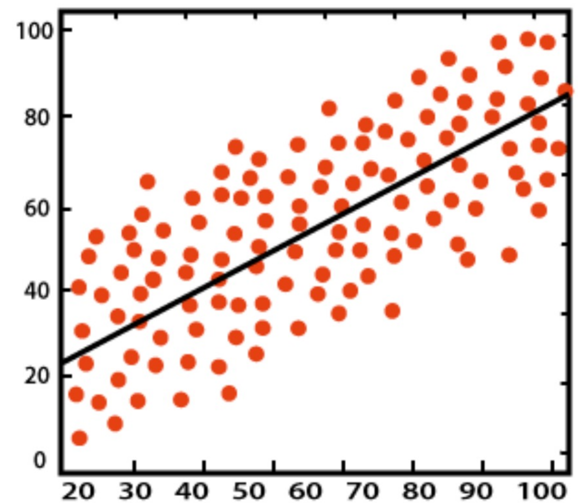
于是，最大似然函数等价于最小化下述目标

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) \Longleftrightarrow \ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i + \ln\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right)$$

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

# 线性模型做“分类”



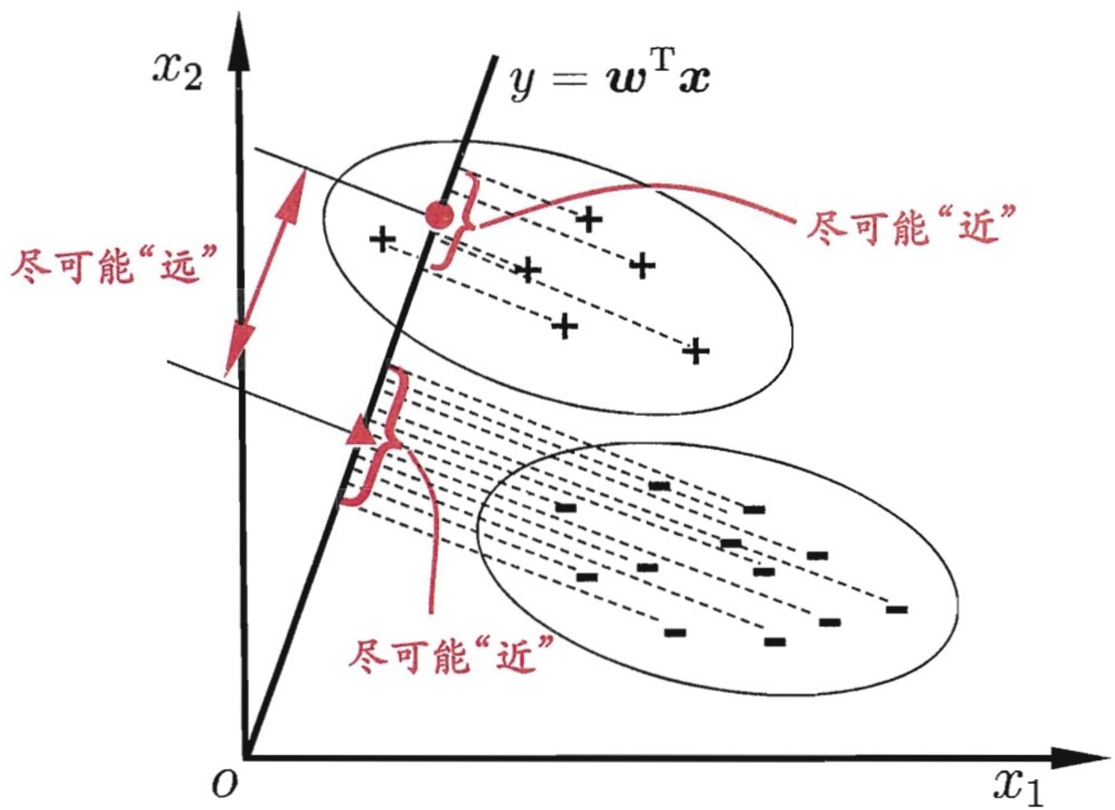




广义线性模型，通过linking function

如何“直接”做分类

对率回归

# 线性判别分析(Linear discriminant Analysis)




LDA的二维示意图. "+", "-"分别代表正例和反例. 椭圆表示数据簇的外轮廓, 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。

# 线性判别分析的目标

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

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

# 线性判别分析的目标

类内离散度矩阵(within-class scatter matrix)

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

类间离散度矩阵(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Rayleigh商**

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

***w*** 的缩放不影响*J*的取值，仅考虑方向

# 求解思路

令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1$ 最大化广义Rayleigh熵的等价形式为

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

由拉格朗日乘子法 $\mathbf{S}_b\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$

注意到 $\mathbf{S}_b\boldsymbol{w}$ 的方向恒为 $\boldsymbol{\mu}_0-\boldsymbol{\mu}_1$, 不妨令 $\mathbf{S}_b\boldsymbol{w}=\lambda(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$

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

实践中通常进行奇异值分解(SVD): $S_w=U\Sigma V^T$

然后: $S_w^{-1}=V\Sigma^{-1}U^T$

LDA可从Bayesian决策理论角度解释：当两类数据同先验、满足高斯分布且协方差相等时，LDA可达最优分类

# 推广到多类

假定有$N$个类，且第$i$类的示例数为$m_i$，定义全局离散度矩阵

$$S_t = \sum_{i=1}^{N} \sum_{x \in X_i} (x - \mu)(x - \mu)' = \sum_{i=1}^{m} \left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)'$$

- 类内离散度矩阵

$$S_w = \sum_{i=1}^{N} \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)'$$

- 类间离散度矩阵

$$S_b = \sum_{i=1}^{N} m_i (\mu_i - \mu)(\mu_i - \mu)'$$

相互联系 $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$

# 推广到多类

假定有 $N$ 个类

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

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

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

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

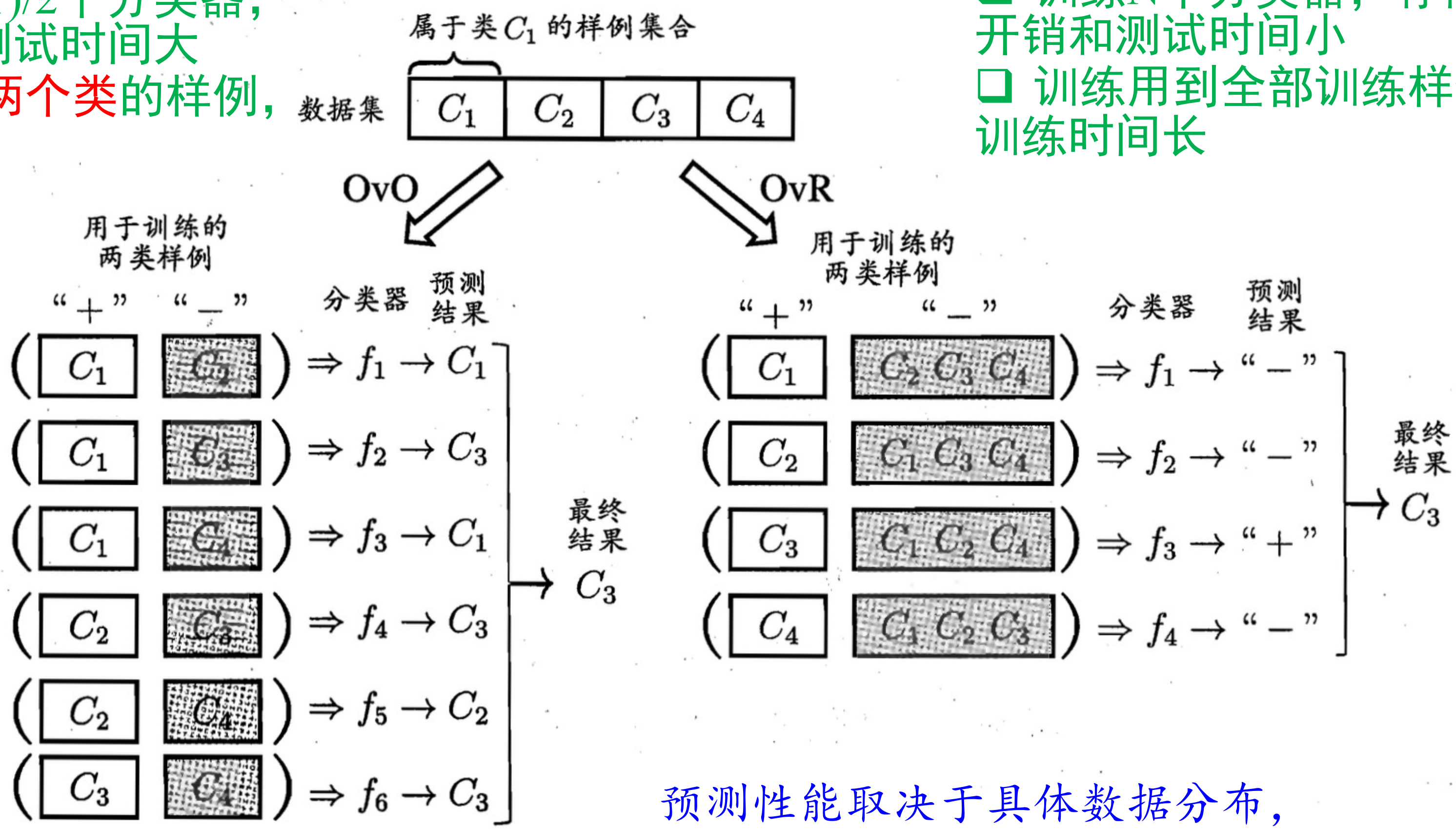
$\mathbf{W}$ 的闭式解是 $\mathbf{S}_w^{-1} \mathbf{S}_b$ 的 $N$-1 个最大广义特征值所对应的特征向量组成的矩阵

# 多分类学习

**拆解法：将一个多分类任务拆分为多个二分类任务求解**

- ❑ 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

# 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. ⇒ 若 $\frac{y}{1-y} > \frac{m^+}{m^-}$ 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)

# 作业

1. 证明极大对数似然估计中目标函数的等价性

$$\ell(\boldsymbol{w}, b)=\sum_{i=1}^{m} \ln p\left(y_i \mid \boldsymbol{x}_i ; \boldsymbol{w}, b\right) \quad \Longleftrightarrow \quad \ell(\boldsymbol{\beta})=\sum_{i=1}^{m}\left(-y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i+\ln \left(1+e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}\right)\right)$$

2. 若学习器 A 的 F1 值比学习器 B 高，试分析A的BEP值是否也比B高？举例说明。

3. 某数据集包含 1200 个样本，三个类别，其中A 类：500 个样本，B类：400 个样本，C类：300 个样本。
现需使用留出法（分层抽样）将其划分为训练集（80%）和测试集（20%），试计算满足条件的划分方式共有多少种？

# 谢谢！