第二章 线性模型

§2.1 背景知识

§2.2 回归问题

§2.3 分类方法

§2.4 复杂情形

§2.1 背景知识

一、问题引入

二、基本形式

问题引入

- 问题回顾
  - 假定收集了一批关于西瓜的数据，包括“色泽”、“根蒂”、“敲声”等属性值，我们如何学习出一个模型，以帮助我们在不剖开西瓜的前提下判断西瓜的成熟度？
  - 如果在预测连续值西瓜成熟度之外，预测的仅是“好”、“坏”这样的离散值，我们又该如何对其进行建模预测？

基本形式

- 基本形式
  - 给定由d个属性所描述的样本$x = (x_1; x_2; \ldots; x_d)$，其中$x_i$是$x$在第$i$个属性上的取值，线性模型试图学习出如下的函数：

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$

  - 一般用向量形式表示为：

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

其中$w = (w_1; w_2; \ldots; w_d)$

基本形式

- 线性模型特点
  - 形式简单、易于建模
  - 引入层级结构或高维映射可实现非线性建模
  - 可解释性强

$$f_{瓜}(x) = 0.2x_{色泽} + 0.5x_{根蒂} + 0.3x_{敲声} + 1$$

则意味着可以综合考虑三项属性判断瓜的好坏，而且根蒂最重要，敲声次之，色泽最次

§2.2 回归问题

一、简单线性回归

二、多元线性回归

三、对数线性回归

四、广义线性模型

线性回归

- 数据集$\mathbf{D}=\{(\mathbf{x_1},\mathbf{y_1}),(\mathbf{x_2},\mathbf{y_2}),\dots,(\mathbf{x_m},\mathbf{y_m})\}$，$\mathbf{x}=(\mathbf{x_{i1}};\mathbf{x_{i2}};\dots;\mathbf{x_{id}})$，$\boldsymbol{y_i}\in\mathbb{R}$，简单线性回归试图学一个线性模型以尽可能准确预测实值输出的标记
- 先考虑一种最简单的情形：输入属性的数目只有一个。为便于讨论，此时我们忽略关于属性的下标，即$\mathbf{D}=\{(\boldsymbol{x_i},\boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{m}}$，其中$\boldsymbol{x_i}\in\mathbb{R}$
- 线性回归试图学得：

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

- 那么要如何确定$\boldsymbol{w}$和$\boldsymbol{b}$呢？

线性回归

- 问题的关键在于如何衡量$f(x_i)$和y_i之间的差异
- 均方误差是线性回归任务中最常用的误差度量准则，因此可试图让均方误差最小化，即

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

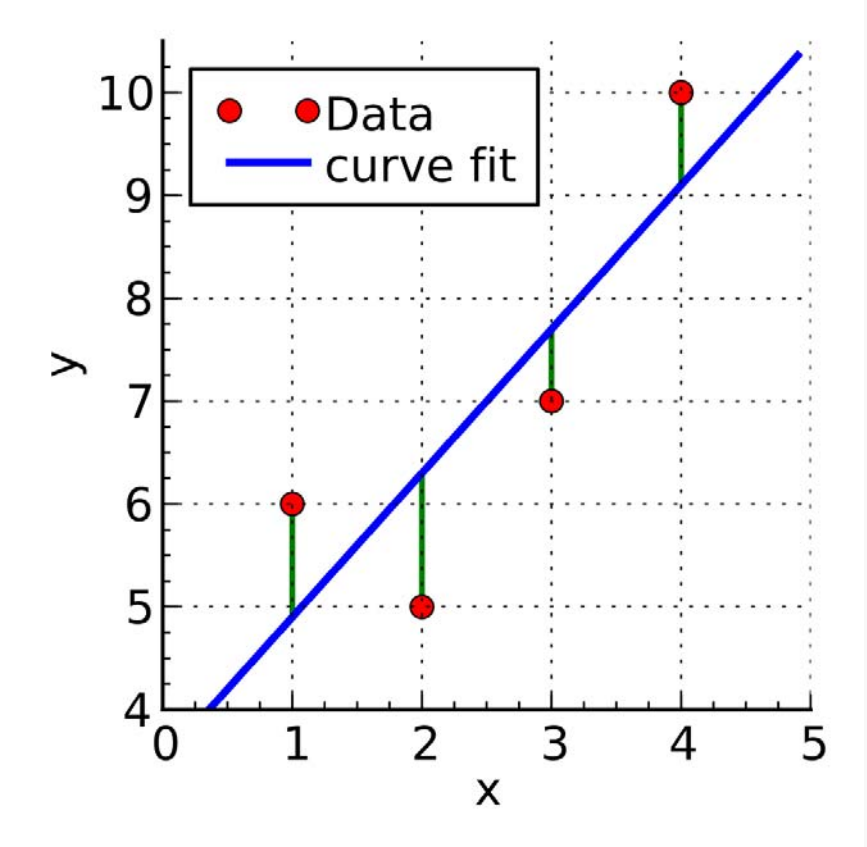
$$= \underset{(w,b)}{\operatorname{argmin}} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

- 均方误差有非常好的几何意义，它对应了常用的欧几里得距离或简称“欧氏距离”（Euclidean distance）

线性回归

- 基于均方误差最小化来进行模型求解的方法，一般称为“**最小二乘法**”（least square method）

- 在线性回归模型中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小



线性回归

- 首先令

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

- 求解w,b使$E(w,b)$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）
- 将$E(w,b)$对w求导，可得到：

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

线性回归

- 再将$E(w,b)$对b求导，可得到

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

- 令上述两式为0，可以得到最优解：

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

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

实质上就是经过样本的均值点

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

多元线性回归

- 更一般的情形：给定数据集D，每个样本由d个属性描述，此时我们试图学得：

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

- 这称为“多元线性回归”（multivariate linear regression），也称“多变量线性回归”。
- 类似的，我们同样可以用最小二乘法进行求解，只需要将之前的过程推广到矩阵形式

多元线性回归

- 我们把w和b吸收入向量形式：

$$\hat{w} = (w; b)$$

- 相应的，我们把数据集表示为一个矩阵：

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} x_1^{\mathrm{T}} & 1 \\ x_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ x_m^{\mathrm{T}} & 1 \end{pmatrix}$$

- 同时将标记也记为向量形式：

$$y = (y_1; y_2; \ldots; y_m),$$

多元线性回归

- 类似于线性回归，有：

$$\hat{\boldsymbol{w}}^* = \underset{\hat{\boldsymbol{w}}}{\operatorname{argmin}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$$

- 我们令：

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

- 同理，我们对$\hat{\boldsymbol{w}}$ 求导，可以得到：

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

- 我们能像线性回归一样求得最优解的闭式解吗？

多元线性回归

- 当$\mathbf{X}^{\mathbf{T}}\mathbf{X}$为**满秩矩阵**(full-rank matrix)或者**正定矩阵**(positive definite matrix)，令上式为0，有：

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

- 其中$\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}$为$\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)$的逆矩阵
- 令$\hat{x_i} = (x_i; 1)$，则最终学得的多元线性回归模型为：

$$\boldsymbol{f}(\hat{x_i}) = \hat{x_i}^{\mathrm{T}}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\boldsymbol{y}$$

- 如果$\mathbf{X}^{\mathbf{T}}\mathbf{X}$不为满秩矩阵呢？

多元线性回归

- 何时会出现非满秩矩阵情况？
- 在许多任务中我们会遇到大量的变量，其数目甚至超过样本数，此时会导致X的列数多于行，进而导致非满秩。
- 出现非满秩矩阵情况会怎样？
- 若X为非满秩矩阵，此时可解出多个最优解$\hat{\boldsymbol{w}}$，而且它们都能使均方误差$E_{\hat{\boldsymbol{w}}}$最小化。
- 面对多个最优解应如何选择？
- 选择哪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。

对数线性回归

- 对于样本$(x, y), y \in \mathbb{R}$，当我们希望线性模型的预测值逼近真实标记时，就得到线性回归模型，我们简写为：

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

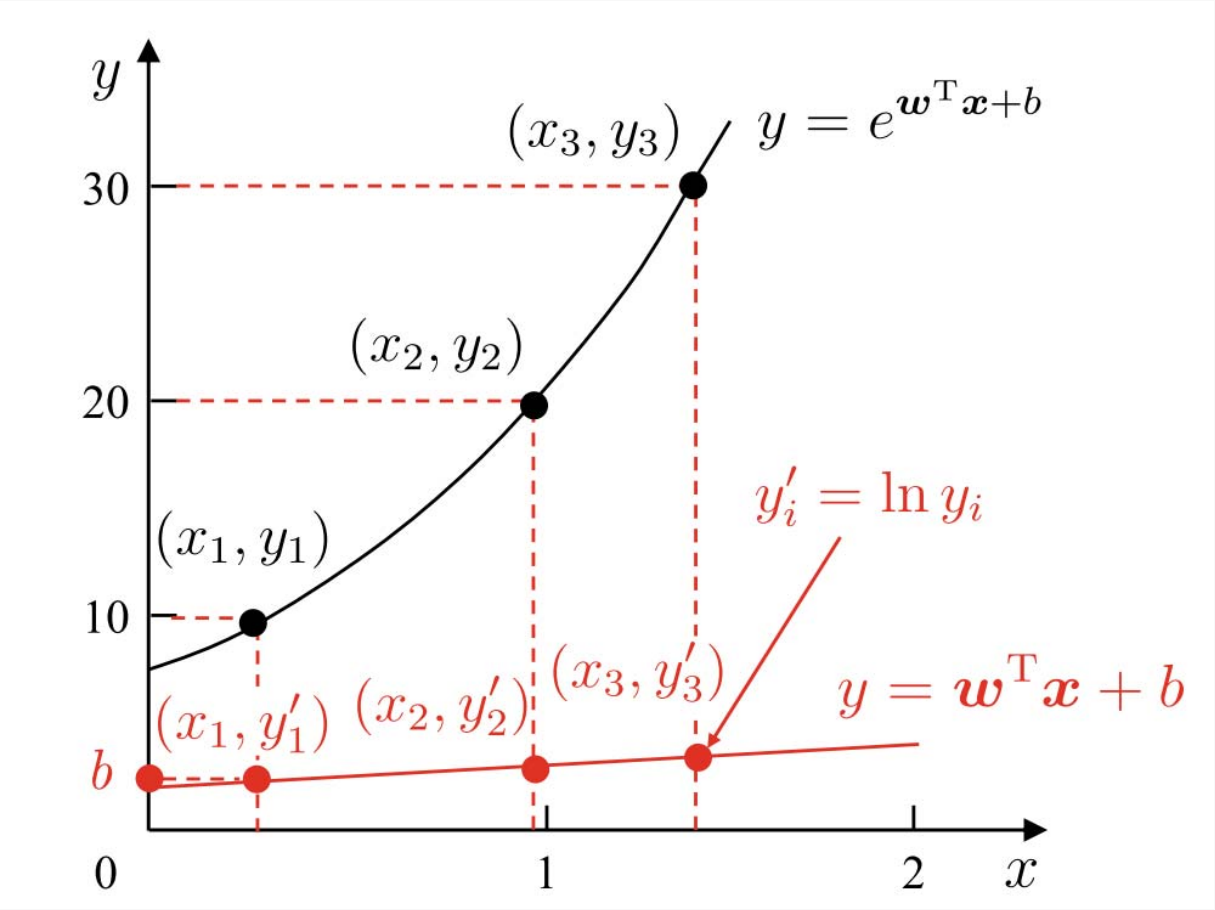
- 能否令模型预测值逼近y的衍生物呢?
- 假设我们认为样本所对应的输出标记是在指数尺度上变化，那就可将输出标记的对数作为线性模型逼近的目标，即：

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

- 这就是“对数线性回归”（log-linear regression）

对数线性回归

- 虽然上式在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
- 如图所示，这里的对数函数起到了将线性回归型的预测值与真实标记联系起来的作用



广义线性回归

- 类似于对数线性回归，更一般地我们可以考虑单调可微函数 $g(\cdot)$, 令：

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

- 其中$g(\cdot)$连续且充分光滑

- 这样得到的模型称为“**广义线性模型**”（generalized linear model），其中函数$g(\cdot)$称为“**联系函数**”（link function）。

- 显然对数线性回归是广义线性模型的特例，此时：

$$g(\cdot) = \ln(\cdot)$$

§2.3 分类方法

一、对数几率回归

二、线性判别分析

对数几率回归

- 我们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该怎么办?
- 此时我们不再需要预测连续实值，而是预测离散标记，我们可以先观察广义线性模型的形式：

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

- 通过广义线性模型，只需要找到一个单调可微函数，将分类任务的真实标记$\boldsymbol{y}$与线性回归模型的预测值联系起来，即可解决分类任务

对数几率回归

- 考虑二分类任务，其输出标记为：

$$y \in \{0, 1\}$$

- 而线性回归模型产生的预测值为实值：

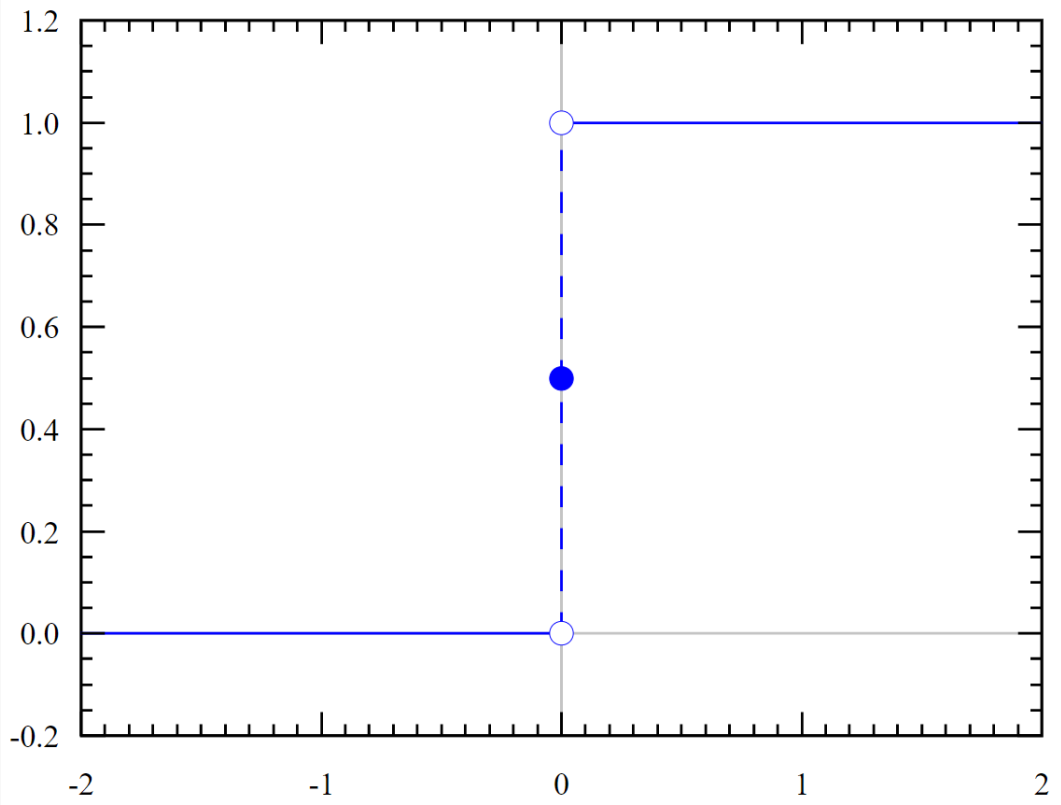
$$z = w^T x + b$$

- 若需将实值转化为标记，我们可以采用“单位阶跃函数”（unit-step function）：

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

对数几率回归

- 如图所示，所以预测值大于零就判为正例，小于零则判为反例，预测值为临界值零则可任意判别
- 单位阶跃函数有什么问题？



对数几率回归

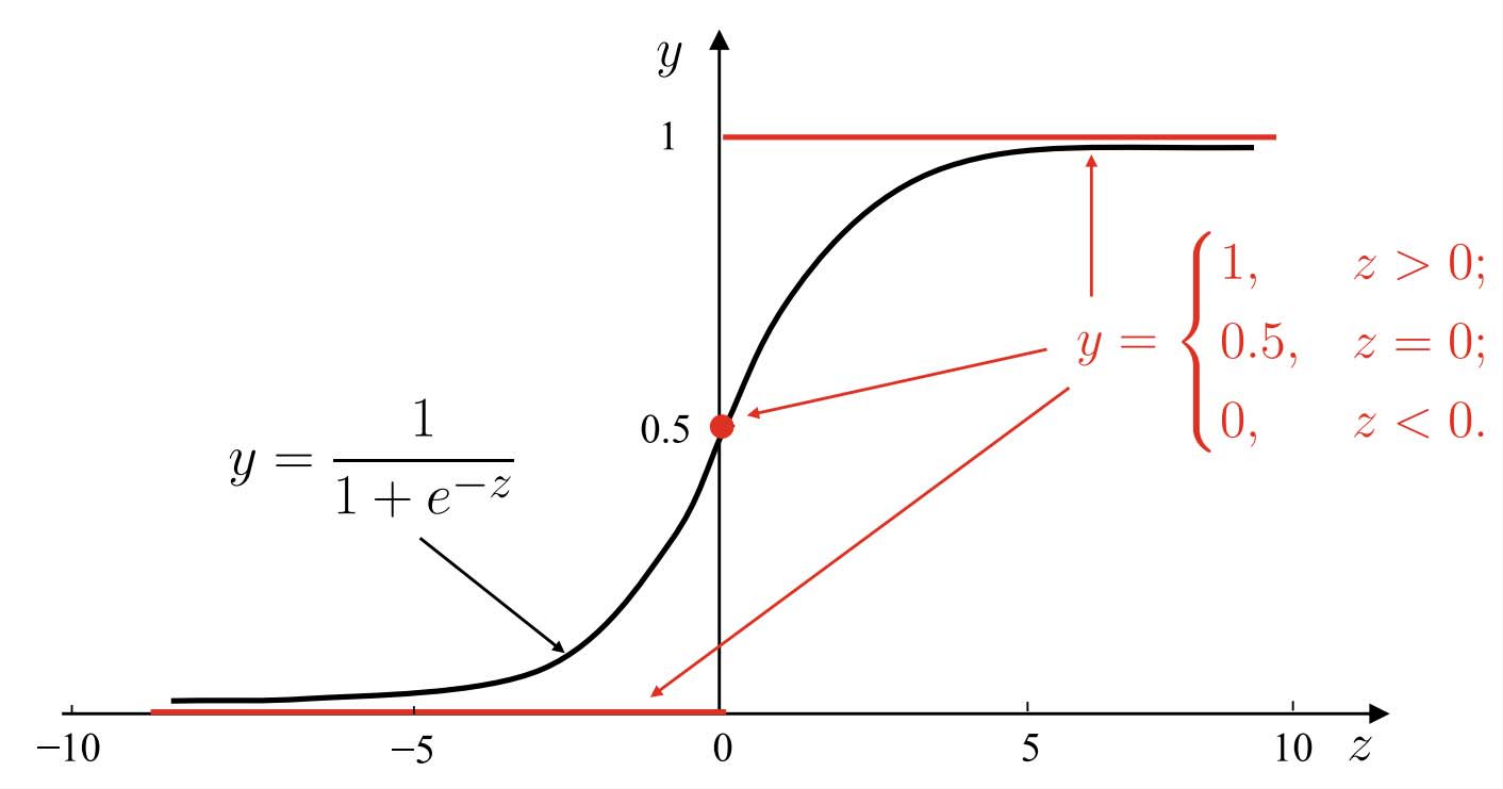
- 单位阶跃函数不连续，求逆运算和微分运算存在一定的问题
- 找到能在一定程度上近似单位阶跃函数的“替代函数”（surrogate function），并希望它单调可微。而对数几率函数(logistic function)正是这样一个常用的替代函数：

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

- 对数几率函数又简称“对率函数”
- 注意对数几率函数与“对数函数”$ln(\cdot)$不同

对数几率回归

- 对数几率函数是一种“Sigmoid函数”，它将z值转化为一个接近0或者1的y值，并且其输出值在$z=0$附近变化很陡：



对数几率回归

- 如果将对数几率函数作为$g^{-}(\cdot)$函数代入广义线性回归模型，我们可以得到：

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

- 类似于对数线性回归，上式可变化为：

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

- 上式有何物理意义？

对数几率回归

- 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，而对几率取对数则得到“对数几率”（log odds，亦称logit）：

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

- 上式在用线性回归模型的预测结果去逼近真实标记的对数几率，因此，其对应的模型称为“对数几率回归”（logistic regression，亦称logit regression）

对数几率回归

- 对数几率回归有哪些特点？

- 虽然名为“回归”，但实际却是一种分类学习方法
- 是直接对分类可能性进行建模，无需事先假设数据分布
- 不仅预测出"类别"，而是可得到近似概率预测
- 对率函数是任意阶可导的凸函数，有很好的数学性质

对数几率回归

- 如何确定对数几率回归模型的参数呢？
- 我们首先得到了对数几率表达式：

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

- 将上式中的y视为类后验概率估计$p(y=1\,|\boldsymbol{x})$，则上式可重写为：

$$\ln \frac{p(y=1\,|\boldsymbol{x})}{p(y=0\,|\boldsymbol{x})} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

- 显然有：

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

对数几率回归

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

- 可通过“极大似然法”（maximum likelihood method）来估计 w和$\boldsymbol{b}$，给定数据集$\{(x_i, y_i)\}_{i=1}^{m}$，对率回归模型最大化“对数似然”（log-likelihood）：

$$\boldsymbol{l}(w,b) = \sum_{i=1}^{m} \boldsymbol{lnp}(\boldsymbol{y_i}|\boldsymbol{x_i}; w, b)$$

- 物理意义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

- 为便于讨论，令：

$$\boldsymbol{\beta} = (w; b), \hat{\boldsymbol{x}} = (\boldsymbol{x}; \mathbf{1})$$

- 则：

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

- 再令：

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

- 我们可以重写似然项：

$$p(y_i|\boldsymbol{x_i}; w, b) = y_i p_1(\hat{\boldsymbol{x_i}}; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x_i}}; \boldsymbol{\beta})$$

对数几率回归

- 将新的似然项代回对数似然，可以发现最大化对数似然等价于最小化下式：

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

- 上式是关于$\boldsymbol{\beta}$的高阶可导连续凸函数，经典的数值优化算法如**梯度下降法**(gradient descent method)、**牛顿法**(Newton method)等都可求得其最优解，于是就得到：

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, l(\boldsymbol{\beta})$$

对数几率回归

- 以牛顿法为例，其第$t+1$轮选代解的更新公式为：

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

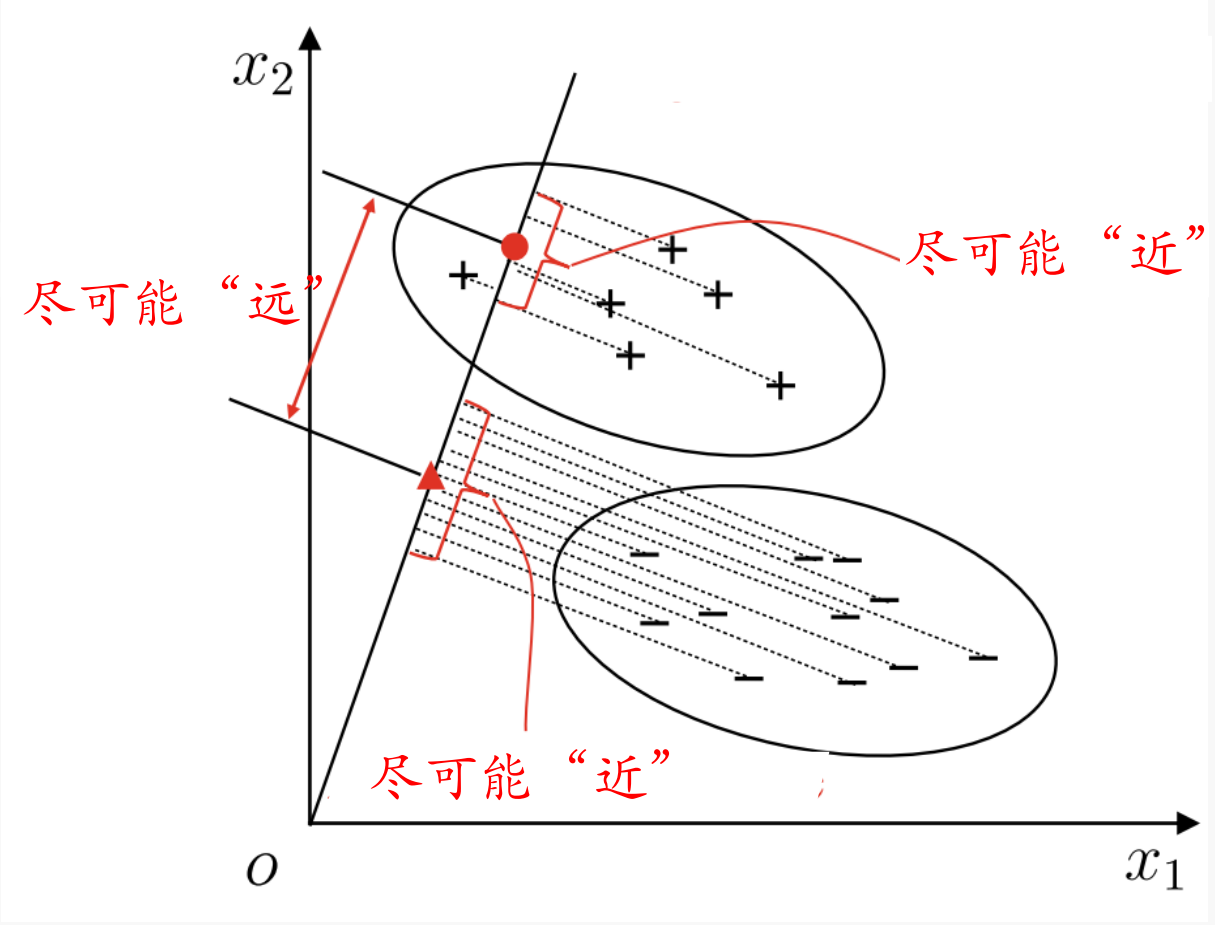
- 其中关于$\boldsymbol{\beta}$的一阶、二阶导数分别为：

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=-\sum_{i=1}^{m} \hat{\boldsymbol{x}}_{i}\left(y_{i}-p_{1}\left(\hat{\boldsymbol{x}}_{i} ; \boldsymbol{\beta}\right)\right)$$

$$\frac{\partial^{2} l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}=\sum_{i=1}^{m} \hat{\boldsymbol{x}}_{i} \hat{\boldsymbol{x}}_{i}^{\mathrm{T}} p_{1}\left(\hat{\boldsymbol{x}}_{i} ; \boldsymbol{\beta}\right)\left(1-p_{1}\left(\hat{\boldsymbol{x}}_{i} ; \boldsymbol{\beta}\right)\right)$$

线性判别分析

- **线性判别分析**（Linear Discriminant Analysis），简称LDA，是一种经典线性学习方法。
- LDA的思想非常朴素：给定训练样本集，将样本投影到一条使得同类样本的投影点尽可能接近、异类样本投影点尽可能远离。



线性判别分析

- 给定数据集$\mathbf{D} = \{(x_i, y_i)\}_{i=1}^{m}, y_i \in \{0, 1\}$，令$X_i$、$\mu_i$、$\Sigma_i$分别表示第$i \in \{0, 1\}$类样本的集合、均值向量、协方差矩阵。
- 若将数据投影到直线w上，则两类样本的中心在直线上的投影分别为：

$$w^{\mathrm{T}}\mu_0、w^{\mathrm{T}}\mu_1$$

- 若将所有样本点投影到直线上，则两类样本的协方差为：

$$w^{\mathrm{T}}\Sigma_0 w、w^{\mathrm{T}}\Sigma_1 w$$

- 直线为一维空间，因此上面四项均为实数。

线性判别分析

- 要想使得同类样本的投影点尽可能接近、异类样本投影点尽可能远离矩阵，我们可以考虑最大化下面的目标：

$$J = \frac{||\boldsymbol{w}^{\mathrm{T}}\mu_0 - \boldsymbol{w}^{\mathrm{T}}\mu_1||_2^2}{\boldsymbol{w}^{\mathrm{T}}\Sigma_0\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\Sigma_1\boldsymbol{w}}$$

$$= \frac{\boldsymbol{w}^{\mathrm{T}}(\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\Sigma_0 + \Sigma_1)\boldsymbol{w}}$$

- 我们定义“**类间散度矩阵**”（between-class scatter matrix）：

$$\mathbf{S}_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}}$$

线性判别分析

- 我们定义“类内散度矩阵”（within-class scatter matrix）：

$$\mathbf{S}_w = \Sigma_0 + \Sigma_1$$

$$= \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^{\mathrm{T}} + \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^{\mathrm{T}}$$

- 则目标可重写为：

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

- 这就是LDA最大化的目标，即$\mathbf{S}_b$和$\mathbf{S}_w$的“广义瑞利商”（generalized Rayleigh quotient）。

线性判别分析

- 注意到目标的分子分母都是关于w的二次项，因此解与w的长度无关，只与其方向有关，所以目标等价于：

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

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

- 由拉格朗日乘子法，上式等价于：

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

- 其中λ是拉格朗日乘子，又有：

$$\mathbf{S}_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}}$$

线性判别分析

- 我们不妨令：

$$\mathbf{S}_b w = \lambda(\mu_0 - \mu_1)$$

- 所以可得：

$$\lambda(\mu_0 - \mu_1) = \lambda \mathbf{S}_w w$$

- 所以可得：

$$w = \mathbf{S}_w^{-1}(\mu_0 - \mu_1)$$

- 后续课程我们会学习，LDA可从贝时斯决策理论的角度来阐释，并可证明，当两类数据同先验、满足高斯分布且协方差相等时，LDA可达到最优分类

线性判别分析

- 如何将LDA 推广到多分类任务中？
- 假定存在N个类别，且第i类样本数为m_i，我们首先定义“全局散度矩阵”：

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$$

$$= \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^{\mathrm{T}}$$

- 其中μ是所有样本的均值向量，接着重新定义类内散度矩阵：

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

线性判别分析

- 其中：

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

- 所以可得：

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

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

- 显然，多分类LDA可以有多种实现方法：使用$\boldsymbol{S_b}$，$\boldsymbol{S_t}$，$\boldsymbol{S_w}$三者中的任何两个即可

线性判别分析

- 常见的一种实现是采用优化目标：

$$\max_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W})}{\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W})}$$

- 其中$\boldsymbol{W} \in \mathbb{R}^{d\times(N-1)}$，$\boldsymbol{tr}(\cdot)$表示矩阵的迹(trace)，上式可通过如下问题求解：

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

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

§3.4 复杂情形

一、类别不平衡问题

二、应用实例

二、思考题

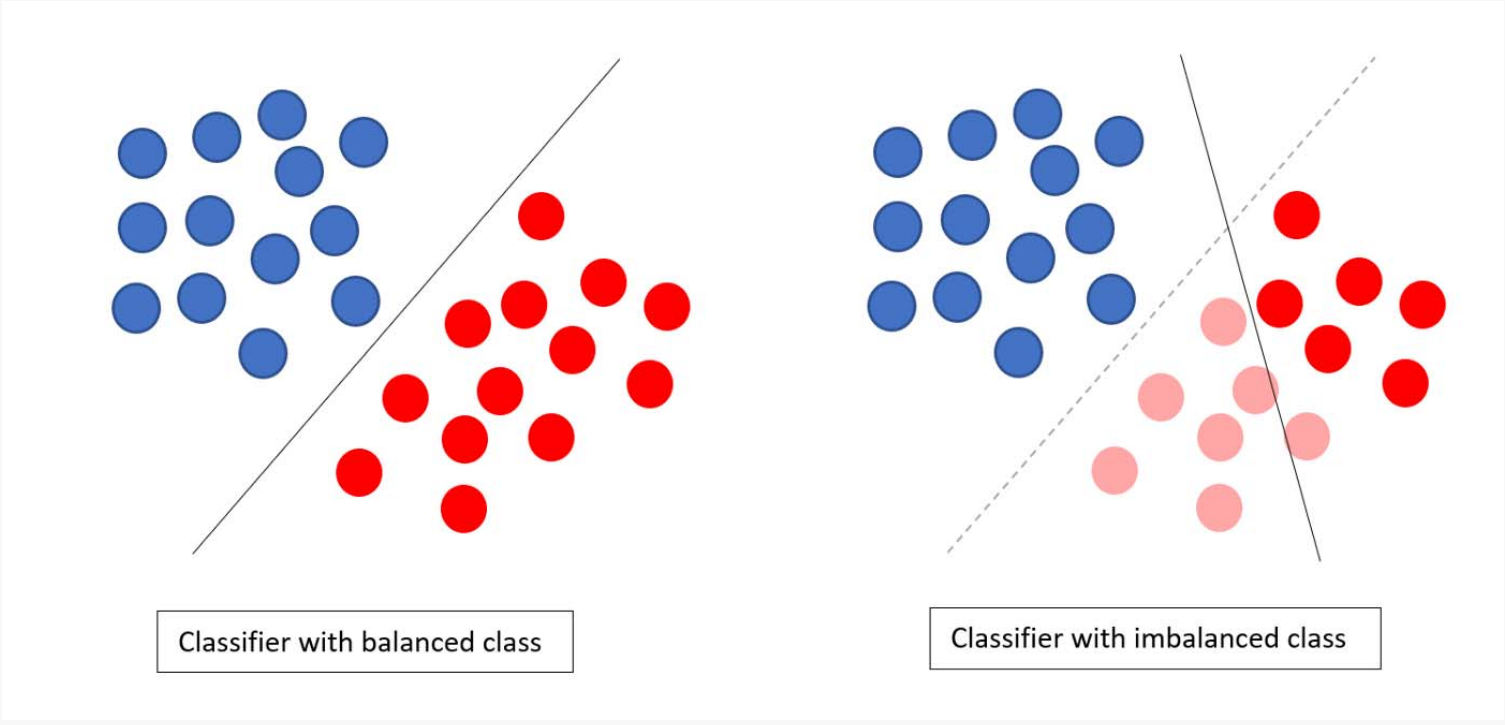
类别不平衡问题

- 前面介绍的分类学习方法都有什么问题？
- 上述分类学习方法基于共同的基本假设：不同类别的训练样例数目相当。如果不同类别的训练样例数目稍有差别，通常影响不大，但若差别很大，则会对学习过程造成困扰。

- 如果训练集内共1000个样本，有998个负样本，但正样本只有2个，会发生什么？

类别不平衡问题

- 类别不平衡(class-imbalance)就是指分类任务中不同类别的训练样本数目差别很大的情况。在现实的分类学习任务中，我们经常会遇到类别不平衡。



类别不平衡问题

- 我们从线性分类器的角度讨论这个问题，在我们用线性回归模型对新样本进行分类时，一般会估计出：

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

- 但我们仍需用预测出的$\boldsymbol{y}$值与一个阈值进行比较，应该如何选取恰当的阈值？
- 通常会设置分类器的决策规则为：

$$\text{若}\ \frac{\boldsymbol{y}}{\boldsymbol{1}-\boldsymbol{y}} > \boldsymbol{1}\ \text{则 预测为正样本}$$

- 此时等价于将阈值设置为0.5

类别不平衡问题

- 当训练集中正、负样本的数目不同时，令m^+表示正样本数目，m^-表示负样本数目，则观测几率为：

$$p = \frac{m^+}{m^-}$$

- 所以为了应对类别不平衡问题，我们应该重新设置分类器的决策规则为：

$$\text{若}\ \frac{y}{1-y} > \frac{m^+}{m^-}\ \text{则 预测为正样本}$$

- 如果仍用通常的决策规则，应如何处理？

类别不平衡问题

- 只需令：

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

- 此即类别不平衡学习的一个基本策略：“再缩放”（rescaling）
- 再缩放思想简单，但实际操作却并不平凡，因为“训练集是真实样本总体的无偏采样”这个假设并不总成立，也就是说我们未必能有效地基于训练集观测几率来推断出真实分布几率
- 再缩放也是“代价敏感学习”（cost-sensitive learning）的基础

类别不平衡问题

- 处理类别不平衡问题的三种方法
  - 若假定正样本较少，负样本较多：
  - “欠采样”（undersampling）直接对训练集里的负样本进行操作，即去除一些负样本，使得正、负样本数目接近，然后再进行学习
  - “过采样”（oversampling）对训练集里的正样本进行操作，即增加一些正样本，使得正、负样本数目接近，然后再进行学习
  - 最后一种是直接基于原始训练集进行学习，但在用训练好的分类器进行预测时，将再缩放的技巧嵌入到其决策过程中，称为“阈值移动”（threshold-moving）

应用实例

- Cost-Sensitive Subspace Learning for Face Recognition
- 在人脸识别系统里，一般将合法人员误识别为非法人员会带来不便，将非法人员合法人员则会带来风险，两种情况要承担代价完全不同的损失。

Jiwen Lu and Yap-Peng Tan, Cost-sensitive subspace learning for face recognition, CVPR 2010.

应用实例

- 我们首先定义出人脸识别系统的损失矩阵，即定义出将第i个人误识别为第j个人的损失。

	G_1	$\cdots$	G_c	I
G_1	0	$\cdots$	C_{GG}	C_{GI}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
G_c	C_{GG}	$\cdots$	0	C_{GI}
I	C_{IG}	$\cdots$	C_{IG}	0

- 之后我们需要定义出重要性函数来刻画出来自不同类别的样本重要性：

$$f(k) = \begin{cases} (c-1)C_{GG} + C_{IG} & \text{if } k = 1, 2, \cdots, c \\ cC_{IG} & \text{otherwise} \end{cases}$$

应用实例

- 接着我们就可以将LDA中的类内、类间距离改写为：

$$\tilde{S}_B = \sum_{k_1=1}^{c+1} \sum_{k_2=1}^{c+1} cost(k_1, k_2)(m_{k_1} - m_{k_2})(m_{k_1} - m_{k_2})^T$$

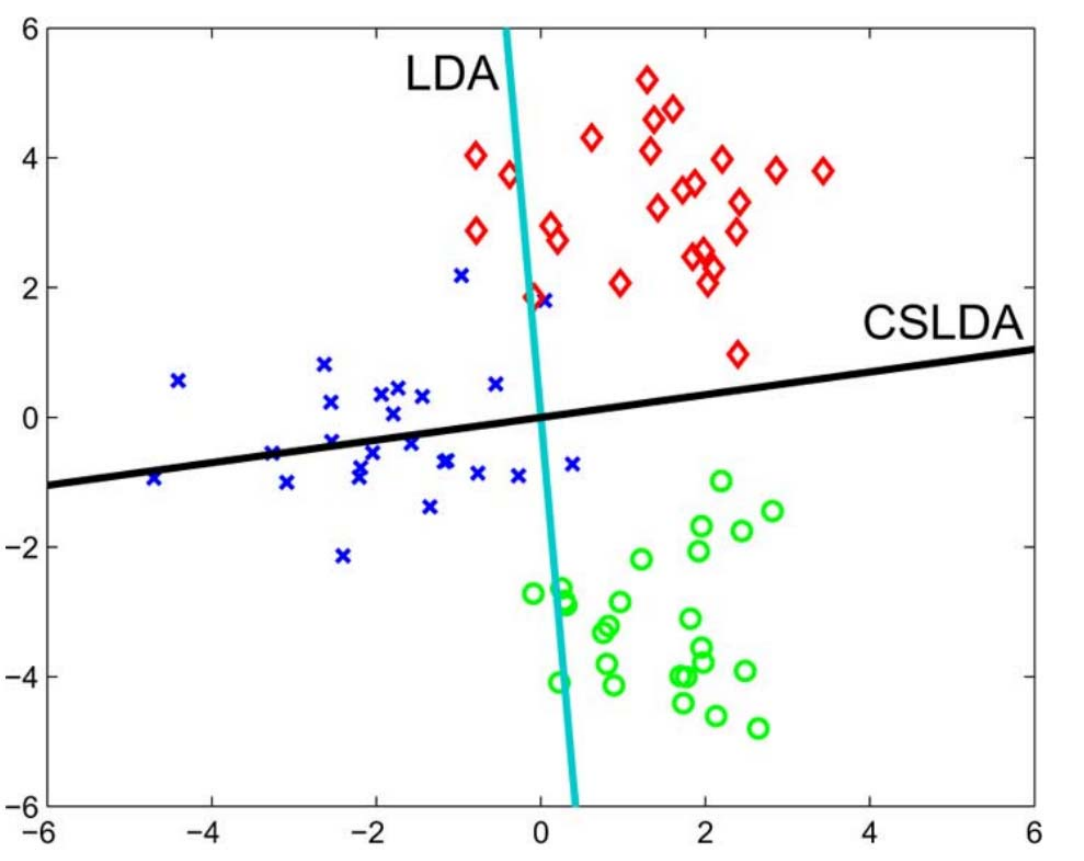
$$\tilde{S}_W = \sum_{k=1}^{c+1} \sum_{l(x_i)=k} f(k)(x_i - m_k)(x_i - m_k)^T$$

- 然后我们就可以类似于LDA对下式求解，求得CSLDA的解：

$$\tilde{S}_B \tilde{w} = \lambda \tilde{S}_W \tilde{w}$$

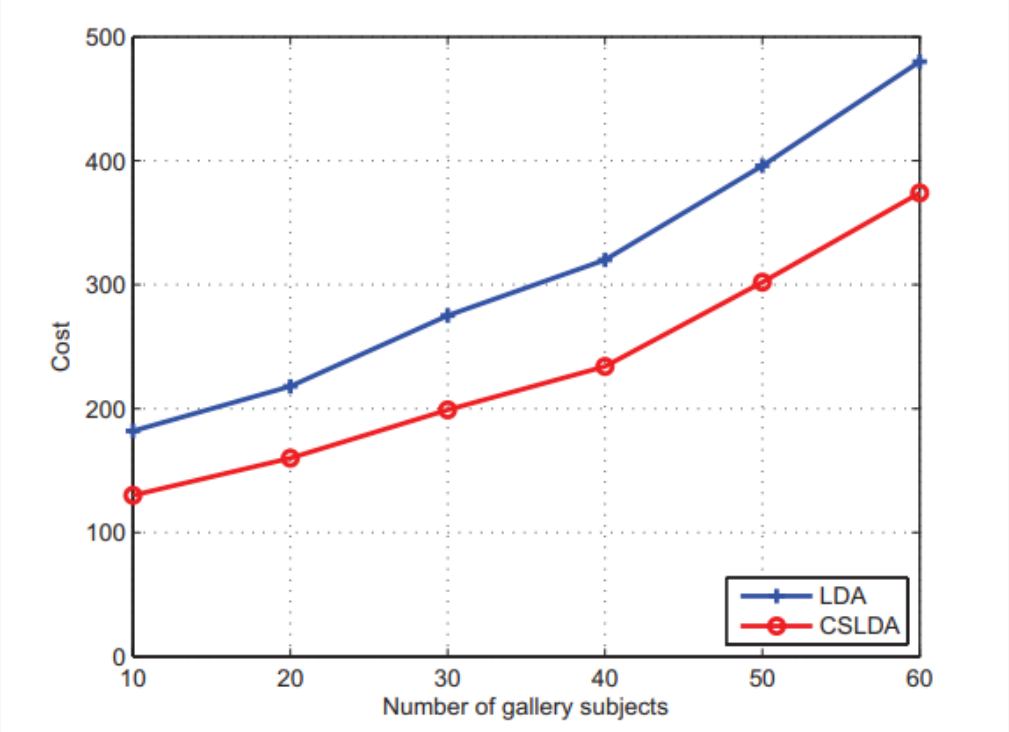
应用实例

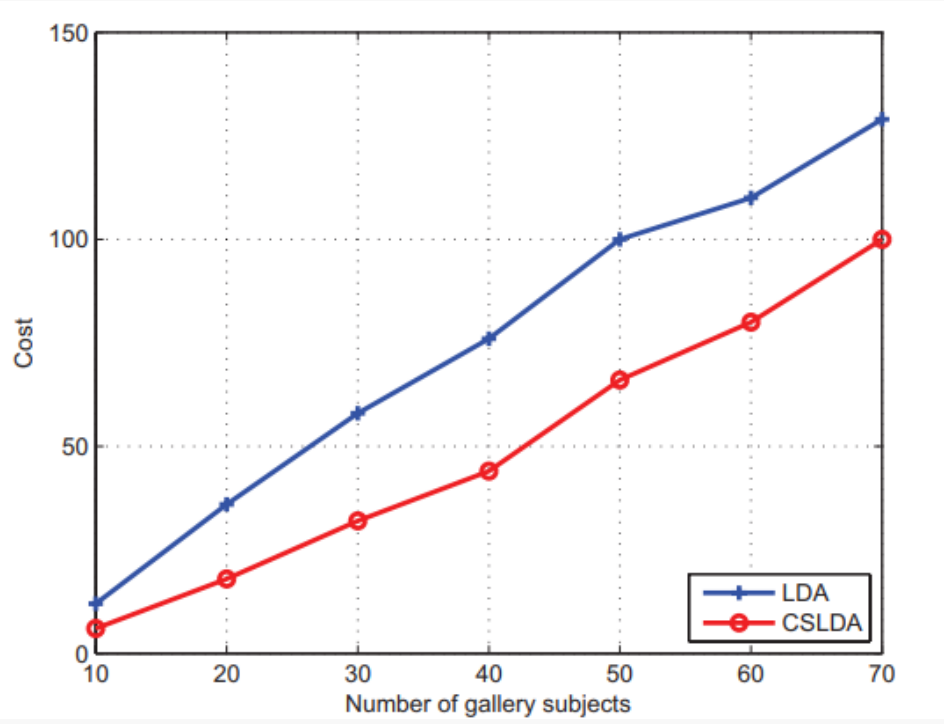
- 如果有A（红）、B（绿）、C（蓝）三类样本，而且C样本误判损失很低，A、B损失很高，CSLDA可以获得比LDA更优分类器



应用实例

- 在AR人脸数据集与FERET人脸数据集上分别计算出不同训练样本条件下的总损失，可以通过对比发现我们的CSLDA方法可以有效降低损失。





思考题1

- 下表展示了一群人$m=15$的身高x_i/体重y_i数据，对$(x_i,y_i)_{i=1}^{n}$进行线性回归

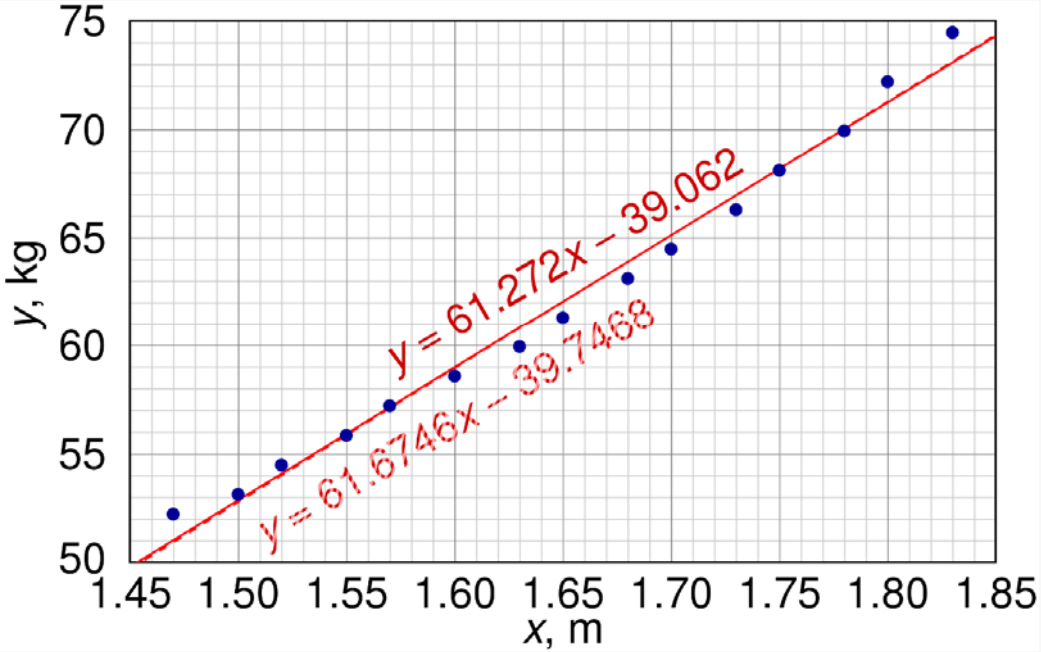
H(m)	1.47	1.50	1.52	1.55	1.57	1.60	1.63	1.65	1.68	1.70	1.73	1.75	1.78	1.80	1.83
M(kg)	52.21	53.12	54.48	55.84	57.20	58.57	59.93	61.29	63.11	64.47	66.28	68.10	69.92	72.19	74.46

$$\bar{x}=\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)=1.65,$$

$$\sum_{i=1}^{m}x_i^2=41.0532$$

$$w=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}=61.272$$

$$b=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)=-39.062$$



思考题2

- 下表展示了一群学生$m = 20$的复习时间x_i和是否通过考试y_i的数据，通过对数几率回归分析复习时间如何影响通过考试的几率

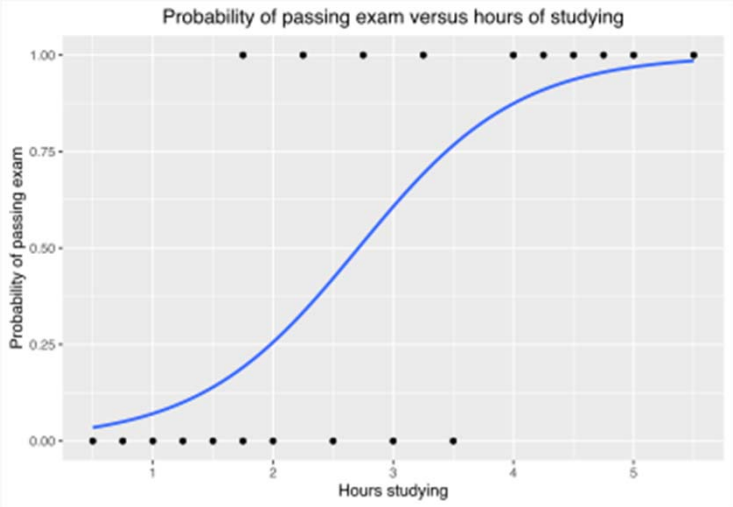
Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1

$$p(y=1\,|\boldsymbol{x}) = \frac{e^{wx+b}}{1+e^{wx+b}} \qquad p(y=0\,|\boldsymbol{x}) = \frac{1}{1+e^{wx+b}}$$

$$l(w,b) = \sum_{i=1}^{m}\Big(y_i \ln(p_i) + \big(1-y_i\big)\ln(1-p_i)\Big)$$

$$\frac{\partial l}{\partial w} = \sum_{i=1}^{m}(y_i - p_i)x_i = 0 \qquad \frac{\partial l}{\partial b} = \sum_{i=1}^{m}(y_i - p_i) = 0$$

$$w \approx 1.5 \qquad b \approx -4.1$$



思考题3

- 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，使其能较好地周于非线性可分数据。