监督学习

主讲：王亚星、刘夏雷、郭春乐

南开大学计算机学院

致谢：本课件主要内容来自浙江大学吴飞教授、

南开大学程明明教授

提纲

机器学习：从数据中学习知识

1. 原始数据中提取特征

2. 学习映射函数 f

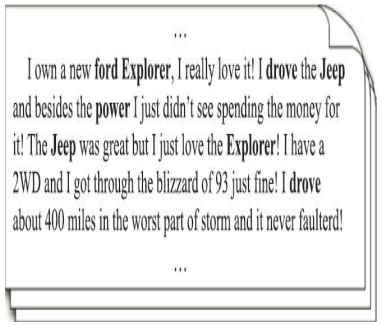
3. 通过映射函数 f 将原始数据映射到语义空间，即寻找数据和任务目标之间的关系

f $\begin{Bmatrix} 81 & 116 & \cdots & 133 \\ 104 & 130 & \cdots & 159 \\ \vdots & \vdots & \ddots & \vdots \\ 155 & 189 & \cdots & 218 \\ 197 & 221 & \cdots & 216 \end{Bmatrix}$

- Person
- Dog
- …

图像数据

类别分类



f {car, money, drive,…}

- 喜悦
- 愤怒
- …

文本数据

情感分类

机器学习的分类

监督学习(supervised learning)
数据有标签、一般为回归或分类等任务

无监督学习(un-supervised learning)
数据无标签、一般为聚类或若干降维任务

强化学习(reinforcement learning)
序列数据决策学习，一般为与从环境交互中学习

半监督学习
(semi-supervised learning)

监督学习：重要元素

标注数据

学习模型

损失函数

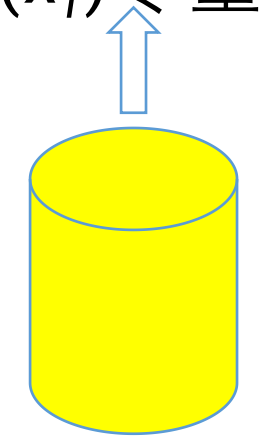
- 标识了类别信息的数据
  学什么
- 如何学习得到映射模型
  如何学
- 如何对学习结果进行度量
  学到否

监督学习：损失函数

- 训练集共有n个标注数据，第i个记为(x_i, y_i)
- 从训练数据中学习映射函数$f(x_i)$
  - 损失函数就是真值y_i与预测值$f(x_i)$之间差值的函数。
- 在训练过程中希望映射函数在训练数据集上得到“损失”最小
  - 即$\min \sum_{i=1}^{n} Loss(f(x_i), y_i)$。

训练映射函数f

使得$f(x_i)$尽量等于y_i



训练数据集

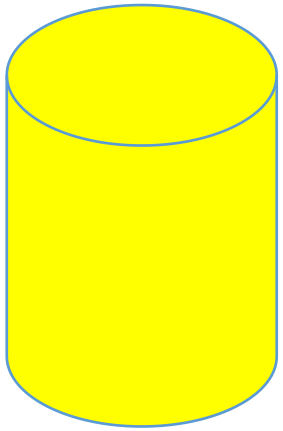
$(x_i, y_i), i = 1, \ldots, n$

监督学习：常见的损失函数

损失函数名称	损失函数定义
0—1损失函数	$Loss(y_i, f(x_i)) = \begin{cases} 1, & f(x_i) \neq y_i \\ 0, & f(x_i) = y_i \end{cases}$
平方损失函数	$Loss(y_i, f(x_i)) = (y_i - f(x_i))^2$
绝对损失函数	$Loss(y_i, f(x_i)) = \lvert y_i - f(x_i) \rvert$
对数损失函数/对数似然损失函数	$Loss(y_i, P(y_i \lvert x_i)) = -logP(y_i \lvert x_i)$

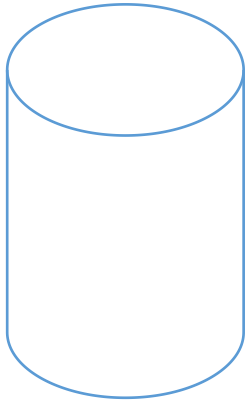
监督学习：训练数据和测试数据

从训练数据集学习
得到映射函数f



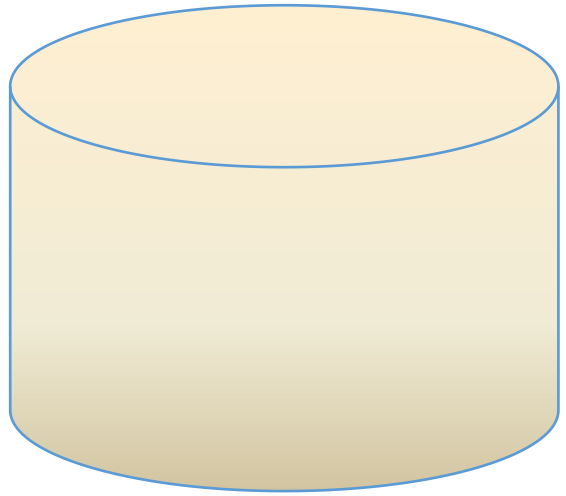
训练数据集
$(x_i, y_i), i = 1, \ldots, n$

在测试数据集
测试映射函数f



测试数据集
$(x_i', y_i'), i = 1, \ldots, m$

未知数据集
上测试映射函数f



监督学习：经验风险和期望风险

从训练数据集学习映射函数f

经验风险(empirical risk)

训练集中数据产生的损失。经验风险越小说明学习模型对训练数据拟合程度越好。

在测试数据集测试映射函数f

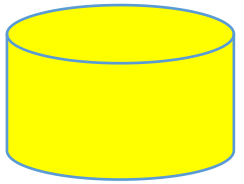
期望风险(expected risk)

当测试集中存在无穷多数据时产生的损失。期望风险越小，学习所得模型越好。

监督学习： 经验风险和期望风险

- **映射函数训练目标： 经验风险最小化**
  - Empirical risk minimization

$$\min_{f \in \Phi} \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, f(x_i))$$



训练数据集

$(x_i, y_i), i = 1, \ldots, n$

选取一个使得训练集所有数据损失平均值最小的映射函数。
这样的考虑是否够？

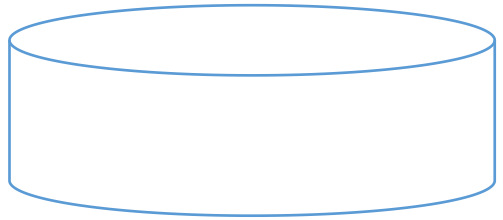
监督学习： 经验风险和期望风险

- **映射函数训练目标： 期望风险最小化**
  - Expected risk minimization

$$\min_{f \in \Phi} \int_{x,y} Loss(y, f(x)) P(x, y)\, dxdy$$

测试数据集数据无穷多

$(x_i', y_i'), i = 1, \ldots, \infty$



监督学习：经验风险和期望风险

- 期望风险是模型关于联合分布期望损失，经验风险是模型关于训练样本集平均损失。
- 根据大数定律，当样本容量趋于无穷时，经验风险趋于期望风险。所以在实践中很自然用经验风险来估计期望风险。
- 由于现实中训练样本数目有限，用经验风险估计期望风险并不理想，要对经验风险进行一定的约束。

监督学习：经验风险和期望风险

- 模型泛化能力与经验风险、期望风险的关系

训练集上表现	测试集上表现	
经验风险小	期望风险小	泛化能力强
经验风险小	期望风险大	过学习 （模型过于复杂）
经验风险大	期望风险大	欠学习
经验风险大	期望风险小	“神仙算法”或“黄粱美梦”

监督学习：结构风险最小化

- 结构风险最小化(structural risk minimization)
  - 为了防止过拟合，在经验风险上加上表示模型复杂度的正则化项(regularizer)或惩罚项(penalty term)：

$$\min_{f \in \Phi} \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, f(x_i)) + \lambda J(f)$$

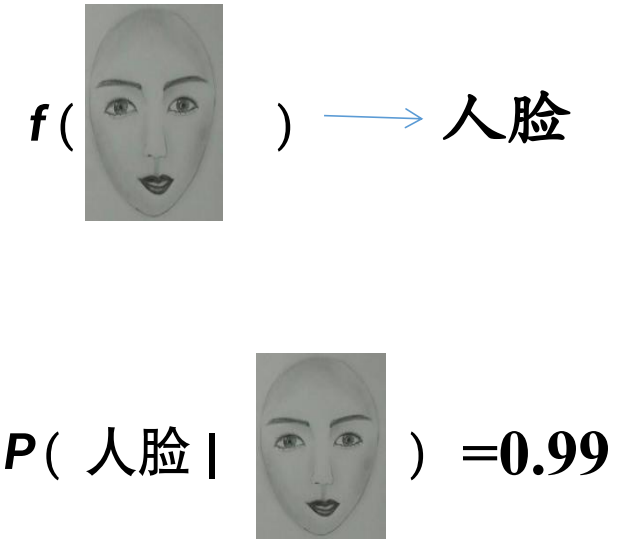
在最小化经验风险与降低模型复杂度之间寻找平衡

监督学习：判别模型与生成模型

- 监督学习方法又可以分为**生成**方法(generative approach)和**判别**方法(discriminative approach)。
- 所学到的模型分别称为生成模型(generative model)和判别模型(discriminative model).

监督学习：判别模型与生成模型

- 判别方法直接学习判别函数$f(X)$或者条件概率分布$P(Y|X)$作为预测的模型，即判别模型。
- 判别模型关心在给定输入数据下，预测该数据的输出是什么。
- 典型判别模型包括回归模型、神经网络、支持向量机和Ada boosting等。



监督学习：判别模型与生成模型

- 生成模型从数据中学习联合概率分布$P(X,Y)$（通过似然概率$P(X|Y)$和类概率$P(Y)$的乘积来求取）

$$P(Y|X)=\frac{P(X,Y)}{P(X)} \text{ 或者 } P(Y|X)=\frac{P(X|Y)\times P(Y)}{P(X)}$$

- 典型方法为贝叶斯方法、隐马尔可夫链
- 授之于鱼、不如授之于“渔”
- 联合分布概率$P(X,Y)$或似然概率$P(X|Y)$求取很困难

似然概率：计算导致样本X出现的模型参数值

$$P(Y|X)=\frac{P(X|Y)\times P(Y)}{P(X)}$$

监督学习：应用

- 分类
- 识别
- 推荐
- …

提纲

线性回归

- 在现实生活中，往往需要分析若干变量之间的关系，如碳排放量与气候变暖之间的关系、某一商品广告投入量与该商品销售量之间的关系等，这种分析不同变量之间存在关系的研究叫回归分析，刻画不同变量之间关系的模型被称为回归模型。如果这个模型是线性的，则称为线性回归模型。

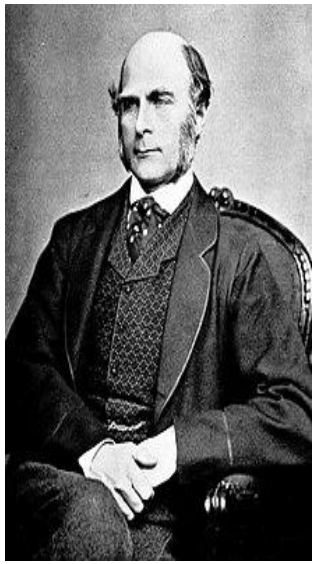
- 一旦确定了回归模型，就可以进行预测等分析工作，如从碳排放量预测气候变化程度、从广告投入量预测商品销售量等。

线性回归：一元线性回归

$$y = 33.73(英寸) + 0.516x$$

y：子女平均身高

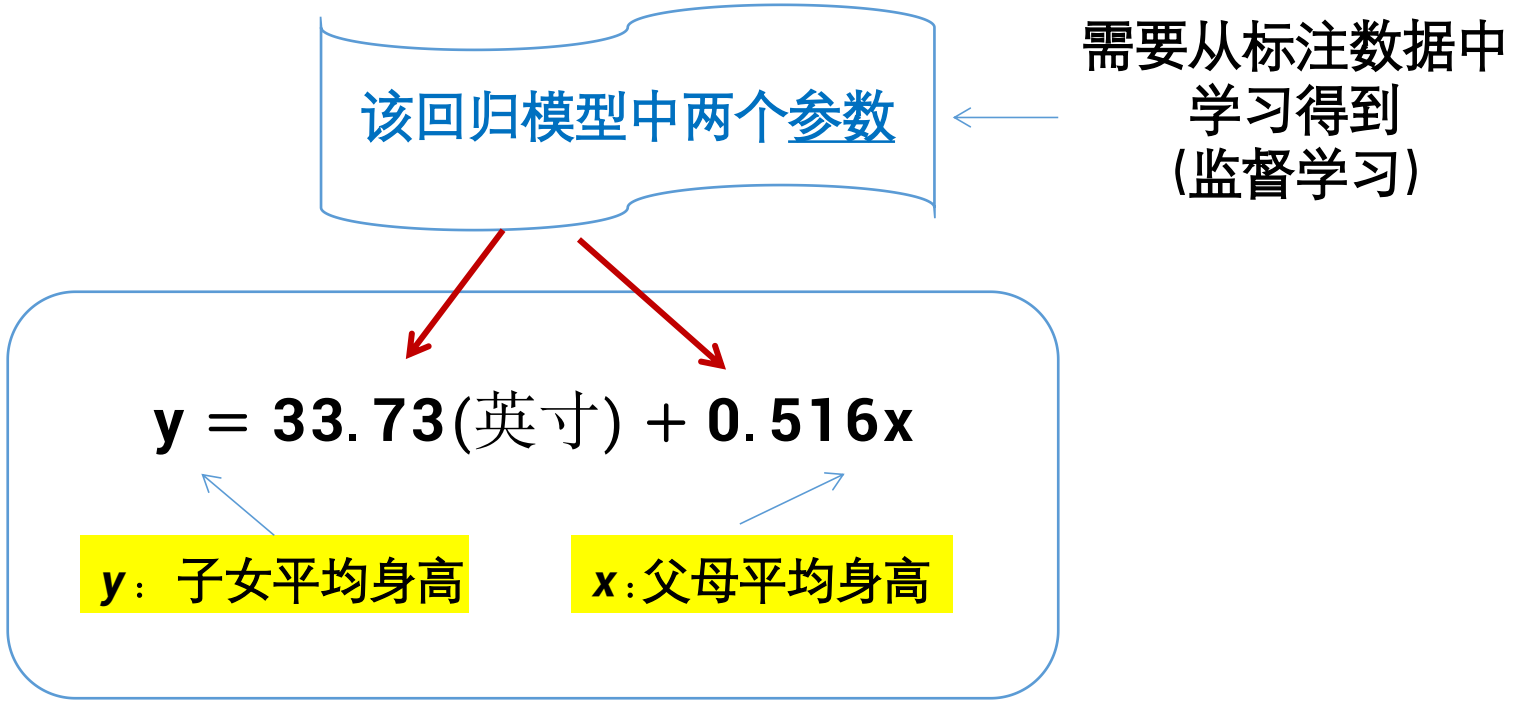
x：父母平均身高



英国著名生物学家兼统计学家高尔顿
Sir Francis Galton
(1822-1911)

- 父母平均身高每增加一个单位，其成年子女平均身高只增加0.516个单位，它反映了这种“衰退(regression)”效应（“回归”到正常人平均身高）。
- 虽然x和y之间并不总是具有“衰退”（回归）关系，但是“线性回归”这一名称就保留了下来了。

线性回归：一元线性回归



- 给出任意一对父母平均身高，则可根据上述方程，计算得到其子女平均身高
- 从父母平均身高来预测其子女平均身高
- 如何求取上述线性方程（预测方程）的参数？

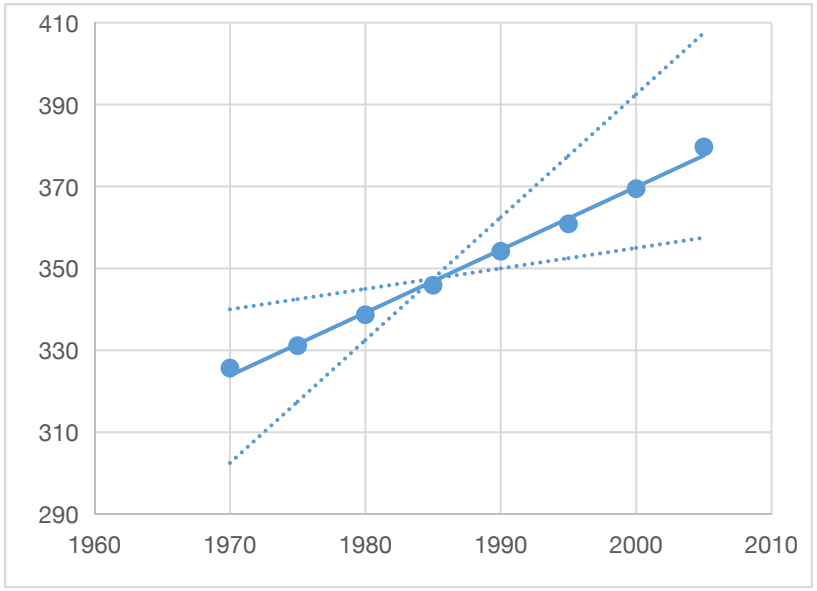
线性回归：一元线性回归

下表给出了莫纳罗亚山（夏威夷岛的活火山）从1970年到2005年每5年的二氧化碳浓度，单位是百万分比浓度 （Parts Per Million，ppm）。

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2 (y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

问题：1) 给出1984年二氧化碳浓度值；2) 预测2010年二氧化碳浓度值

线性回归：一元线性回归



莫纳罗亚山地区时间年份与二氧化碳浓度之间的一元线性回归模型（实线为最佳回归模型）

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2 (y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

⇩ 代入

回归模型：$y = ax + b$

求取：最佳回归模型是最小化残差平方和的均值，即要求8组(x, y)数据得到的残差平均值$\frac{1}{N}\sum(y - \tilde{y})^2$最小。残差平均值最小只与参数$a$和$b$有关，最优解即是使得残差最小所对应的$a$和$b$的值。

线性回归：一元线性回归

回归模型参数求取：$y_i = ax_i + b\ (1 \le i \le n)$

- 记在当前参数下第i个训练样本x_i的预测值为$\hat{y}_i$
- x_i的标注值（实际值）y_i与预测值$\hat{y}_i$之差记为$(y_i - \hat{y}_i)^2$
- 训练集中n个样本所产生误差总和为：$L(a,b) = \sum_{i=1}^{n}(y_i - a \times x_i - b)^2$

目标：寻找一组a和b，使得误差总和$L(a,b)$值最小。在线性回归中，解决如此目标的方法叫最小二乘法。

一般而言，要使函数具有最小值，可对$L(a,b)$参数a和b分别求导，令其导数值为零，再求取参数a和b的取值。

线性回归：一元线性回归

回归模型参数求取：$y_i = ax_i + b\ (1 \le i \le n)$

$$\min_{a,b} L(a,b) = \sum_{i=1}^{n} (y_i - a \times x_i - b)^2$$

$$\frac{\partial L(a,b)}{\partial b} = \sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 0$$

$$\rightarrow \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

$$\rightarrow \sum_{i=1}^{n} (y_i) - a\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0$$

$$\rightarrow n\bar{y} - an\bar{x} - nb = 0$$

$$b = \bar{y} - a\bar{x}$$

可以看出：只要给出了训练样本$(x_i, y_i)(i = 1, \ldots, n)$，我们就可以从训练样本出发，建立一个线性回归方程，使得对训练样本数据而言，该线性回归方程预测的结果与样本标注结果之间的差值和最小。

线性回归：一元线性回归

回归模型参数求取：$y_i = ax_i + b\ (1 \le i \le n)$

$$\min_{a,b} L(a,b) = \sum_{i=1}^{n} (y_i - a \times x_i - b)^2$$

$$\frac{\partial L(a,b)}{\partial a} = \sum_{i=1}^{n} 2(y_i - ax_i - b)(-x_i) = 0$$

将$b = \bar{y} - a\bar{x}$ $(\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n})$
代入上式

$$\rightarrow \sum_{i=1}^{n} (y_i - ax_i - \bar{y} + a\bar{x})(x_i) = 0$$

$$\rightarrow \sum_{i=1}^{n} (y_i x_i - ax_i x_i - \bar{y}x_i + a\bar{x}x_i) = 0$$

$$\rightarrow \sum_{i=1}^{n} (y_i x_i - \bar{y}x_i) - a\sum_{i=1}^{n} (x_i x_i - \bar{x}x_i) = 0$$

$$\rightarrow (\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}) - a(\sum_{i=1}^{n} x_i x_i - n\bar{x}^2) = 0$$

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i x_i - n\bar{x}^2}$$

线性回归：一元线性回归

回归模型参数求取：$y_i = ax_i + b\ (1 \le i \le n)$ $\min\limits_{a,b} \quad L(a,b) = \sum\limits_{i=1}^{n} (y_i - a \times x_i - b)^2$

$b = \bar{y} - a\bar{x}$

$a = \dfrac{\sum_{i=0}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i x_i - n\bar{x}^2}$

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2 (y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

训练样本数据

$$a = \frac{x_1 y_1 + x_2 y_2 + \ldots + x_8 y_8 - 8\bar{x}\bar{y}}{x_1^2 + x_2^2 + \ldots + x_8^2 - 8\bar{x}^2} = 1.5344$$

$$b = \bar{y} - a\bar{x} = -2698.9$$

预测莫纳罗亚山地区二氧化碳浓度的一元线性回归模型为“二氧化碳浓度 = 1.5344 × 时间年份 − 2698.9”，即$y = 1.5344x - 2698.9$。

线性回归：多元线性回归

多元线性回归模型例子

接下来扩展到数据特征的维度是多维的情况，在上述数据中增加一个影响火灾影响面积的潜在因素—风力。

气温x	5.1	8.2	11.5	13.9	15.1	16.2	19.6	23.3
风力z	4.5	5.8	4	6.3	4	7.2	6.3	8.5
火灾影响面积y	2.14	4.62	8.24	11.24	13.99	16.33	19.23	28.74

多维数据特征中线性回归的问题定义如下：假设总共有m个训练数据$\{(x_i, y_i)\}_{i=1}^{m}$，其中$x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,D}] \in \mathbb{R}^D$，$D$为数据特征的维度，线性回归就是要找到一组参数$a = [a_0, a_1, \ldots, a_D]$，使得线性函数：

$$f(\boldsymbol{x_i}) = a_0 + \sum_{j=1} a_j x_{i,j} = a_0 + \boldsymbol{a^T x_i}$$

线性回归：多元线性回归

最小化均方误差函数：

$$J_m = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - f(\boldsymbol{x_i})\right)^2$$

为了方便，使用矩阵来表示所有的训练数据和数据标签。

$$X = [x_1, \ldots, x_m], \quad \boldsymbol{y} = [y_1, \ldots, y_m]$$

其中每一个数据x_i会扩展一个维度，其值为1，对应参数a_0。均方误差函数可以表示为：

$$J_m(\boldsymbol{a}) = \left(\boldsymbol{y} - X^T\boldsymbol{a}\right)^T\left(\boldsymbol{y} - X^T\boldsymbol{a}\right)$$

均方误差函数$J_n(\boldsymbol{a})$对所有参数$\boldsymbol{a}$求导可得（勘误：P119）：

$$\nabla J(\boldsymbol{a}) = \left(-2X(y - X^T\boldsymbol{a})\right)^T$$

因为均方误差函数$J_n(\boldsymbol{a})$是一个二次的凸函数，所以函数只存在一个极小值点，

也同样是最小值点，所以令$\nabla J(\boldsymbol{a}) = 0$可得

$$XX^T\boldsymbol{a} = X\boldsymbol{y}$$

$$\boldsymbol{a} = (XX^T)^{-1}X\boldsymbol{y}$$

$$\boldsymbol{y} = \boldsymbol{A} \rightarrow \frac{\delta y}{\delta x} = \boldsymbol{0}$$

$$\boldsymbol{y} = \boldsymbol{Ax} \rightarrow \frac{\delta y}{\delta x} = \boldsymbol{A}$$

$$\boldsymbol{y} = \boldsymbol{xA} \rightarrow \frac{\delta y}{\delta x} = \boldsymbol{A^T}$$

$$\boldsymbol{y} = \boldsymbol{x^TAx} \rightarrow \frac{\delta y}{\delta x} = \boldsymbol{2x^TA}$$

MORE VIDEOS

线性回归：多元线性回归

对于上面的例子，转化为矩阵的表示形式为：

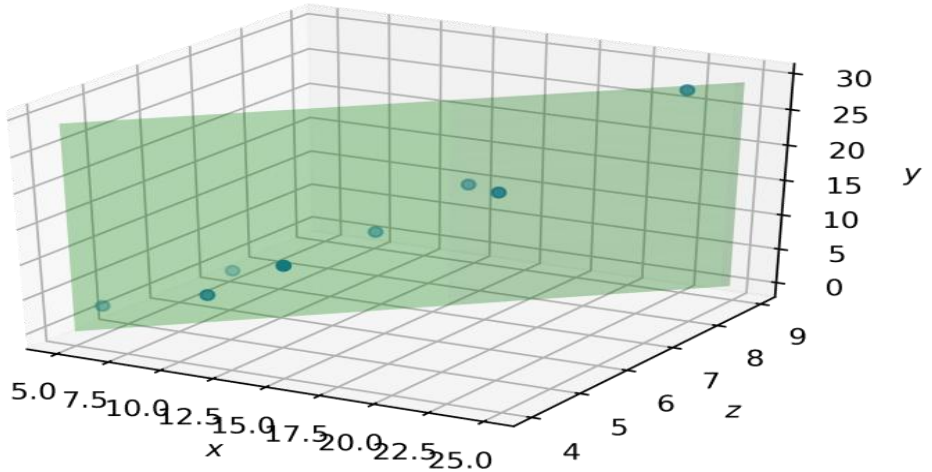
$$X = \begin{bmatrix} 5.1 & 8.2 & 11.5 & 13.9 & 15.1 & 16.2 & 19.6 & 23.3 \\ 4.5 & 5.8 & 4. & 6.3 & 4. & 7.2 & 6.3 & 8.5 \\ 1. & 1. & 1. & 1. & 1. & 1. & 1. & 1. \end{bmatrix}$$

$$\mathbf{y} = [2.14 \quad 4.62 \quad 8.24 \quad 11.24 \quad 13.99 \quad 16.33 \quad 19.23 \quad 28.74]^T$$

其中矩阵X多出一行全1，是因为常数项a_0，可以看作是数值为全1的特征的对应系数。计算可得

$$\boldsymbol{a} = [1.312 \quad 0.626 \quad -9.103]$$

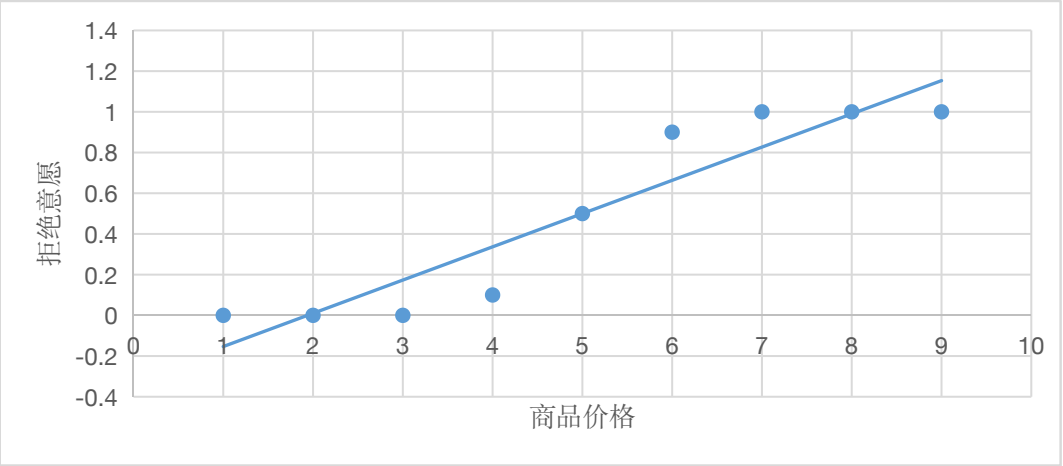
$$\boldsymbol{y} = -\mathbf{9.103} + \mathbf{1.312}\boldsymbol{x} + \mathbf{0.626}\boldsymbol{z}$$



线性回归：逻辑回归/对数几率回归

逻辑回归/对数几率回归模型例子

线性回归一个明显的问题是对离群点（和大多数数据点距离较远的数据点，outlier）非常敏感，导致模型建模不稳定，使结果有偏，为了缓解这个问题（特别是在二分类场景中）带来的影响，可考虑逻辑回归(logistic regression)[Cox 1958]。

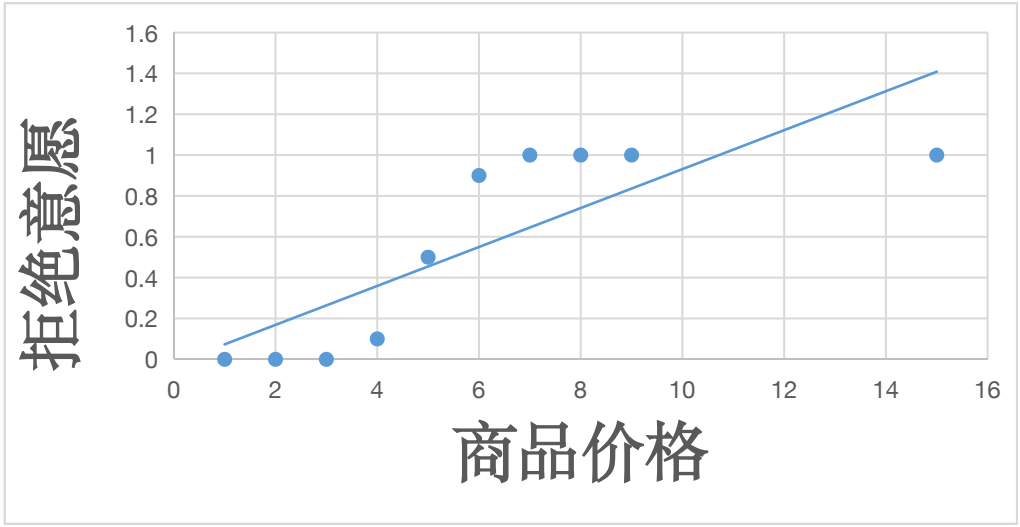


用户对某件商品拒绝购买意愿（选择不购买商品的人数/受调查的总人数）与商品价格之间的关系

线性回归：逻辑回归/对数几率回归

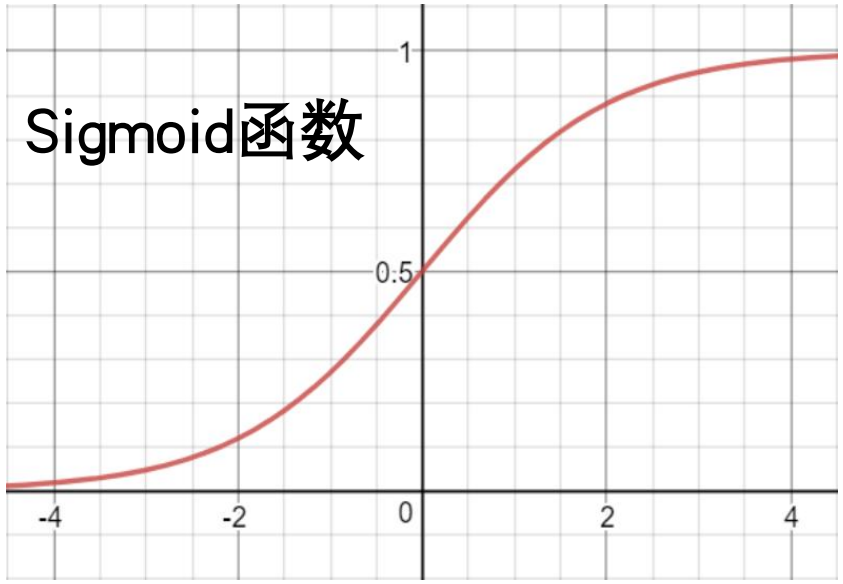
逻辑回归/对数几率回归模型例子

线性回归一个明显的问题是对离群点（和大多数数据点距离较远的数据点，outlier）非常敏感，导致模型建模不稳定，使结果有偏，为了缓解这个问题（特别是在二分类场景中）带来的影响，可考虑逻辑回归(logistic regression)[Cox 1958]。



加入一个离群点，该点表示当商品价格为15时，用户拒绝意愿为1（即用户不愿意购买该商品）

线性回归：逻辑回归/对数几率回归



逻辑回归(logistic regression)就是在回归模型中引入 sigmoid函数的一种非线性回归模型。Logistic回归模型可如下表示：

$$y = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+b)}} \quad , \text{其中} y \in (0,1),\ z = \boldsymbol{w}^T\boldsymbol{x} + \mathrm{b}$$

这里$\frac{1}{1+e^{-z}}$是sigmoid函数、$\boldsymbol{x} \in \mathbb{R}^d$是输入数据、$\boldsymbol{w} \in \mathbb{R}^d$和$b \in \mathbb{R}$是回归函数的参数。

线性回归：逻辑回归/对数几率回归

逻辑回归虽可用于对输入数据和输出结果之间复杂关系进行建模，但由于逻辑回归函数的输出具有概率意义，使得逻辑回归函数更多用于二分类问题（y = 1表示输入数据$\boldsymbol{x}$属于正例，y = 0表示输入数据$\boldsymbol{x}$属于负例）。

$y = \frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+b)}}$可用来计算输入数据$\boldsymbol{x}$属于正例概率，这里y理解为输入数据$\boldsymbol{x}$为正例的概率、1 − y理解为输入数据x为负例的概率，即$p(y=1|\boldsymbol{x})$。我们现在对比值$\frac{p}{1-p}$取对数(即$\log\left(\frac{p}{1-p}\right)$)来表示输入数据$\boldsymbol{x}$属于正例概率。$\frac{p}{1-p}$被称为几率(odds)，反映了输入数据$\boldsymbol{x}$作为正例的相对可能性。$\frac{p}{1-p}$的对数几率(log odds)或logit函数可表示为$\log\left(\frac{p}{1-p}\right)$。

显然，可以得到$p(y=1|\boldsymbol{x}) = \mathrm{h}_\theta(x) = \frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+b)}}$和$p(y=0|\boldsymbol{x}) = 1 - \mathrm{h}_\theta(x) = \frac{e^{-(\boldsymbol{w}^T\boldsymbol{x}+\mathrm{b})}}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+\mathrm{b})}}$。$\theta$表示模型参数（$\theta = \{\boldsymbol{w}, b\}$）。于是有：

$$\mathrm{logit}(p(y=1|\boldsymbol{x})) = \log\left(\frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})}\right) = \log\left(\frac{p}{1-p}\right) = \boldsymbol{w}^T\boldsymbol{x} + \mathrm{b}$$

线性回归：逻辑回归/对数几率回归

- 如果输入数据$\boldsymbol{x}$属于正例的概率大于其属于负例的概率，即$p(y=1|\boldsymbol{x})>0.5$，则输入数据$\boldsymbol{x}$可被判断属于正例。这一结果等价于$\frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})}>1$，即$\log\left(\frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})}\right)>\log 1=0$，也就是$\boldsymbol{w}^T\boldsymbol{x}+\mathrm{b}>0$成立。

- 从这里可以看出，logistic回归是一个线性模型。在预测时，可以计算线性函数$\boldsymbol{w}^T\boldsymbol{x}+\mathrm{b}$取值是否大于0来判断输入数据$\boldsymbol{x}$的类别归属。

线性回归：逻辑回归/对数几率回归

模型参数的似然函数被定义为$\mathbb{L}(\theta|\mathbb{D}) = \mathrm{p}(\mathbb{D}|\theta)$，其中$\mathbb{D} = \{(x_i, y_i\}(1 \leq \mathrm{i} \leq \mathrm{n})$表示所有观测数据（或训练数据），$\theta$表示模型参数（$\theta = \{\boldsymbol{w}, b\}$）。在最大化对数似然函数过程中，一般假设观测所得每一个样本数据是独立同分布（independent and identically distributed，i.i.d），于是可得：

$$\mathbb{L}(\theta|\mathbb{D}) = \mathrm{p}(\mathbb{D}|\theta) = \prod_{i=1}^{n} \mathrm{p}(y_i|\mathrm{x}, \theta) = \prod_{i=1}^{n} \left(\mathrm{h}_\theta(x_i)\right)^{y_i} \left(1 - \mathrm{h}_\theta(x_i)\right)^{1-y_i}$$

对上述公式取对数：

$$\mathrm{l}(\theta) = \log\left(L(\theta|\mathbb{D})\right) = \sum_{i=1}^{n} y_i \log\left(\mathrm{h}_\theta(x_i)\right) + (1 - y_i)\log(1 - \mathrm{h}_\theta(x_i))$$

线性回归：逻辑回归/对数几率回归

最大似然估计目的是计算似然函数的最大值，而分类过程是需要损失函数最小化。因此，在上式前加一个负号得到损失函数(交叉熵)：

$$\mathrm{T}(\theta) = -\mathrm{I}(\theta) = -\log\left(L(\theta|\mathbb{D})\right)$$

$$= -\left(\sum_{i=1}^{n} y_i \log\left(\mathrm{h}_\theta(x_i)\right) + (1-y_i)\log(1-\mathrm{h}_\theta(x_i))\right)$$

$\mathrm{T}(\theta)$等价于：$\mathrm{T}(\theta) = \begin{cases} -\log\left(\mathrm{h}_\theta(x_i)\right) & if\ y=1 \\ -\log\left(1-\mathrm{h}_\theta(x_i)\right) & if\ y=0 \end{cases}$

线性回归：逻辑回归/对数几率回归

需要最小化损失函数来求解参数。数损失函数对参数θ的偏导如下（其中，$h'_\theta(x) = h_\theta(x)(1-h_\theta(x)), \log' x = \frac{1}{x}$ ）

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= -\sum_{i=1}^{n} \left(y_i \frac{1}{h_\theta(x_i)} \frac{\partial h_\theta(x_i)}{\partial \theta_j} + (1-y_i) \frac{1}{1-h_\theta(x_i)} \frac{\partial (1-h_\theta(x_i))}{\partial \theta_j} \right) \\
&= -\sum_{i=1}^{n} \frac{\partial h_\theta(x_i)}{\partial \theta_j} \left(\frac{y_i}{h_\theta(x_i)} - \frac{1-y_i}{1-h_\theta(x_i)} \right) \\
&= -\sum_{i=1}^{n} x_i h_\theta(x_i)(1-h_\theta(x_i)) \left(\frac{y_i}{h_\theta(x_i)} - \frac{1-y_i}{1-h_\theta(x_i)} \right) \\
&= -\sum_{i=1}^{n} x_i (y_i(1-h_\theta(x_i)) - (1-y_i) h_\theta(x_i)) \\
&= \sum_{i=1}^{n} (h_\theta(x_i) - y_i) x_i
\end{aligned}
$$

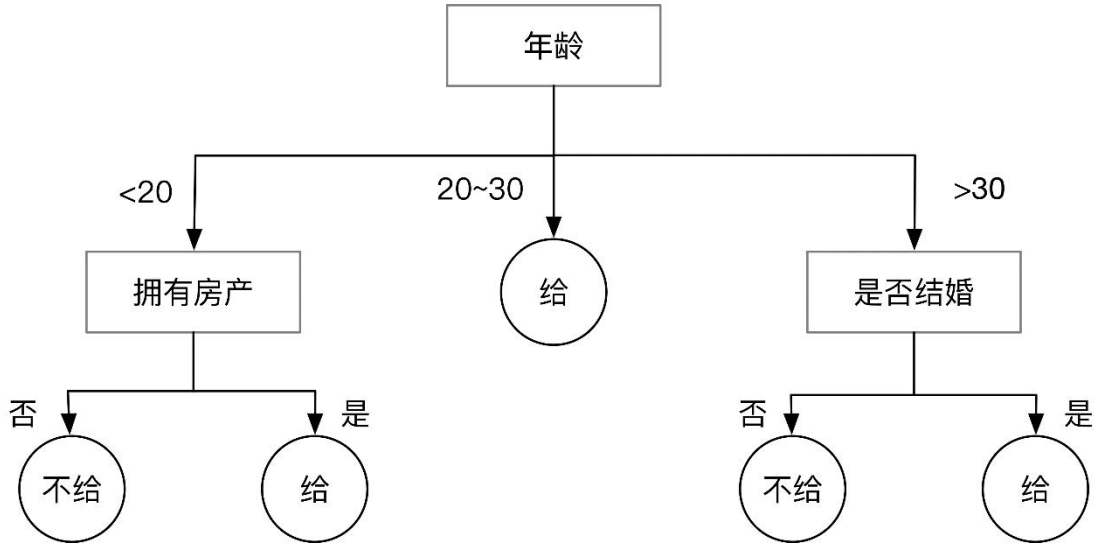
将求导结果代入梯度下降迭代公式得：（勘误：P124）

$$
\theta_j = \theta_j - \eta \sum_{i=1}^{n} (h_\theta(x_i) - y_i) x_i
$$

提纲

决策树

决策树是一种通过树形结构来进行分类的方法。在决策树中，树形结构中每个非叶子节点表示对分类目标在某个属性上的一个判断，每个分支代表基于该属性做出的一个判断，最后树形结构中每个叶子节点代表一种分类结果，所以决策树可以看作是一系列以叶子节点为输出的决策规则（Decision Rules）[Quinlan 1987]。



决策树：案例

决策树分类案例

序号	年龄	银行流水	是否结婚	拥有房产	是否给予贷款
1	>30	高	否	是	否
2	>30	高	否	否	否
3	20~30	高	否	是	是
4	<20	中	否	是	是
5	<20	低	否	是	是
6	<20	低	是	否	否
7	20~30	低	是	否	是
8	>30	中	否	是	否
9	>30	低	是	是	是
10	<20	中	否	是	是
11	>30	中	是	否	是
12	20~30	中	否	否	是
13	20~30	高	是	是	是
14	<20	中	否	否	否

决策树：信息熵

信息熵（entropy）

假设有K个信息，其组成了集合样本D，记第k个信息发生的概率为$p_k(1 \le k \le K)$”。如下定义这K个信息的信息熵：

$$E(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

$E(D)$值越小，表示D包含的信息越确定，也称D的纯度越高。需要指出，所有p_k累加起来的和为1。

要点：构建决策树时划分属性的顺序选择是重要的。性能好的决策树随着划分不断进行，决策树分支结点样本集的“纯度”会越来越高，即其所包含样本尽可能属于相同类别。

年龄属性划分后子样本集情况统计

年龄属性取值 a_i	“>30”	“20~30”	“<20”
对应样本数 $\lvert D_i \rvert$	5	4	5
正负样本数量	(2+，3−)	(4+，0−)	(3+，2−)

决策树：信息熵

年龄属性划分后子样本集情况统计

年龄属性取值	“>30”	“20~30”	“<20”
对应样本数	5	4	5
正负样本数量	(2+，3−)	(4+，0−)	(3+，2−)

$$\text{“年龄} > 30\text{”}: \quad Ent(D_0) = -\left(\frac{2}{5}\times\log_2\frac{2}{5}+\frac{3}{5}\times\log_2\frac{3}{5}\right) = 0.971$$

$$\text{“年龄20~30”}: \quad Ent(D_1) = -\left(\frac{4}{4}\times\log_2\frac{4}{4}+0\right) = 0$$

$$\text{“年龄} < 20\text{”}: \quad Ent(D_2) = -\left(\frac{3}{5}\times\log_2\frac{3}{5}+\frac{2}{5}\times\log_2\frac{2}{5}\right) = 0.971$$

决策树：信息增益

得到上述三个的信息熵后，可进一步计算使用年龄属性对原样本集进行划分后的信息增益，计算公式如下：

$$Gain(D, A) = Ent(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} Ent(D_i)$$

将 $A = \text{年龄}$ 代入。于是选择年龄这一属性划分后的信息增益为：

$$Gain(D, \text{年龄}) = 0.940 - \left(\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 \right) = 0.246$$

同理，可以计算银行流水、是否结婚、是否拥有房产三个人物属性的信息增益。通过比较四种属性信息增益的高低来选择最佳属性对原样本集进行划分，得到最大的“纯度”。如果划分后的不同子样本集都只存在同类样本，那么停止划分。

决策树：构建决策树

$info$和$Gain-ratio$计算公式如下：

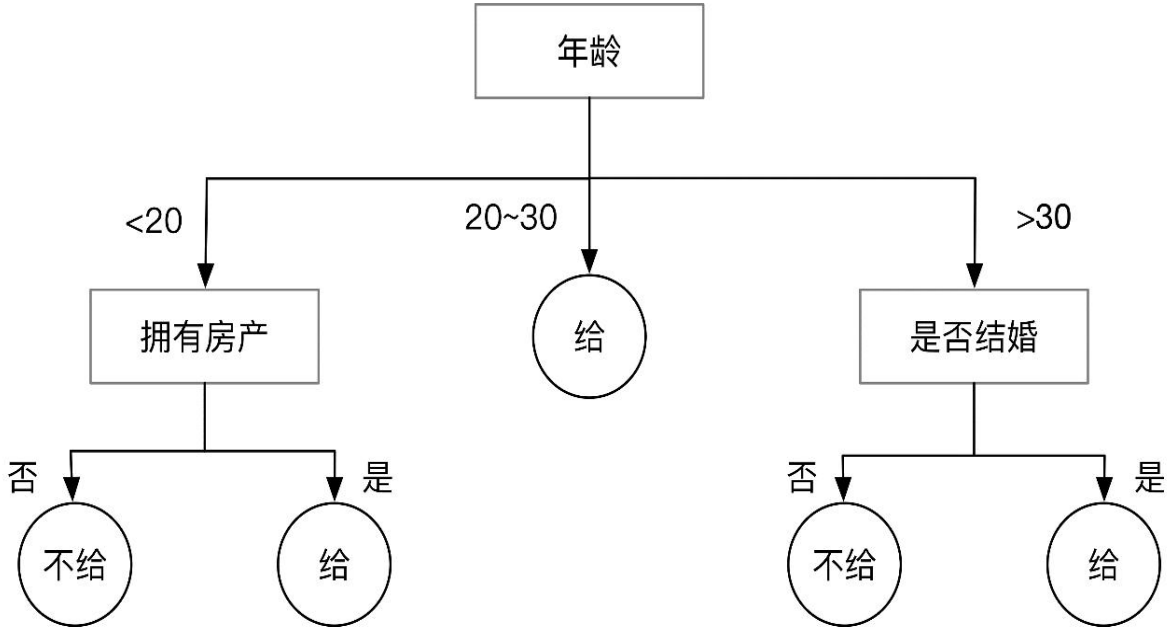
$$info = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

$$Gain-ratio = Gain(D,A)/info$$

另一种计算更简的度量指标是如下的Gini系数：

$$\text{Gini(D)} = 1 - \sum_{k=1}^{K} p_k^2$$

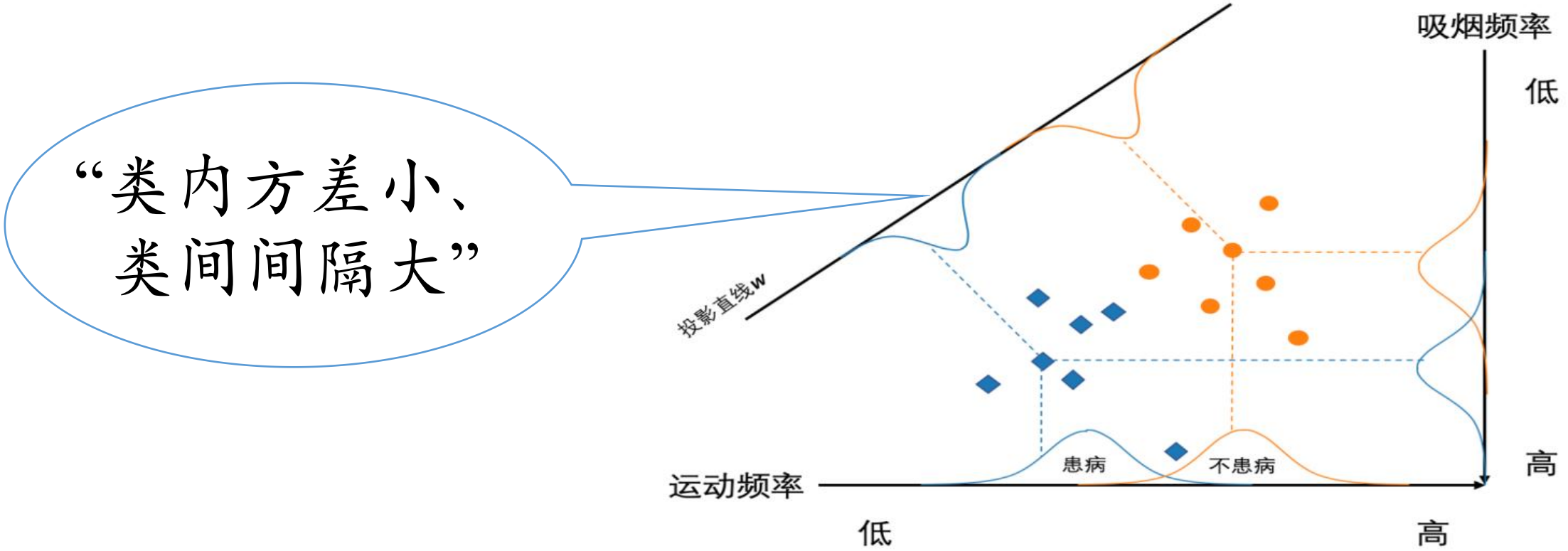
相对于信息熵的计算$E(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$，不用计算对数log，计算更为简易。



提纲

线性区别分析 (linear discriminant analysis，LDA)

- **一种基于监督学习的降维方法**
  - 也称为Fisher线性判别分析 (FDA) [Fisher 1936]
  - LDA利用类别信息，将高维数据样本线性投影到一个低维空间



线性区别分析：符号定义

- **假设样本集为**$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$，$\boldsymbol{x}_i \in \mathbb{R}^d$
  - 其中，y_i的取值范围是$\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_K\}$，即一共有$K$类样本
  - 定义$\boldsymbol{X}$为所有样本构成集合、$\boldsymbol{X}_i$为第i类样本的集合
  - N_i为第i个类别所包含样本个数
  - $\boldsymbol{m}$为所有样本的均值向量、$\boldsymbol{m}_i$为第i类样本的均值向量
- Σ_i**为第**i**类样本的协方差矩阵，定义为:**

$$\Sigma_i = \sum_{\boldsymbol{x} \in \boldsymbol{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

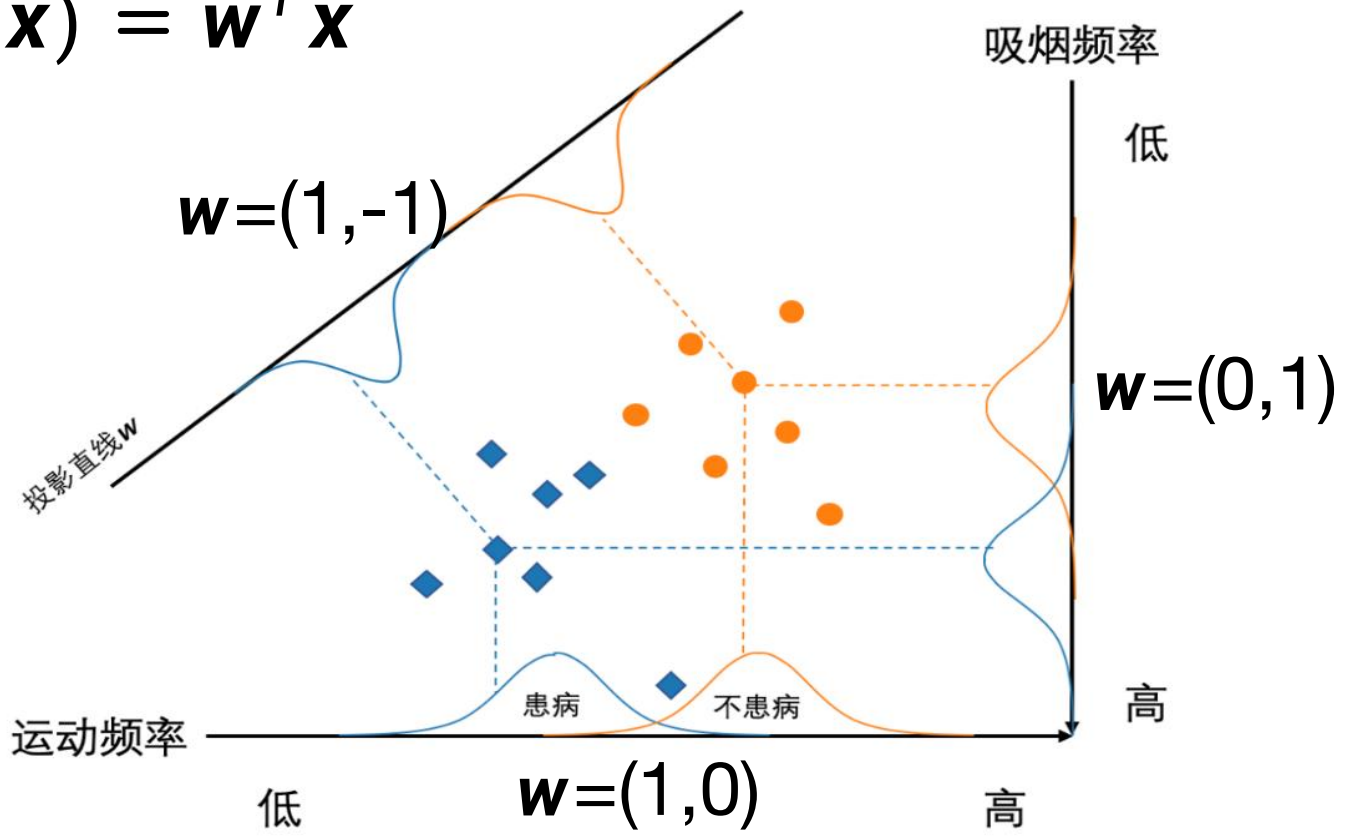
线性区别分析：二分类问题

- **先来看$K = 2$的情况：训练样本归属于$\mathbb{C}_1$或$\mathbb{C}_2$两个类别**
  - 过如下的线性函数投影到一维空间上（其中$\boldsymbol{w} \in \mathbb{R}^n$）

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$$

节点(1,1), (2,2), (3,3), (4,4)

都会投影到同一个点。



线性区别分析：二分类问题

- 先来看$K = 2$的情况：训练样本归属于$\mathbb{C}_1$或$\mathbb{C}_2$两个类别
  - 过如下的线性函数投影到一维空间上（其中$\boldsymbol{w} \in \mathbb{R}^n$）

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

- 投影之后类别$\mathbb{C}_1$的协方差矩阵s_1为：

$$s_1 = \sum_{\boldsymbol{x} \in \mathbb{C}_1} \left(\boldsymbol{w}^T \boldsymbol{x} - \boldsymbol{w}^T \boldsymbol{m}_1\right)^2 = \boldsymbol{w}^T \sum_{\boldsymbol{x} \in \mathbb{C}_1} \left[(\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^T\right] \boldsymbol{w}$$

- 同理可得到投影之后类别$\mathbb{C}_2$的协方差矩阵s_2

线性区别分析：二分类问题

- **投影后两个协方差矩阵为$s_1 = \boldsymbol{w}^T \Sigma_1 \boldsymbol{w}$和$s_2 = \boldsymbol{w}^T \Sigma_2 \boldsymbol{w}$**
  - 为了使同类本尽可能靠近(分散程度低)，需要最小化$s_1 + s_2$
- **投影后，归属于两个类别的数据样本中心为：**

$$m_1 = \boldsymbol{w}^T \boldsymbol{m}_1, \quad m_2 = \boldsymbol{w}^T \boldsymbol{m}_2$$

- **为使不同类样本尽可能彼此远离，需要最大化**

$$\|m_2 - m_1\|_2^2$$

- **总体需要最大化的目标$J(\boldsymbol{w})$定义为**

$$J(\boldsymbol{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1 + s_2}$$

线性区别分析：二分类问题

$$J(\boldsymbol{w}) = \frac{\|\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)\|_2^2}{\boldsymbol{w}^T\Sigma_1\boldsymbol{w} + \boldsymbol{w}^T\Sigma_2\boldsymbol{w}} = \frac{\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w}}{\boldsymbol{w}^T(\Sigma_1 + \Sigma_2)\boldsymbol{w}} = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

- **S_b称为类间散度矩阵(between-class scatter matrix)**
  - 衡量两个类别均值点之间的“分离”程度：

$$\boldsymbol{S}_b = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$$

- **S_w称为类内散度矩阵(within-class scatter matrix)**
  - 衡量每个类别中数据点的“分离”程度：

$$\boldsymbol{S}_w = \Sigma_1 + \Sigma_2$$

- **由于$J(\boldsymbol{w})$的分子和分母都是关于$\boldsymbol{w}$的二项式，因此解只与$\boldsymbol{w}$的方向有关，与$\boldsymbol{w}$的长度无关，因此可令分母$\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w} = 1$，用拉格朗日乘子法来求解**

线性区别分析：二分类问题

- 对应拉格朗日函数为：

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

- 对$\boldsymbol{w}$求偏导并使其求导结果为零，可得$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{w}$
  - λ和$\boldsymbol{w}$分别是$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的特征根和特征向量
  - $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{w}$也被称为Fisher线性判别

$$\boldsymbol{S}_w{}^{-1}\boldsymbol{S}_b\boldsymbol{w} = \boldsymbol{S}_w{}^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \times \lambda_w = \lambda\boldsymbol{w}$$

- 由于对$\boldsymbol{w}$的放大缩小不影响结果，可约去未知数λ和λ_w：

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

线性区别分析：多分类问题

假设n个原始高维数据所构成的类别种类为K、每个原始数据被投影映射到低维空间中的维度为r。

令投影矩阵$\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_r)$，可知$\boldsymbol{W}$是一个$n\times r$矩阵。于是，$\boldsymbol{W}^T\boldsymbol{m}_i$为第$i$类样本数据中心在低维空间的投影结果，$\boldsymbol{W}^T\Sigma_i\boldsymbol{W}$为第$i$类样本数据协方差在低维空间的投影结果。

类内散度矩阵$\boldsymbol{S}_w$重新定义如下：

$$\boldsymbol{S}_w=\sum_{i=1}^{K}\Sigma_i，\text{其中}\Sigma_i=\sum_{\boldsymbol{x}\in class\ i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T$$

在上式中，$\boldsymbol{m}_i$是第i个类别中所包含样本数据的均值。
类间散度矩阵$\boldsymbol{S}_b$重新定义如下：

$$\boldsymbol{S}_b=\sum_{i=1}^{K}\frac{N_i}{N}(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T$$

线性区别分析：多分类问题

将多类LDA映射投影方向的优化目标$J(\boldsymbol{W})$改为：

$$J(\boldsymbol{W}) = \frac{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}}$$

其中，$\prod_{diag} \mathbb{A}$为矩阵$\mathbb{A}$主对角元素的乘积。

继续对$J(\boldsymbol{W})$进行变形：

$$J(\boldsymbol{W}) = \frac{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}} = \frac{\prod_{i=1}^{r} \boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\prod_{i=1}^{r} \boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i} = \prod_{i=1}^{r} \frac{\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i}$$

显然需要使乘积式子中每个$\frac{\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i}$取值最大，这就是二分类问题的求解目标，即每一个$\boldsymbol{w}_i$都是$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{W} = \lambda \boldsymbol{W}$的一个解。

线性区别分析：线性判别分析的降维步骤

对线性判别分析的降维步骤描述如下：

1. 计算数据样本集中每个类别样本的均值

2. 计算类内散度矩阵S_w和类间散度矩阵S_b

3. 根据$S_w^{-1}S_bW = \lambda W$来求解$S_w^{-1}S_b$所对应前r个最大特征值所对应特征向量$(w_1, w_2, \ldots, w_r)$，构成矩阵W

4. 通过矩阵W将每个样本映射到低维空间，实现特征降维。

线性区别分析：与主成分分析法的异同

	线性判别分析	主成分分析
是否需要样本标签	监督学习	无监督学习
降维方法	优化寻找特征向量$\boldsymbol{w}$	优化寻找特征向量$\boldsymbol{w}$
目标	类内方差小、类间距离大	寻找投影后数据之间方差最大的投影方向
维度	LDA降维后所得到维度是与数据样本的类别个数K有关	PCA对高维数据降维后的维数是与原始数据特征维度相关

谢谢!