机器学习初步

——线性分类模型

李爽

助理教授，特聘副研究员

数据科学与知识工程研究所

E-mail: shuangli@bit.edu.cn

Homepage: shuangli.xyz

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

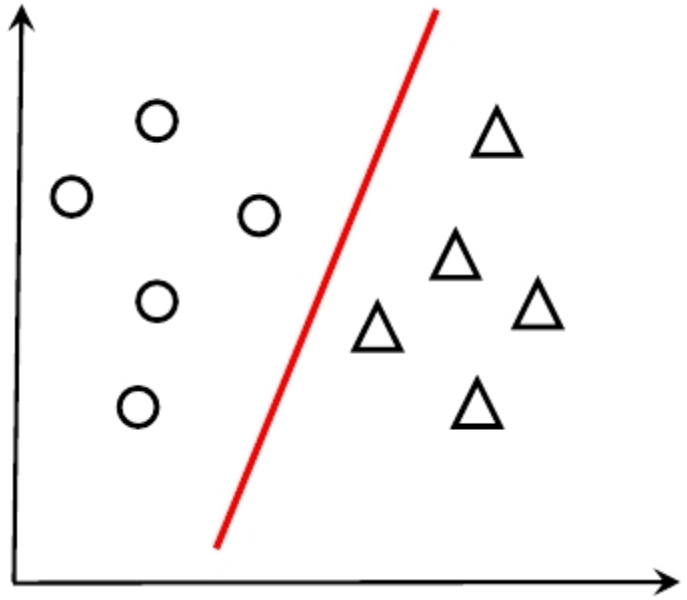
线性分类模型

- 线性模型基础
- 线性回归
- 对数几率回归
- 线性判别分析
- 多分类学习
- 类别不平衡问题

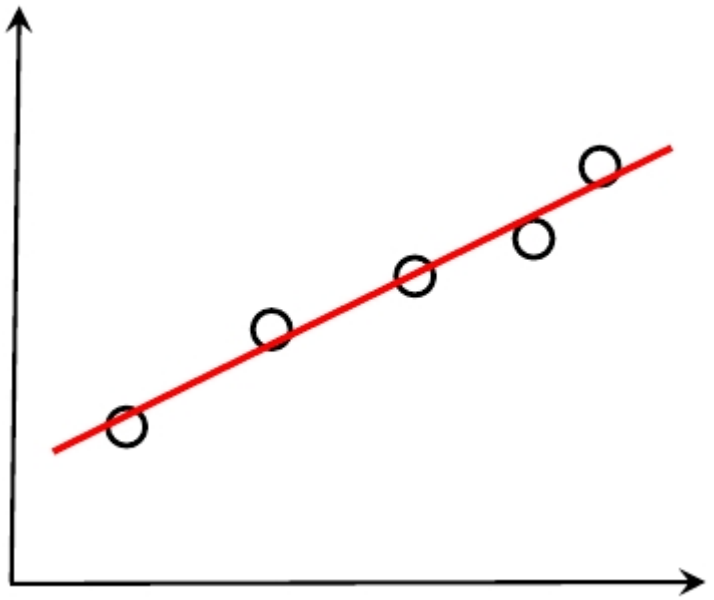
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

一、线性模型基础

1. 线性模型应用：



线性模型应用：分类



线性模型应用：回归

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

一、线性模型基础

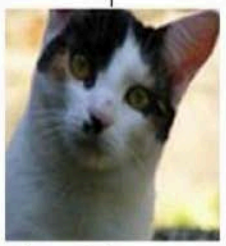
图像分类

(32x32x3)

image parameters

$f(x,W)$

输出图片在每个类别上的得分



input image

stretch pixels into single column

0.2	-0.5	0.1	2.0
1.5	1.3	2.1	0.0
0	0.25	0.2	-0.3

W

56
231
24
2

x_i

+

1.1
3.2
-1.2

b

-96.8	cat score
437.9	dog score
61.95	ship score

$f(x_i; W, b)$

一、线性模型基础

2. 线性模型基本形式：

- 给定有d个属性描述的示例$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$，其中$x_i$是$\boldsymbol{x}$在第$i$个属性上的取值，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

$$f(\boldsymbol{x}) = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_d x_d + b$$

- 一般用向量形式表示写成：

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

- 其中$\boldsymbol{\omega} = (\omega_1; \omega_2; \ldots; \omega_d)$. $\boldsymbol{\omega}$和b学得之后，模型就得以确定

一、线性模型基础

3. 线性模型特点：

- 线性模型简单、易于建模；
- $\boldsymbol{\omega}$直观表达了各属性在预测中的重要性，使得线性模型有很好的解释性
  - 例如：西瓜有3个属性(色泽，根蒂，敲声)，我们学得如下模型

$$f_{\text{妃瓢}}(\boldsymbol{x}) = 0.2 \cdot x_{\text{艸汊}} + 0.5 \cdot x_{\text{枞酋}} + 0.3 \cdot x_{\text{數壺}} + 1$$

则意味着可通过综合考虑色泽、根蒂和敲声来判断瓜好不好，其中根蒂最要紧，而敲声比色泽更重要

线性分类模型

- 线性模型基础
- 线性回归
- 对数几率回归
- 线性判别分析
- 多分类学习
- 类别不平衡问题

二、线性回归

1. 回归

起源：19世纪80年代, 英国统计学家弗朗西斯. 高尔顿提出

研究：父代身高与子代身高之间的关系

结论：子代的身高有向族群平均身高“回归”的趋势。

➢身高有回归于“中心”的趋势

大自然有一种约束力，使人类身高在一定时期是相对稳定的。

➢实际身高与“中心”存在偏差

二、线性回归

回归模型的一般形式：

- 设因变量y，自变量向量$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$，则刻画$y$与$\boldsymbol{x}$关系的回归模型的一般形式为

$$y = f(\boldsymbol{x}) + \varepsilon$$

其中，ε为随机误差，它表示除了$\boldsymbol{x}$外的其它随机干扰因素。

二、线性回归

2. 线性回归

- 定义：

  线性回归（linear regression）指的是y与$\boldsymbol{x}$之间是线性关系

- 目标：

  线性回归试图学得一个线性模型以尽可能准确地预测实值输出标记

- 参数定义：

  数据集$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$,其中$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id}), y_i \in \mathbb{R}$

- 常见的线性回归有两种：一元线性回归 和 多元线性回归

二、线性回归

3. 一元线性回归

- 一元线性回归输入属性的数目只有一个，即$D = \{(x_i, y_i)\}_{i=1}^{m}$，则线性回归试图学得：

$$f(x_i) = \omega x_i + b, 使得 f(x_i) \simeq y_i$$

- 如何求解ω和b?

答：最小化均方误差(MSE)，即

$$(\omega^*, b^*) = \arg\min_{(\omega,b)} \sum_{i=1}^{m} (y_i - \omega x_i - b)^2$$

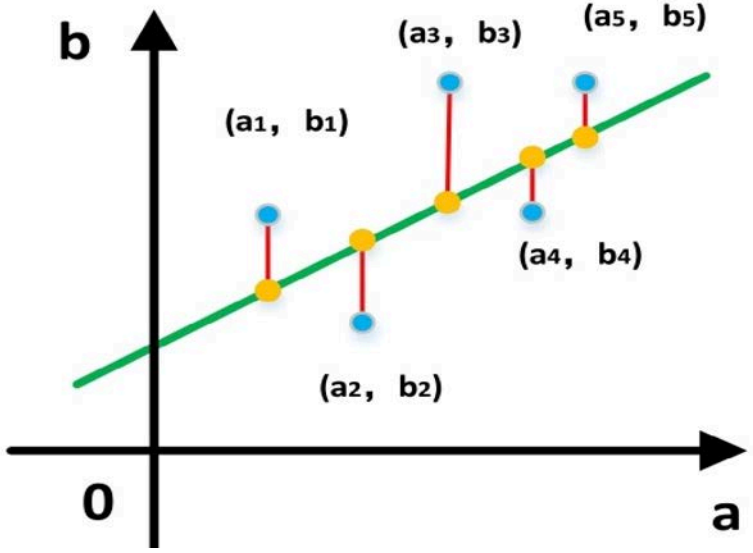
ω^*, b^*表示ω咒b的解

$$= \arg\min_{(\omega,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

二、线性回归

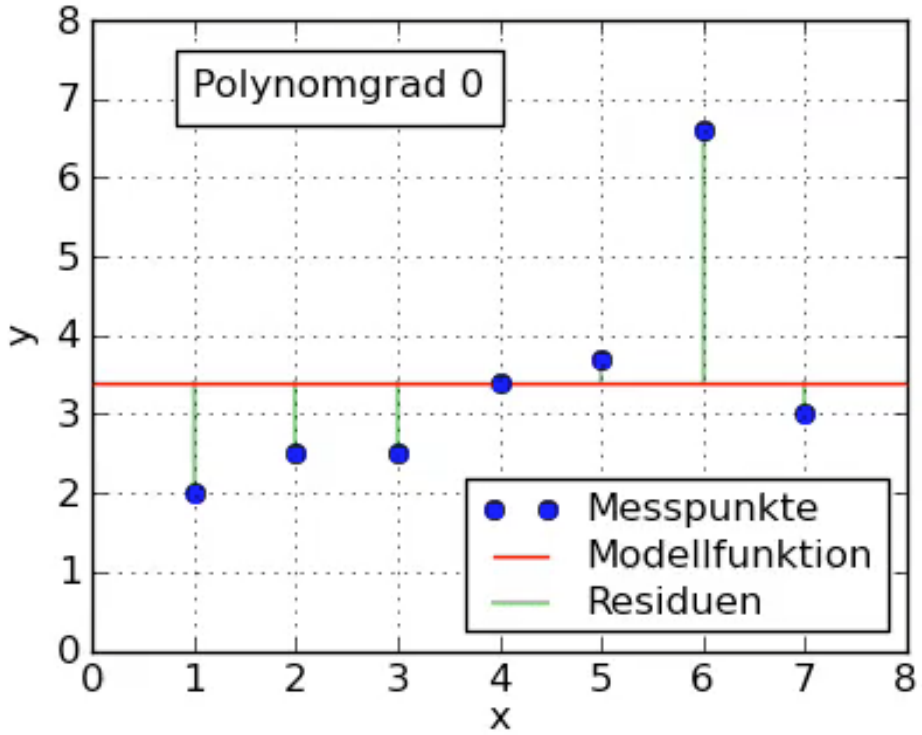
最小二乘法

- 均方误差是回归任务中最常用的性能度量，其具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为“欧式距离”。
- 基于均方误差最小化来进行模型求解的方法称为“最小二乘法”。在线性回归中，最小二乘法就是试图找到一条直线，是所有样本到直线上的欧氏距离之和最小



二、线性回归

- 同一组数据，选择不同的$f(x)$，通过最小二乘法可以得到不一样的拟合曲线, 如下图所示



二、线性回归

- 求解ω和b使$E_{(\omega,b)}=\sum_{i=1}^{m}(y_i-\omega x_i-b)^2$最小化的过程，称为线性回归模型的最小二乘“参数估计”。将$E_{(\omega,b)}$分别对ω和b求导，得到

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

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

二、线性回归

- 令上述导数为零，可得到ω和b 的闭式(closed-form)解

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

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

其中 $\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ 为x的均值

二、线性回归

4. 多元线性回归

- 多元线性回归样本由d个属性描述，即$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$，数据集为 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$，此时我们试图学得

$$f(\boldsymbol{x}_i) = \omega^T \boldsymbol{x}_i + b, \text{佼值} f(\boldsymbol{x}_i) \simeq y_i$$

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

二、线性回归

- 类似的，利用最小二乘法来对ω和b进行估计。
- 为了便于讨论，把ω和b吸收入向量形式$\hat{\omega} = (\omega, b)$,同时把数据集$D$表示为一个$m \times (d+1)$大小的矩阵$\boldsymbol{X}$, 每一行对应于一个示例，该行前$d$个元素对一个于示例的$d$个属性值，最后一个元素恒置为1，即

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

二、线性回归

- 将标签也写成向量形式$\boldsymbol{y} = (y_1; y_2; \ldots; y_m)$, 采用最小二乘法有：

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

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

$$\frac{\partial E_{\hat{\omega}}}{\partial \hat{\omega}} = 2\,\boldsymbol{X}^T (\boldsymbol{X}\hat{\omega} - \boldsymbol{y})$$

令上式为零可得$\hat{\omega}$。

二、线性回归

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

由于上式涉及矩阵逆的计算，需要讨论：

- 当$\boldsymbol{X}^T\boldsymbol{X}$为满秩矩阵或正定矩阵时，上式为零可得

$$\hat{\omega}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- 而现实任务中，$\boldsymbol{X}^T\boldsymbol{X}$往往不是满秩矩阵。
- 例如在许多任务中会遇到变量数大于样例数，导致$\boldsymbol{X}$的列数多于行数，$\boldsymbol{X}^T\boldsymbol{X}$显然不满秩，此时会解出多个$\hat{\omega}$，它们都能使均方误差最小化。
- 选择哪一个解作为输出？

  常用做法：引入正则化(regularization)项

二、线性回归

5. 对数线性回归

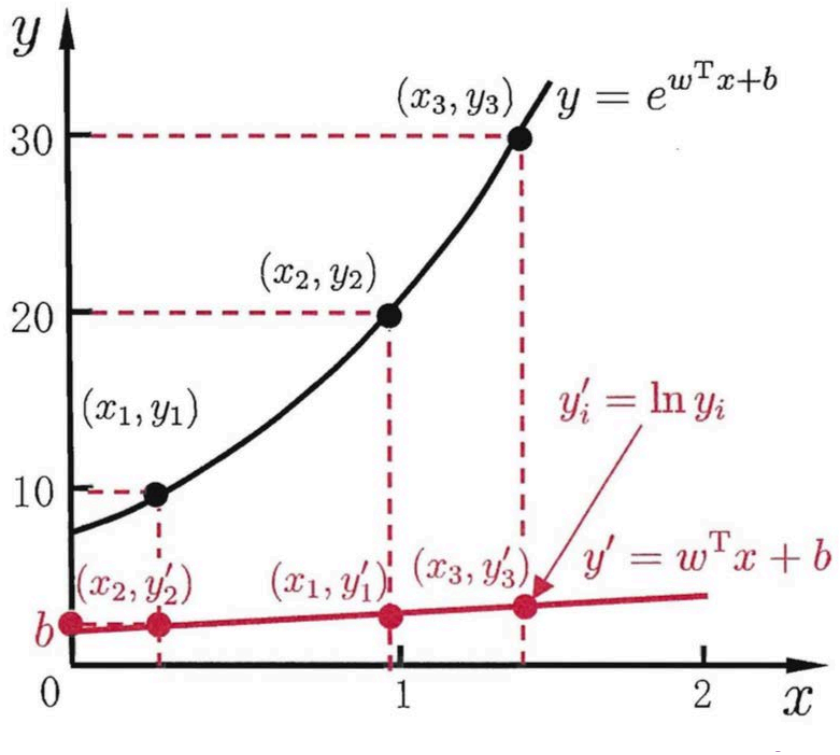
- 对于样例$(\boldsymbol{x}, y)$, $y \in R$, 若希望线性模型的预测值接近真实标记，则得到线性回归模型

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

- 思考：能否令模型的预测值逼近y的衍生物呢？

  若令 $\ln y = \omega^T \boldsymbol{x} + b$, 则得到“对数线性回归”

- 它实际上是在试图让$e^{\omega^T \boldsymbol{x}+b}$逼近$y$. $\ln y = \omega^T \boldsymbol{x} + b$ 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，如右图所示. 这里的对数函数起到了将线性回归模型的预测值与真实标记联系起来的作用.



二、线性回归

6. 广义线性模型

- 考虑单调可微函数$g(\cdot)$,令

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

- 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”.
- 显然，对数线性回归是广义线性模型在$g(\cdot) = \ln(\cdot)$时的特例

线性分类模型

- 线性模型基础
- 线性回归
- 对数几率回归
- 线性判别分析
- 多分类学习
- 类别不平衡问题

三、对数几率回归

1. 线性模型解决分类任务

- 思考：上一节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?

  答：答案蕴涵在此公式中：

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

- 只需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.

三、对数几率回归

- 以二分类任务为例，其输出标记$y \in \{0,1\}$

线性回归模型预测值 $z = \boldsymbol{\omega}^T \boldsymbol{x} + b$ —— 如何转化？ ——> 0/1 值

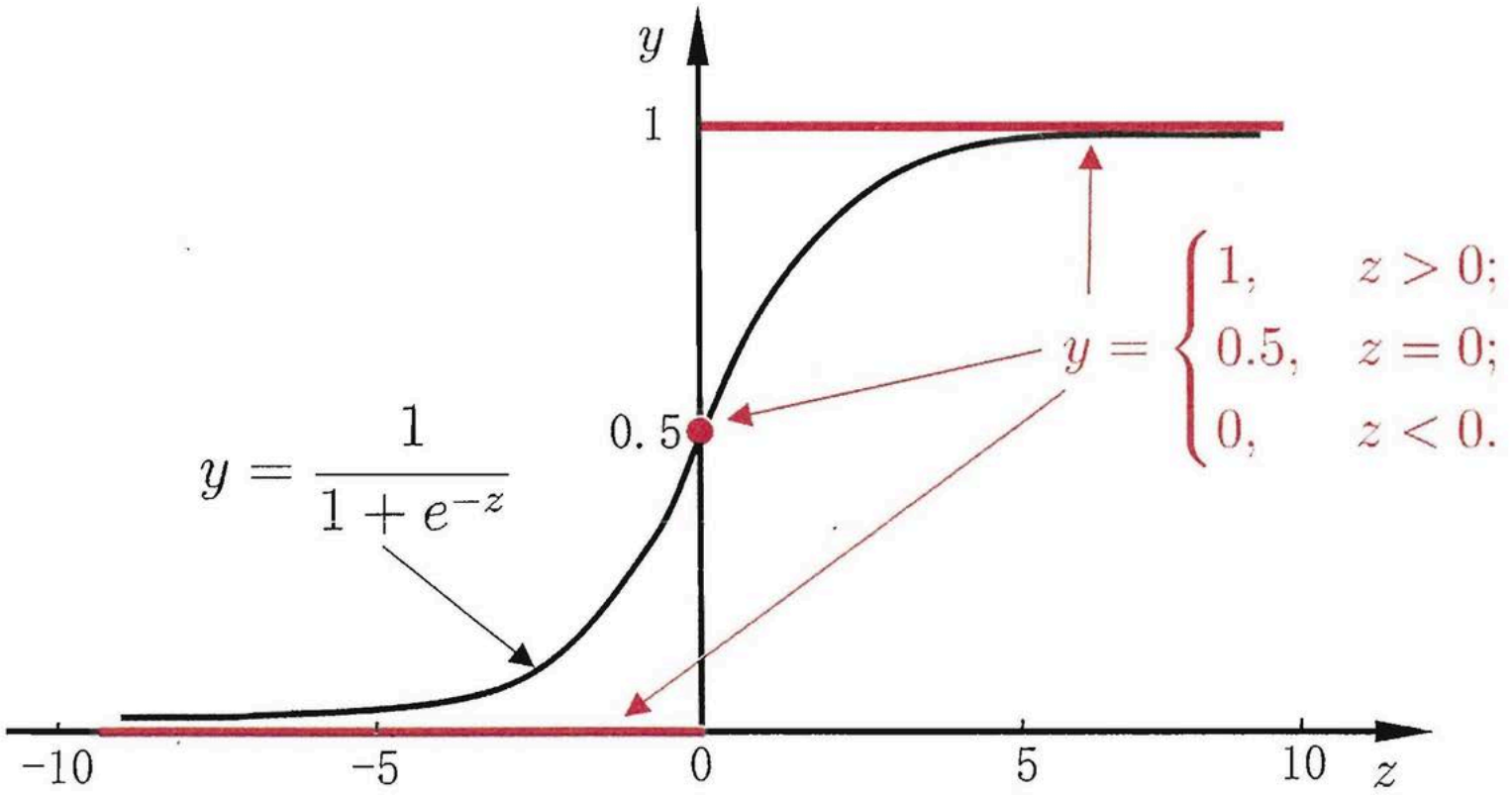
单位阶跃函数

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

- 即若预测值z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，如下图所示.

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

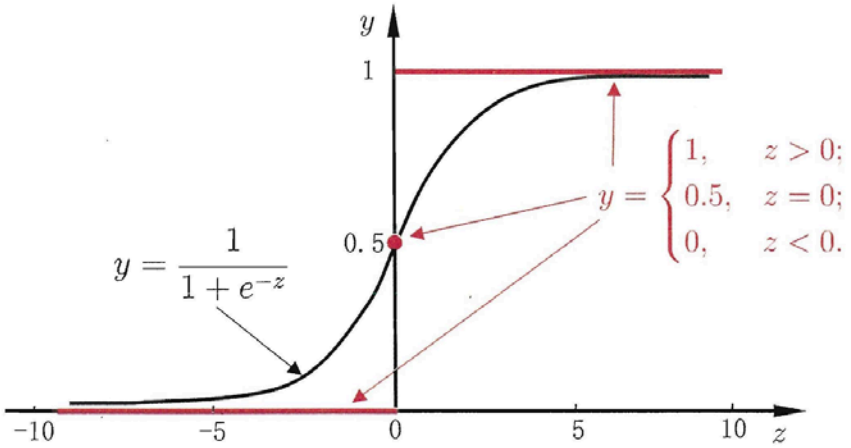
三、对数几率回归



红色线（点）为单位跃阶函数，黑色线为对数几率函数

三、对数几率回归

2. 对数几率函数



单位跃阶函数存在的问题：

- 单位阶跃函数不连续，不能直接用作式$y = g^{-1}(\boldsymbol{\omega}^T \boldsymbol{x} + b)$中的$g^{-1}(\cdot)$.

解决思路：

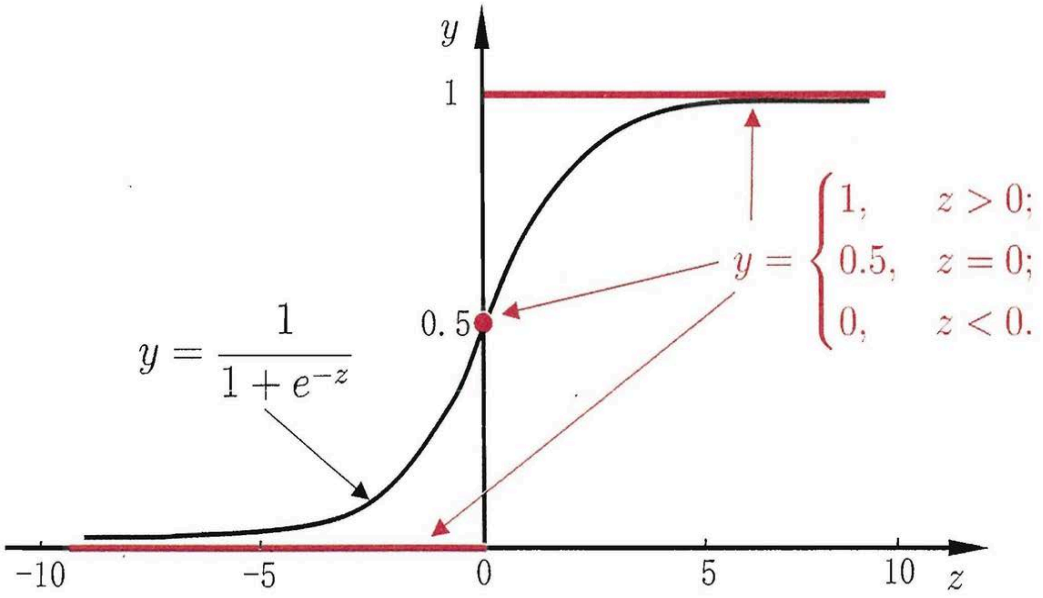
- 找一个在一定程度上能够近似单位跃阶函数的替代函数，并且它单调可微
- 这个替代函数就是对数几率函数，简称为对率函数 $y = \frac{1}{1+e^{-z}}$

三、对数几率回归

对数几率函数是一种“Sigmoid函数”

- Sigmoid函数基本性质
- 定义域：$(-\infty,+\infty)(-\infty,+\infty)$
- 值域：$(-1,1)(-1,1)$
- 函数在定义域内为连续和光滑函数

- Sigmoid函数特性：
- 函数的取值在0-1之间，
- 在0.5处为中心对称，
- 越靠近x=0的取值斜率越大。



三、对数几率回归

- 将对数几率函数作为$g = (\cdot)$代入式$y = g^{-1}(\boldsymbol{\omega}^T\boldsymbol{x} + b)$, 得到

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

变形得：

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

三、对数几率回归

- 若将y视为样本$\boldsymbol{x}$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$\boldsymbol{x}$作为正例的相对可能性。对几率取对数则得到“对数几率”(log odds,亦称logit)

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

三、对数几率回归

- 结论：

公式 $y=\frac{1}{1+e^{-(\omega^T x+b)}}$ 实际上是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”.

对数几率回归实际是一种分类学习方法

- 对数几率回归的优点：

- 直接对分类可能性进行建模，无需事先假设数据分布
- 它不是仅预测出“类别”，而是可得到近似概率预测，这对许多需利用概率辅助决策的任务很有用；
- 此外，对率函数是任意阶可导的凸函数，有很好的数学性质，现有的许多数值优化算法都可直接用于求取最优解.

三、对数几率回归

4. 对数几率回归参数求解

- 将公式 $y = \frac{1}{1+e^{-(\boldsymbol{\omega}^T \boldsymbol{x}+b)}}$中的$y$视为类后验概率估计$p(y = 1|\boldsymbol{x})$, 则式$\ln \frac{y}{1-y} = \boldsymbol{\omega}^T \boldsymbol{x} + b$可重写为

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

显然有

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

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

三、对数几率回归

- 可通过“极大似然法”来估计ω和b. 给定数据集$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$, 对率回归模型最大化“对数似然”,

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

似然项

- 目的是令每个样本属于其真实标记的概率越大越好.

三、对数几率回归

- 为便于讨论, 令$\boldsymbol{\beta} = (\boldsymbol{\omega}; b)$, $\hat{\boldsymbol{x}} = (\boldsymbol{x}; 1)$, 则$\boldsymbol{\omega}^T\boldsymbol{x} + b$可简写为$\boldsymbol{\beta}^T\hat{\boldsymbol{x}}$. 再令

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

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

- 则上述的似然项可重写为

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

三、对数几率回归

经过推导，得到如下3个公式↓

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

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

将左侧3个公式带入下式，问题从最大化转化成最小化

最大化似然函数

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

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

最小化

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

三、对数几率回归

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

- 上式是关于$\boldsymbol{\beta}$的高阶可导连续凸函数, 根据凸优化理论, 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

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

三、对数几率回归

- 以牛顿法为例，第$t+1$轮迭代解的更新公式为

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

- 其中关于$\boldsymbol{\beta}$的一阶、二阶导数分别为

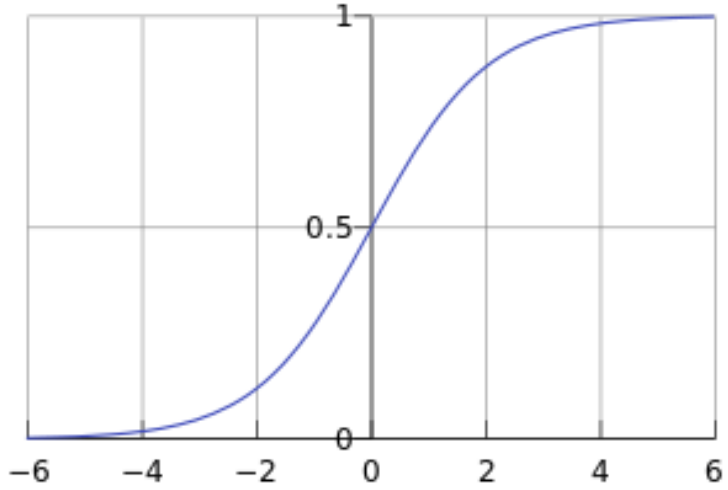
$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i (y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}))$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^T p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})(1 - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}))$$

三、对数几率回归

5. 实例

- 在电商推荐系统中的应用



- 对数几率回归可用于估计某事件的可能性，如某用户购买某商品的可能性、广告被某用户点击的可能性等。如图所示，对数几率回归公式将事件可能性限制在0到1之间，对应事件发生的概率。

实现方法：

- 通过对数几率回归，得到某用户购买某品牌的可能性，最终按照这个可能性排序来取top-k进行推荐。

三、对数几率回归

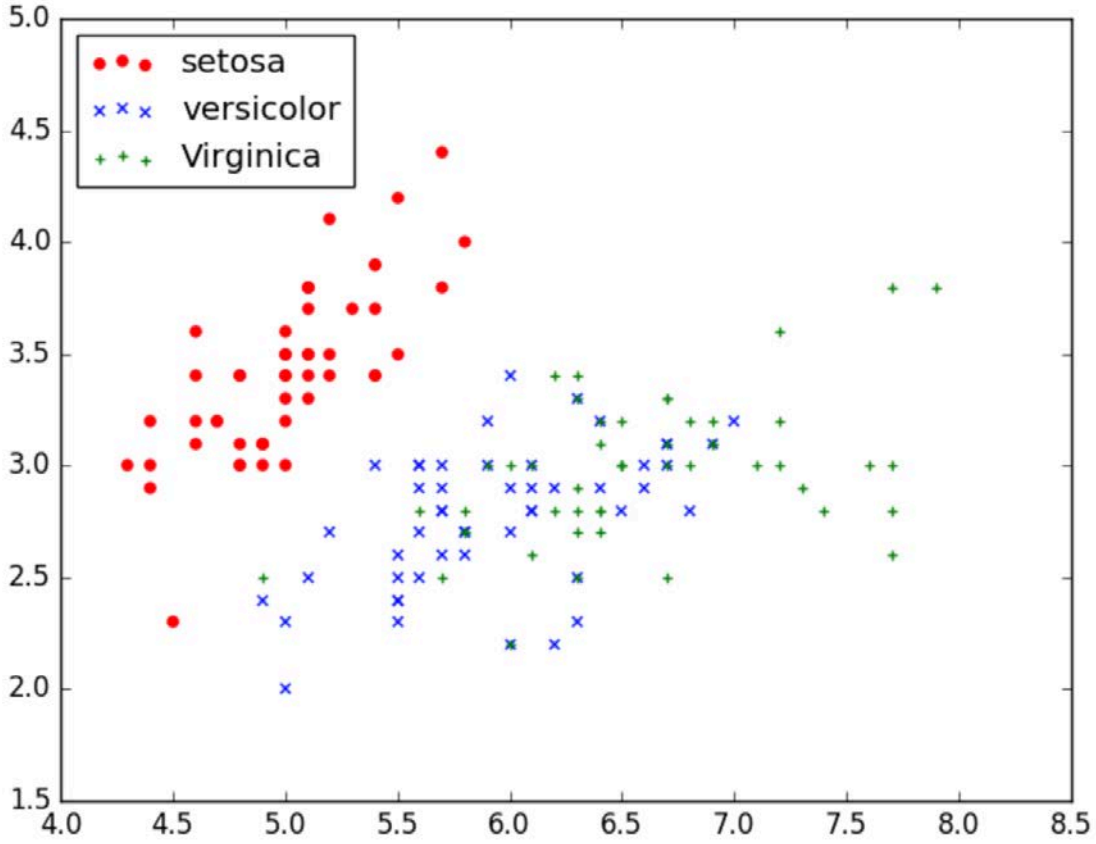
- 莺尾花数据集分类

① 数据集：样本x有4个属性值，类别标签y有3个，共150条数据

列名	说明	类型
SepalLength	花萼长度	float
SepalWidth	花萼宽度	float
PetalLength	花瓣长度	float
PetalWidth	花瓣宽度	float
Class	类别变量。0 表示山鸢尾，1 表示变色鸢尾，2 表示维吉尼亚鸢尾。	int

三、对数几率回归

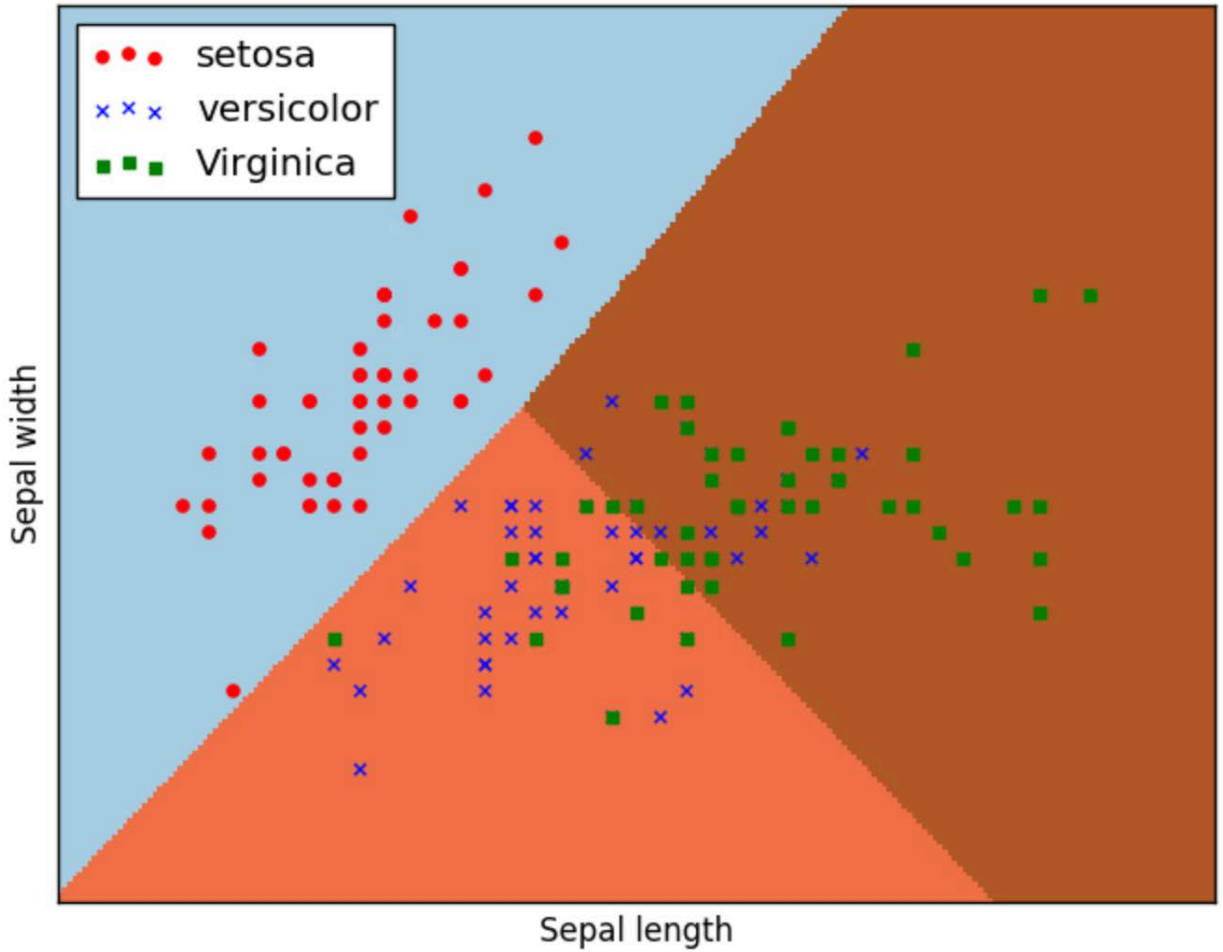
② 散点图：从4个属性值中取2个值来画散点图



从图中可以看出，数据集线性可分的，可以划分为3类，分别对应三种类型的鸢尾花，可采用对数几率回归对其进行分类预测

三、对数几率回归

③ 分类结果：



线性分类模型

- 线性模型基础
- 线性回归
- 对数几率回归
- 线性判别分析
- 多分类学习
- 类别不平衡问题

四、线性判别分析

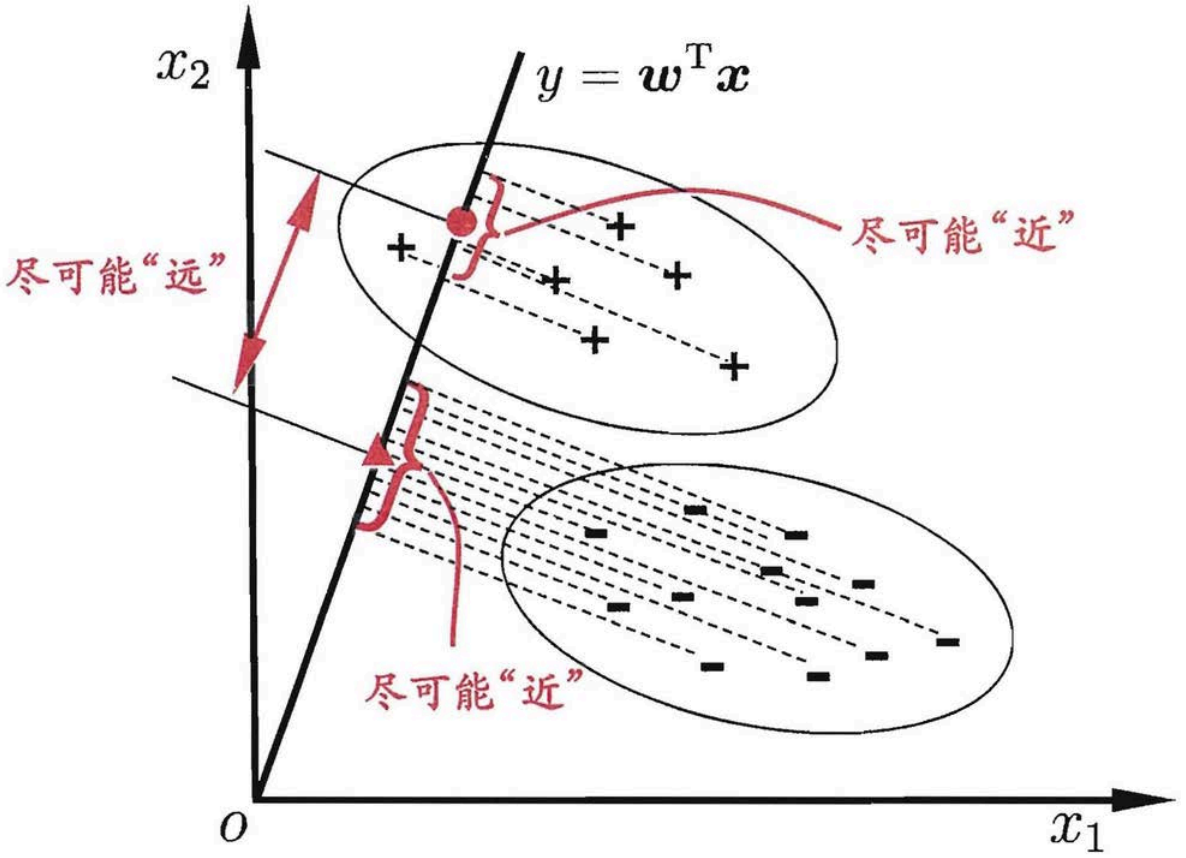
1. 介绍

- 线性判别分析(Linear Discriminant Analysis, 简称LDA)是一种经典的线性学习方法，在二分类问题上因为最早由[Fisher, 1936]提出, 亦称“Fisher 判别分析”

LDA的思想:

- 给定训练样例集，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;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
- 简言之，“投影后类内方差最小，类间方差最大”

四、线性判别分析



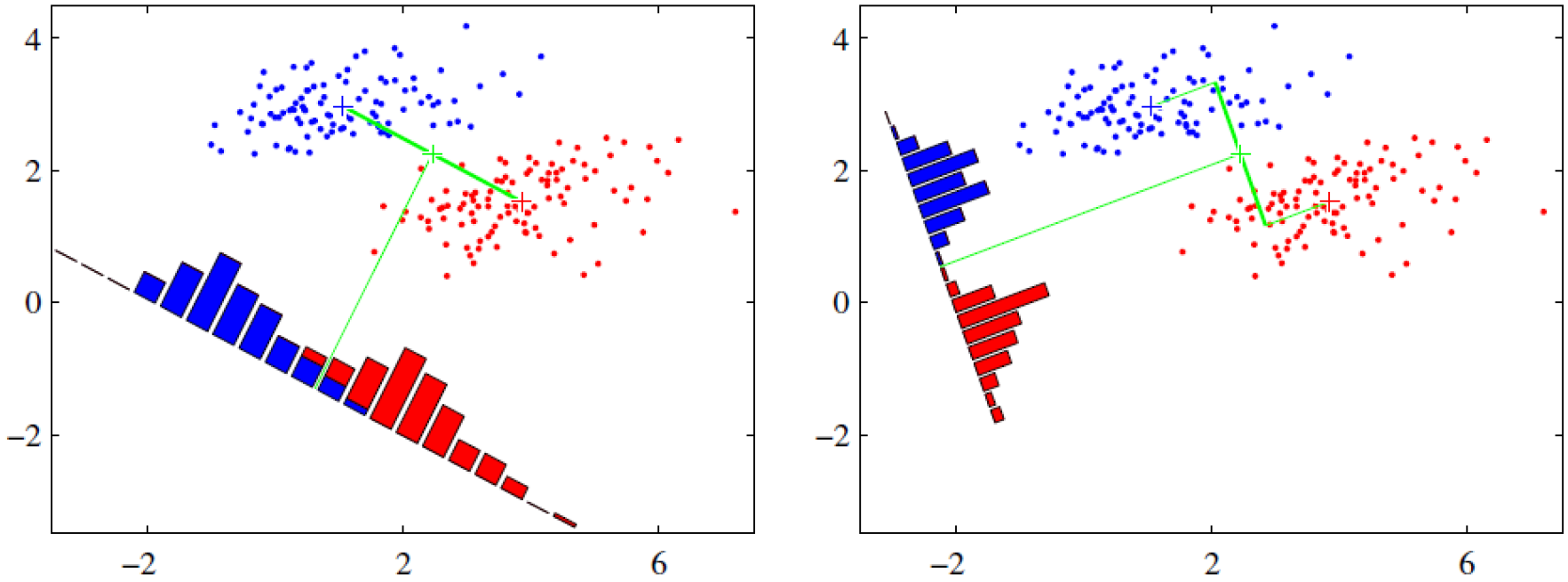
LDA的二维示意图

如图所示，我们要将数据在低维度上进行投影，投影后希望同一种类别数据的投影点尽可能的接近，而不同类别的数据的类别中心之间的距离尽可能的大。

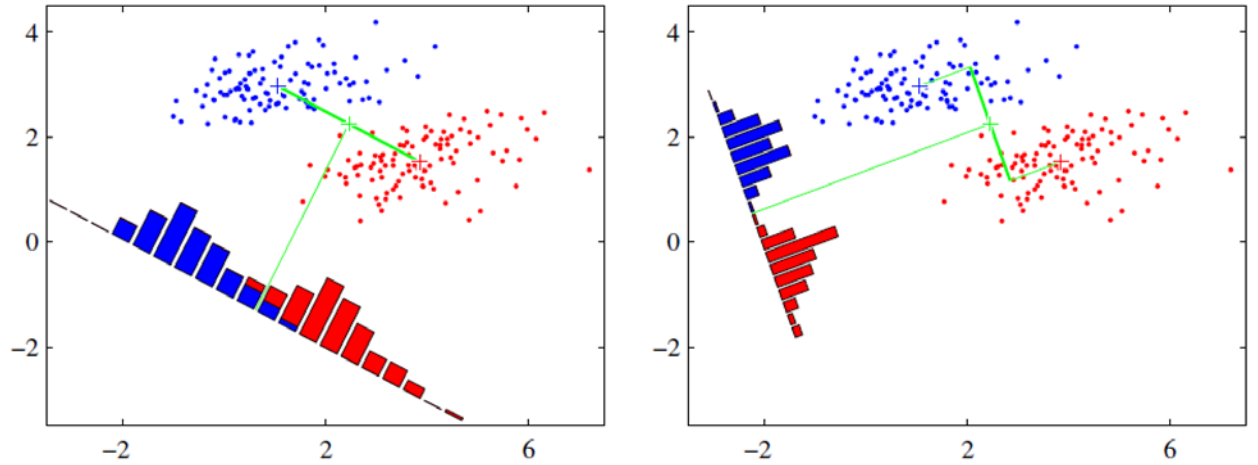
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

四、线性判别分析

- 假设我们有两类数据，分别为红色和蓝色，如下图所示，这些数据特征是二维的，我们希望将这些数据投影到一维的一条直线，让每一种类别数据的投影点尽可能的接近，而红色和蓝色数据中心之间的距离尽可能的大。



四、线性判别分析



两种投影方式，哪一种能更好的满足我们的标准呢？

从直观上可以看出，右图要比左图的投影效果好，因为

- 右图的黑色数据和蓝色数据各个较为集中，且类别之间的距离明显。
- 左图在边界处数据混杂。

四、线性判别分析

- LDA vs.PCA：

相同点：

- 两者均可以对数据进行降维。
- 两者在降维时均使用了矩阵特征分解的思想。
- 两者都假设数据符合高斯分布。

不同点：

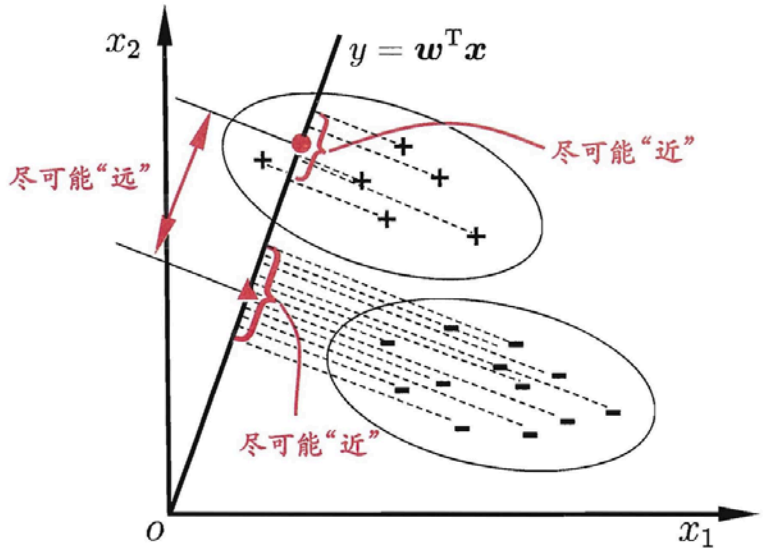
- LDA是有监督的降维方法，而PCA是无监督的降维方法。
- LDA降维最多降到类别数k-1的维数，而PCA没有这个限制。
- LDA除了可以用于降维，还可以用于分类。
- LDA选择分类性能最好的投影方向，而PCA选择样本点投影具有最大方差的方向。

四、线性判别分析

2. LDA解决二分类任务

参数定义：

- 数据集$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}, y_i \in \{0,1\}$
- X_i表示第$i \in \{0,1\}$类示例的集合
- $\boldsymbol{\mu}_i$表示第$i \in \{0,1\}$类示例的均值向量
- Σ_i表示第$i \in \{0,1\}$类示例的协方差矩阵



实现方式：将数据投影到直线$y = \boldsymbol{\omega}^T\boldsymbol{x} + b$上，可以得到4个值

- 两类样本的中心在直线上的投影分别为$\boldsymbol{\omega}^T\boldsymbol{\mu}_0$，$\boldsymbol{\omega}^T\boldsymbol{\mu}_1$
- Σ_i两类样本的协方差分别为$\boldsymbol{\omega}^T\Sigma_0\boldsymbol{\omega}$，$\boldsymbol{\omega}^T\Sigma_1\boldsymbol{\omega}$

四、线性判别分析

目标：

- 使同类样例的投影点尽可能接近
- 使异类样例的投影点尽可能远离

方法：

- 可以让同类样例投影点的协方差尽可能小，即$\boldsymbol{\omega}^T\Sigma_0\boldsymbol{\omega}+\boldsymbol{\omega}^T\Sigma_1\boldsymbol{\omega}$尽可能小；
- 可以让类中心之间的距离尽可能大，即$\|\boldsymbol{\omega}^T\boldsymbol{\mu}_0-\boldsymbol{\omega}^T\boldsymbol{\mu}_1\|_2^2$尽可能大.

同时考虑两者，得到最大化目标：

$$J=\frac{\|\boldsymbol{\omega}^T\boldsymbol{\mu}_0-\boldsymbol{\omega}^T\boldsymbol{\mu}_1\|_2^2}{\boldsymbol{\omega}^T\Sigma_0\boldsymbol{\omega}+\boldsymbol{\omega}^T\Sigma_1\boldsymbol{\omega}}=\frac{\boldsymbol{\omega}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\omega}}{\boldsymbol{\omega}^T(\Sigma_0+\Sigma_1)\boldsymbol{\omega}}$$

四、线性判别分析

定义两个矩阵：

- 类内散度矩阵$\boldsymbol{S}_\omega$，$\boldsymbol{S}_\omega = \Sigma_0 + \Sigma_1 = \sum_{\boldsymbol{x}\in X_0} (\boldsymbol{x}-\boldsymbol{\mu}_0)(\boldsymbol{x}-\boldsymbol{\mu}_0)^T + \sum_{\boldsymbol{x}\in X_1} (\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^T$
- 类间散度矩阵$\boldsymbol{S}_b$，$\boldsymbol{S}_b = (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T$

最大化目标重写为：

$$J = \frac{\boldsymbol{\omega}^T \boldsymbol{S}_b \boldsymbol{\omega}}{\boldsymbol{\omega}^T \boldsymbol{S}_\omega \boldsymbol{\omega}}$$

此公式就是LDA最大化的目标，即$\boldsymbol{S}_b$与$\boldsymbol{S}_\omega$的“广义瑞利商”

四、线性判别分析

- 瑞利商：是指这样的函数$R(A,x)$ $$R(A,x)=\frac{x^H Ax}{x^H x}$$

- 其中x为非零向量，而A为 $n\times n$ 的 Hermitan矩阵。所谓的 Hermitan矩阵就是满足共轭转置矩阵和自己相等的矩阵，即$A^H=A$。如果我们的矩阵A是实矩阵，则满足$A^T=A$的矩阵即为 Hermitan矩阵。

- 广义瑞利商：是指这样的函数$R(A,B,x)$ $$R(A,B,x)=\frac{x^H Ax}{x^H Bx}$$

- 其中x为非零向量，而A,B为 $n\times n$ 的 Hermitan矩阵。B为正定矩阵。

- 广义瑞利商可通过标准化转化为瑞利商

四、线性判别分析

最大化目标：$J = \frac{\boldsymbol{\omega}^T \boldsymbol{S}_b \boldsymbol{\omega}}{\boldsymbol{\omega}^T \boldsymbol{S}_\omega \boldsymbol{\omega}}$

- 如何求解ω？
- 分析：注意到最大化目标中的分子和分母都是关于$\boldsymbol{\omega}$的二次项，因此最大化目标的解与$\boldsymbol{\omega}$的长度无关，只与其方向有关。不失一般性，令$\boldsymbol{\omega}^T \boldsymbol{S}_\omega \boldsymbol{\omega} = 1$
- 所以最大化目标等价于：

$$\min_{\boldsymbol{\omega}} \quad -\boldsymbol{\omega}^T \boldsymbol{S}_b \boldsymbol{\omega}$$
$$s.t. \quad \boldsymbol{\omega}^T \boldsymbol{S}_\omega \boldsymbol{\omega} = 1$$

四、线性判别分析

$$\begin{aligned} \min_{\boldsymbol{\omega}} \quad & -\boldsymbol{\omega}^T \boldsymbol{S}_b \boldsymbol{\omega} \\ s.t. \quad & \boldsymbol{\omega}^T \boldsymbol{S}_\omega \boldsymbol{\omega} = 1 \end{aligned}$$

由拉格朗日乘子法，上式等价于

$$\boldsymbol{S}_b \boldsymbol{\omega} = \boldsymbol{\lambda} \boldsymbol{S}_\omega \boldsymbol{\omega} \qquad (*)$$

其中λ是拉格朗日乘子。注意到$\boldsymbol{S}_b \boldsymbol{\omega}$的方向恒为$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$, 不妨令

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

代入到（*）中，得

$$\boldsymbol{\omega} = \boldsymbol{S}_\omega^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

四、线性判别分析

3. LDA解决多分类任务

参数定义：

- 数据集$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}, y_i \in \{C_1, C_2, \ldots, C_k\}$
- X_i表示第$i \in \{C_1, C_2, \ldots, C_k\}$类样本的集合
- $\boldsymbol{\mu}_i$表示第$i \in \{C_1, C_2, \ldots, C_k\}$类样本的均值向量
- Σ_i表示第$i \in \{C_1, C_2, \ldots, C_k\}$类样本的协方差矩阵

定义：

- 全局散度矩阵$\boldsymbol{S}_t$，$\boldsymbol{S}_t = \boldsymbol{S}_b + \boldsymbol{S}_\omega = \sum_{i=1}^{m}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$
- 类内散度矩阵$\boldsymbol{S}_\omega$，$\boldsymbol{S}_\omega = \sum_{i=1}^{N}\boldsymbol{S}_{\omega i}$
- 类间散度矩阵$\boldsymbol{S}_b$，$\boldsymbol{S}_b = \boldsymbol{S}_t - \boldsymbol{S}_\omega = \sum_{i=1}^{N} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$

其中μ是所有示例的均值向量，$\boldsymbol{S}_{\omega i} = \sum_{\boldsymbol{x} \in X_i}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$

四、线性判别分析

- 多分类LDA有多种实现方法，常见的一种实现方法是采用如下的优化目标：

$$\max_{W} \frac{tr(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W})}{tr(\boldsymbol{W}^T \boldsymbol{S}_\omega \boldsymbol{W})}$$

- 其中$\boldsymbol{W} = \mathbb{R}^{d \times (N-1)}$, $tr(\cdot)$表示矩阵的迹（trace），上述优化目标可通过如下广义特征值问题求解：

$$\boldsymbol{S}_b \boldsymbol{W} = \boldsymbol{\lambda} \boldsymbol{S}_\omega \boldsymbol{W}$$

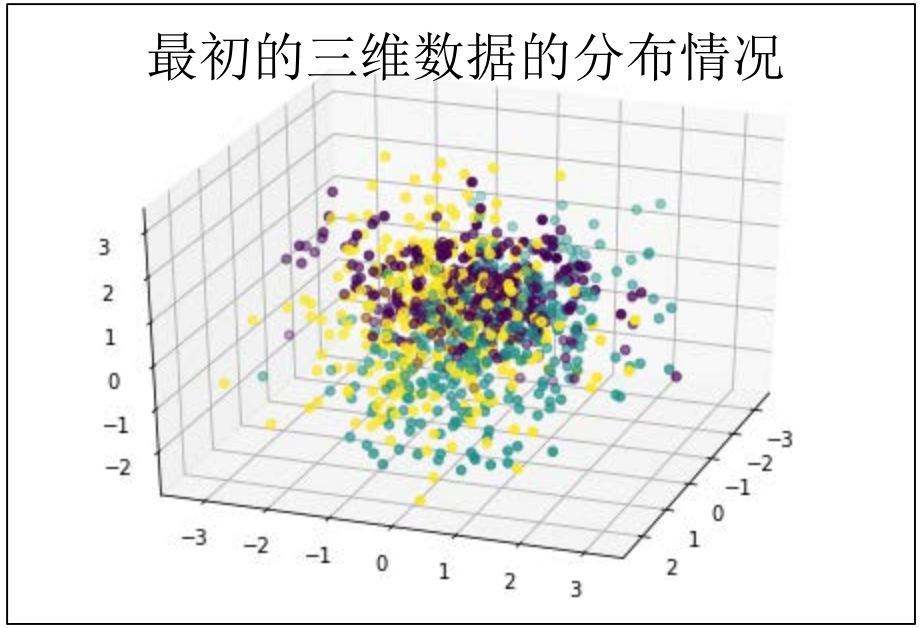
- $\boldsymbol{W}$的闭式解则是$\boldsymbol{S}_\omega^{-1} \boldsymbol{S}_b$的$N-1$个最大广义特征值所对应的特征向量组成的矩阵。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

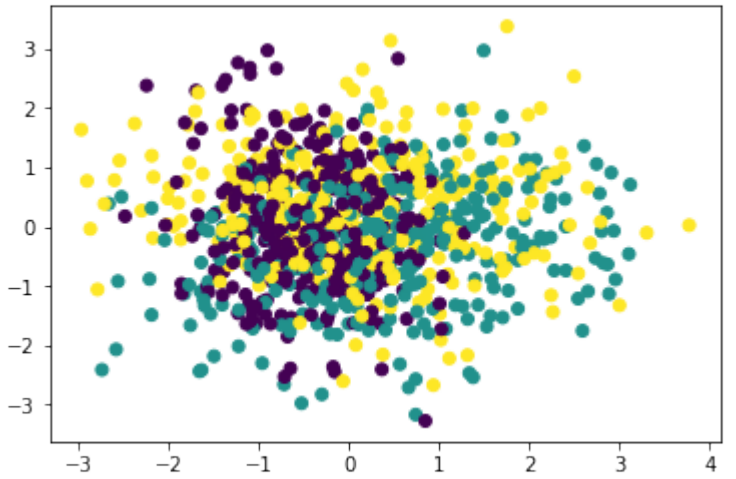
四、线性判别分析

4. 实例：降维

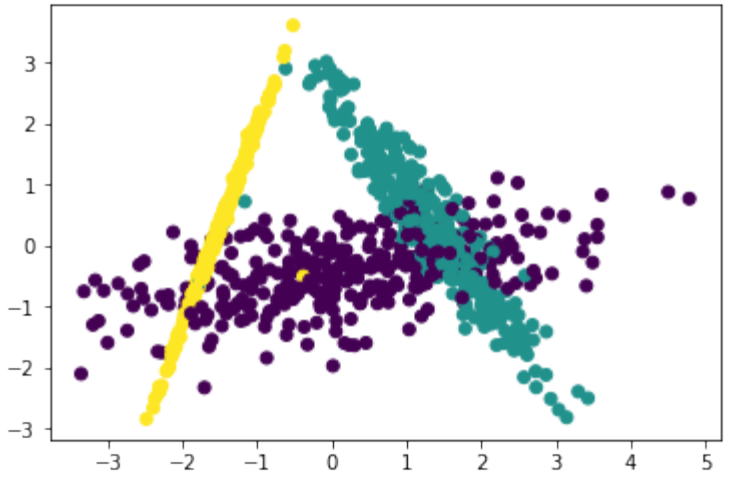
LDA vs.PCA：



降维



PCA



LDA

目录

五、多分类学习

问题：现实中经常遇到多分类学习，该如何将解决？

答：

- 方法一：用二分类学习方法可推广到多分类,如上一节的LDA的推广。
- 方法二：基于一些基本策略,利用二分类学习器来解决多分类问题.
- 在多数情形下，常用方法二来解决多分类学习任务。多分类学习的基本思路是“拆解法”，即将多分类任务拆为若干个二分类任务求解。

具体来说：

- 先对问题进行拆分，为拆出的每个二分类任务训练一个分类器；
- 在测试时，对这些分类器的预测结果进行集成以获得最终的多分类结果.

五、多分类问题

- 拆分策略：

要获得最终的多分类结果关键是如何对多分类任务进行拆分，以及如何对多个分类器进行集成.

最经典的拆分策略有三种：

- “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).

五、多分类问题

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

- 一对一拆分策略OvO：

OvO将N个类别两两配对，从而产生$N(N-1)/2$个二分类任务，例如OvO将为区分类别C_i咒C_j训练一个分类器，该分类器把D中的C_i类样例作为正例，C_j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，于是我们将得到$N(N-1)/2$个分类结果，最终结果可以通过投票产生。

五、多分类问题

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

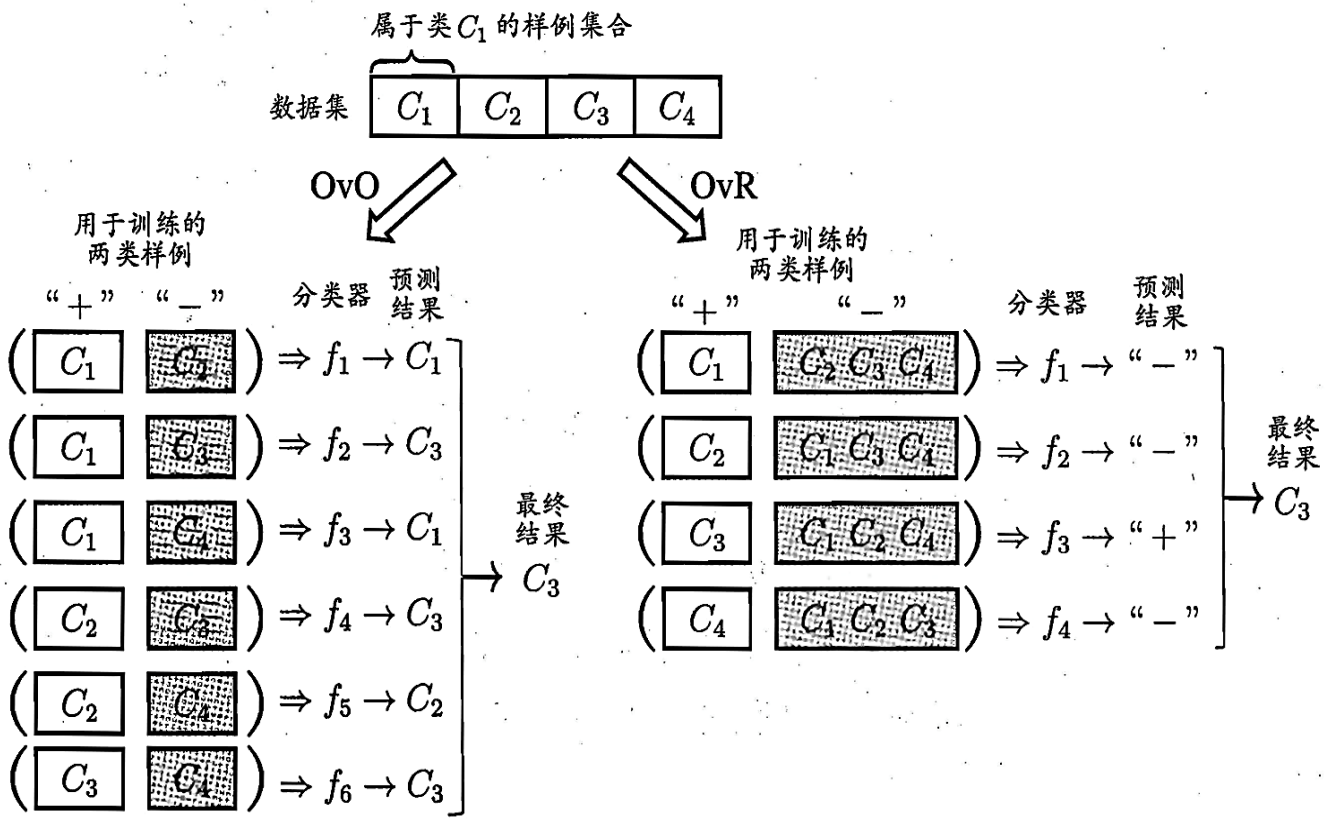
- 一对其余拆分策略OvR：

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

① 若仅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;

② 若有多个分类器预测为正类, 选择置信度最大的类别标记作为分类结果.

五、多分类问题



OvO与OvR示意图

OvO与OvR比较：

- OvR只需训练N个分类器，而OvO需训练$N(N-1)/2$个分类器，因此，OvO的存储开销和测试时间开销通常比OvR更大
- 但在训练时，OvR的每个分类器均使用全部训练样例，而OvO的每个分类器仅用到两个类的样例，因此，在类别很多时，OvO的训练时间开销通常比OvR更小

五、多分类问题

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

- 多对多拆分策略MvM：

MvM每次将若干个类作为正类，若干个其它类作为反类。显然OvO和OvR是MvM的特例。MvM的正、反类构造必须有特殊的设计。这里介绍一种最常用的MvM技术：纠错输出码（ECOC）

五、多分类问题

- 纠错输出码

ECOC是将编码的思想引入类别拆分，并尽可能在解码的过程中具有容错性。

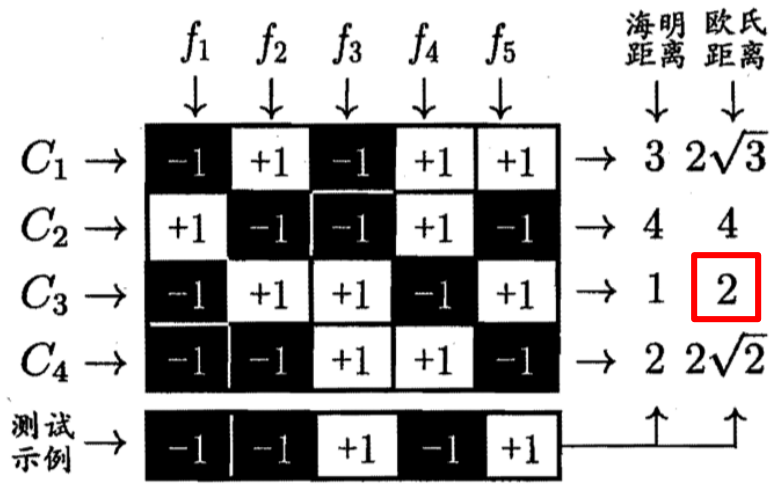
ECOC工作过程主要分为两步：

- 编码：对N个类别做M次划分，每次划分将一部分类别划为正类，一部分划为反类，从而形成一个二分类训练集，这样一共产生M个训练集，可训练出M个训练器。
- 解码：M个分类器分别对测试样本进行预测，这些预测标记组成一个编码。将这个预测编码与每个类别各自的编码进行比较，返回其中距离最小的类别最为最终预测结果。

五、多分类问题

ECOC是通过“编码矩阵”来划分类别。编码矩阵有多种形式，常见的有二元码和三元码。

- 二元码：将每个类别分别指定为正类和反类
- 三元码：在正、反类之外，还可以指定“停用类”



(a) 二元 ECOC 码

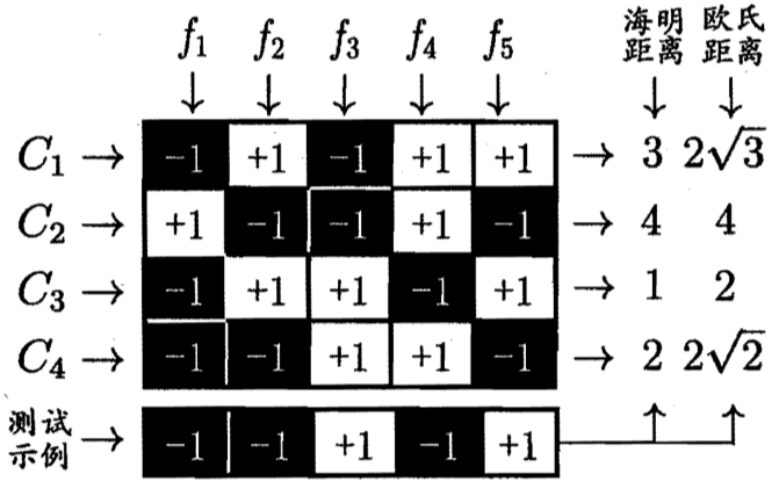
左图所示，分类器f_2将C_1类和C_3类的样例作为正例，C_2类和C_4类的样例作为反例；

在解码阶段，各分类器的预测结果联合起来形成了测试示例的编码，该编码与各类所对应的编码进行比较，将距离最小的编码所对应的类别作为预测结果。左图预测结果为C_3

五、多分类问题

思考：为什么叫“纠错输出码”？

答：因为在测试阶段，ECOC编码对分类器的错误有一定的容忍和修正能力。例如右图中对测试示例的正确预测编码是$(-1,+1,\ +1,-1,\ +1)$，假设在预测时某个分类器出错了，例如f_2出错从而导致了错误编码$(-1,-1,\ +1,-1,\ +1)$，但基于这个编码仍能产生正确的最终分类结果C_3.



(a) 二元 ECOC 码

- 一般来说，对同一个学习任务，ECOC编码越长，纠错能力越强。然而，编码越长，意味着所需训练的分类器越多，计算、存储开销都会增大；
- 另一方面，对有限类别数，可能的组合数目是有限的，码长超过一定范围后就失去了意义。

目录

六、类别不平衡问题

- 概念

- 类别不平衡(class-imbalance)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.

- 例如：在通过拆分法解决多分类问题时，即使原始问题中不同类别的训练样例数目相当,在使用OvR、MvM策略后产生的二分类任务仍可能出现类别不平衡现象

Next，从线性分类器角度来介绍处理类别不平衡的基本方法

六、类别不平衡问题

- 处理方法

新样本$\boldsymbol{x}$

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + b \longrightarrow \text{预测值}y\text{与某一阈值比较} \longrightarrow \begin{cases} \text{步侑}, & y \geq \text{阎偷} \\ \text{叟侑}, & y < \text{阎偷} \end{cases}$$

- y实际上表达了正例的可能性，几率$\frac{y}{1-y}$则反映了正例可能性与反例可能性之比值。通常情况下，阈值设置为0.5恰表明分类器认为真实正、反例可能性相同，即分类器决策规则为

$$\text{若}\frac{y}{1-y} > 1\text{，则预测为正例}$$

六、类别不平衡问题

- 当训练集中正、反例的数目不同时，令m^+表示正例数目，m^-表示反例数目，则观测几率是$\frac{m^+}{m^-}$，由于我们通常假设训练集是真实样本总体的无偏采样，因此观测几率就代表了真实几率。
- 于是，只要分类器的预测几率高于观测几率就应判定为正例，即

$$若\frac{y}{1-y} > \frac{m^+}{m^-}，则预测为正例$$

无偏采样意味着真实样本总体的类别比例在训练集中得以保持.

六、类别不平衡问题

然而，当训练集中正、反例的数目不同时，分类器仍按照下式进行决策。

$$若\frac{y}{1-y} > 1，则预测为正例$$

因此，对预测值进行微调，令

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

这就是类别不平衡学习的一个基本策略——“ 再缩放”.

六、类别不平衡问题

- 再缩放中存在问题及解决方法
  - 再缩放是在“训练集是真实样本总体的无偏采样”假设下产生的，然而这个假设往往并不成立。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几率.
- 针对此问题，产生了三种做法：

① 直接对训练集里的反类样例进行“欠采样”， 即去除一些反例使得正、反例数目接近，然后再进行学习；

② 对训练集里的正类样例进行“过采样”， 即增加一些正例使得正、 反例数目接近，然后再进行学习；

③ 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，将式 $\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$ 嵌入到其决策过程中，称为“阈值移动”

谢谢！

李爽

E-mail: shuangli@bit.edu.cn

Homepage: shuangli.xyz

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY