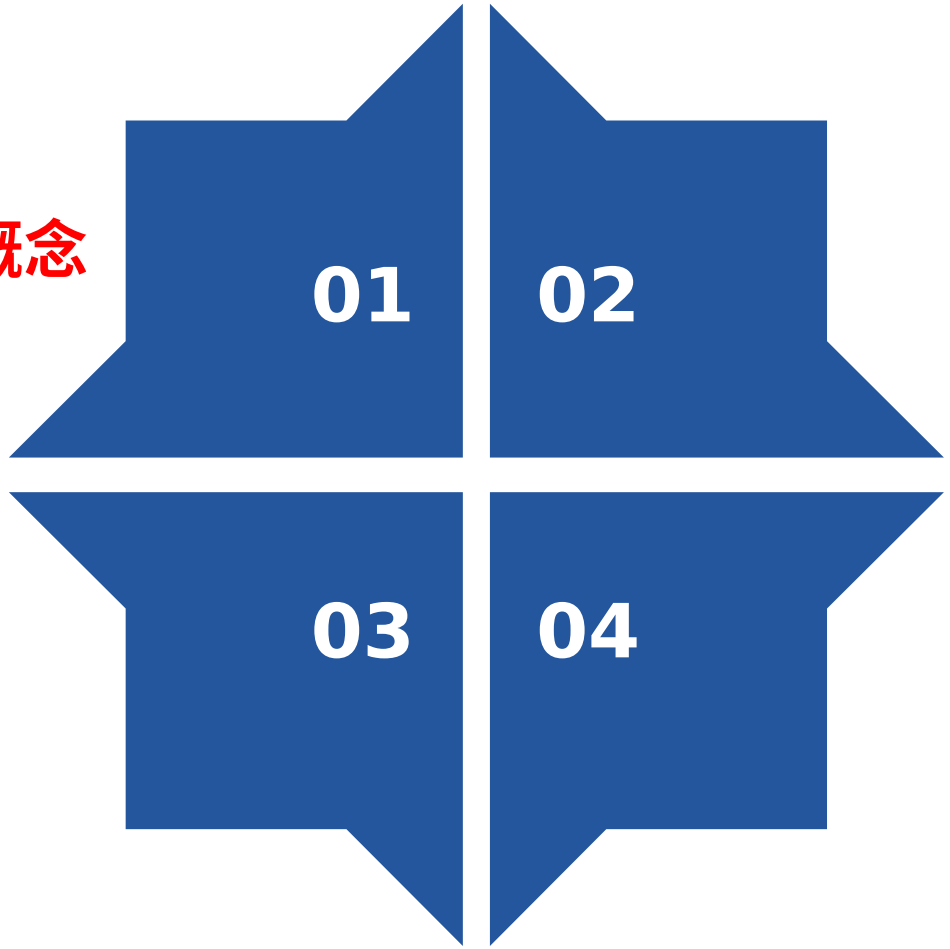
课程《人工智能原理与技术》

同济大学 TONGJI UNIVERSITY

第七周 机器学习与有监督学习

同济大学 TONGJI UNIVERSITY

目录



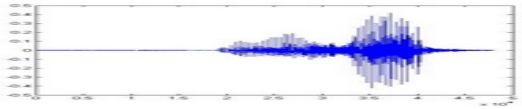
机器学习基本概念

有监督学习

回归分析

决策树

机器学习 ≈ 构建一个映射函数

语音识别 $f($  $)=$ “你好”

图像识别 $f($  $)=$ “猫”

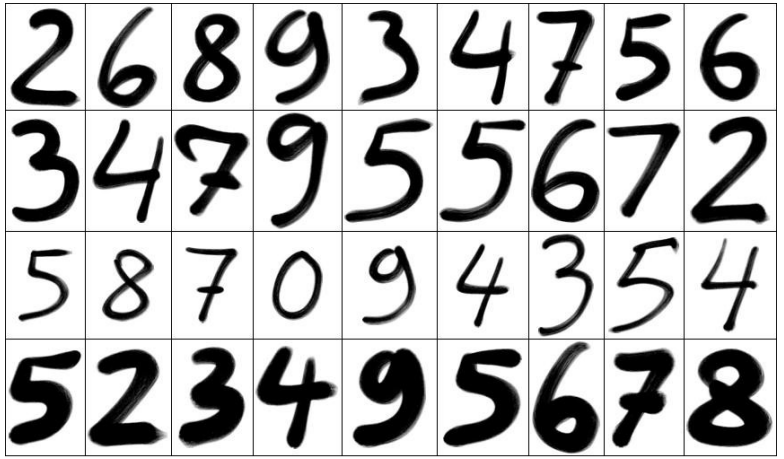
围棋 $f($  $)=$ “5-5”（落子位置）

对话系统 $f($ “你好” $)=$ “今天天气真不错”

用户输入　　机器

为什么要“机器学习”？

现实世界的问题都比较复杂
很难通过规则来手工实现



机器学习概念

- 机器学习通过对数据的优化学习，**建立能够刻画数据中所蕴含语义概念或分布结构等信息的模型**。

  在模型学习过程中，采用合适手段来利用有标签数据或无标签数据， 对模型参数不断进行优化，从而提升模型性能。

- IBM 公司的工程师阿瑟·塞缪尔第一次使用了“机器学习”（1959 年 7 月）

  让机器自行学习，而不需要明确编程

  编程：只能按部就班完成预设任务

  像人一样具有自我学习能力

机器学习类型

- 有监督学习
  数据集中包含标签
- 无监督学习
  从无标签数据出发学习映射函数
- 半监督学习
  学习映射函数过程中使用的一部分数据有标签、一部分数据没有标签
- 强化学习

没有免费的午餐

- 为了在训练优化针对不同的任务，往往需要采用不同机器学习模型， 1995 年， David Wolpert 等学者在所提出了“没有免费午餐定理 (No Free Lunch Theorem)” 指出：**任何一个机器学习模型如果在一些训练集以外的样本误差小（ off-training set error ），那么必然在另外一些训练集以外的样本上表现欠佳，任何模型在平均意义上而言其性能都是一样的，即没有放之四海而皆准的最好算法**。
- 似乎这一定理给机器学习带来了一个令人沮丧的事实（即针对某一域的所有问题，所有算法的期望性能是相同的），但是这一定理也告诉我们，离开具体场景和问题去讨论采用哪种机器学习算法是毫无意义的，应该在机器学习中合理引入已有先验假设对模型进行约束，以提升模型效果，如在自然语言理解中引入句子中单词和单词之间的上下文关联（诸如 n-gram 文法）、在视觉图像分析引入像素点之间的空间依赖（诸如卷积算子）等。

模型评估与参数估计手段：损失函数

表 4.1　常见损失函数的定义

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

将映射函数记为f、第i个训练数据记为(x_i, y_i)以及f对x_i的预测结果记为$\hat{y}_i$（即$\hat{y}_i = f(x_i)$，可定义损失函数$Loss(f(x_i), y_i)$来估量预测值$\hat{y}_i$和真实值y_i之间差异。很显然，在训练过程中希望映射函数在训练集上累加差异最小，即$\min \sum_{i=1}^{n} Loss(f(x_i), y_i)$。

经验风险与期望风险

同济大学 TONGJI UNIVERSITY

经验风险

映射函数f在训练集上所产生损失一般被称为经验风险$\Re_{emp}$ (empirical risk)。经验风险越小说明模型对训练集数据拟合程度越好。经验风险被定义为：

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

$\neq$

期望风险

如果知道某一任务包含的所有数据，则可以从所有数据中计算模型产生的损失，这一误差损失被称为期望风险$\Re$（expected risk），即真实风险或真实误差。记该任务中所有数据的联合分布为$P(x,y)$，期望风险被定义为：

$$\int_{x\times y} Loss\big(y, f(x)\big)P(x,y)dxdy$$

经验风险最小化

当然，由于无法事先就得到任何任务所对应的所有数据分布（如无法采取世界中所有人脸图像来笃信完成人脸识别），使得计算期望风险这一目标可望不可及。因此，机器学习中模型优化目标一般为**经验风险最小化（empirical risk minimization）**，虽然机器学习的目标是追求期望风险最小化，即不断提升模型泛化能力

期望风险$\mathfrak{R}$与经验风险$\mathfrak{R}_{emp}$之间存在如下关系：

$$\underbrace{\mathfrak{R}}_{\text{期望风险}} \leq \underbrace{\mathfrak{R}_{emp}}_{\text{经验风险}} + err$$

其中err取值与机器学习模型的复杂程度和训练集样本数目有关。在模型训练过程中，如果使用同一批训练数据反复训练，模型会变得越复杂，虽然经验风险$\mathfrak{R}_{emp}$会降低，但是err取值会越大，导致期望风险$\mathfrak{R}$增加，这一现象被称为过学习（overfitting）。

模型泛化能力与经验风险和期望风险之间关系

- 泛化能力
  - 模型在训练集上所取得性能与在测试集上所取得性能保持一致

经验风险	期望风险	模型泛化能力
经验风险小 （训练集上表现好）	期望风险小 （所有数据上表现好）	泛化能力强
经验风险小 （训练集上表现好）	期望风险大 （所有数据上表现不好）	过学习 （模型过于复杂）
经验风险大 （训练集上表现不好）	期望风险大 （所有数据上表现不好）	欠学习
经验风险大 （训练集上表现不好）	期望风险小 （所有数据上表现好）	“神仙算法”或“黄粱美梦”

模型泛化能力与经验风险和期望风险之间关系

优化　　正则化

经验风险最小　降低模型复杂度

如前“没有免费午餐定理”所指出，在模型优化中引入恰当先验约束可提升模型性能。为了防止过学习，结构风险最小化 (structural risk minimization) 引入正则化 (regularizer) 或惩罚项 (penalty term) 来降低模型模型复杂度，既最小化经验风险、又力求降低模型复杂度，在两者之间寻找平衡：

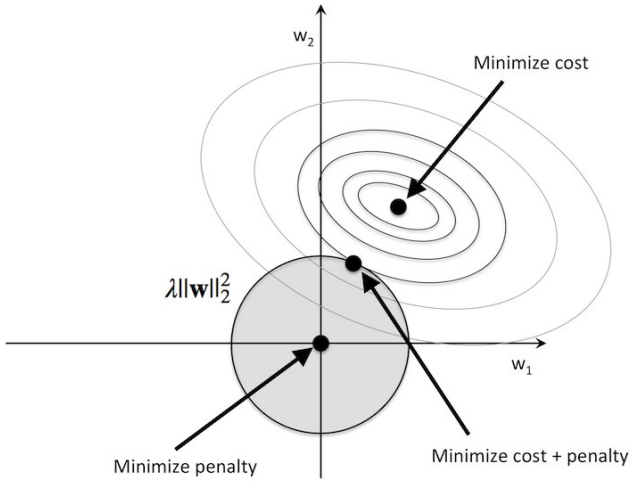
$$\frac{1}{n}\sum_{i=1}^{n} Loss(y_i, f(x_i)) + \lambda J(f)$$

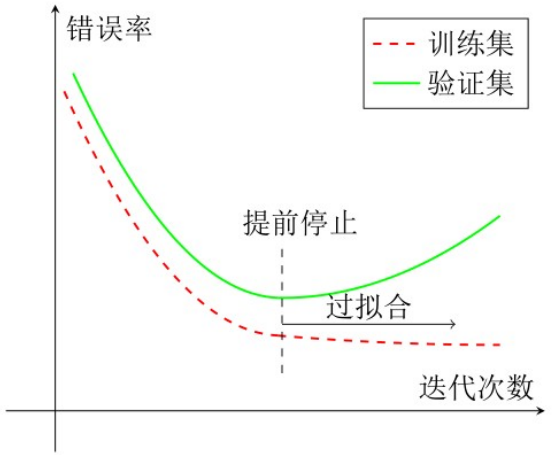
其中$J(f)$是正则化因子或惩罚项因子，λ是用来调整惩罚强度的系数。哲学领域的奥卡姆剃刀定律（Occam's Razor, Ockham's Razor）阐明了“如无必要，勿增实体”的意义，即“简单有效原理”。老子《道德经》曾写道，“万物之始，大道至简，衍化至繁”。在模型中加入约束（如约束模型系数稀疏等），使得从数据到模型的建模过程中，能够“化繁为简、大巧不工”。

正则化

所有损害优化的方法都是正则化

- 增加优化约束：L1/L2约束、数据增强
- 干扰优化过程：权重衰减、随机梯度下降、提前停止





模型度量方法

机器学习模型需要若干性能度量指标来判断其性能优劣。

下面以二分类问题（正类、负类）为例，介绍几种主要度量方法。

n为训练样例的总数，正例总数和负例总数分别是P(positive)和N(negative)。

机器模型预测类别可分为如下四类:

真正例（True Positive, TP）

假正例（False Positive, FP）

真反例（True Negative, TN）

假反例（False Negative, FN）

		预测情况	
		正例（Positive）	反例（Negative）
真实情况	正例（Positive）	TP（True Positive）	FN（False Negative）
	反例（Negative）	FP（False Positive）	TN（True Negative）

令TP、FP、TN、FN分别表示其对应的样例数。

模型度量方法

- **准确率(accuracy)**：$ACC = \frac{TP+TN}{P+N}$。很显然，如果正负样本比例不平衡，ACC 不是一个度量模型好的方法。比如，某一种恶性疾病很罕见（如1万个疑似患者中仅有1人罹患该疾病），机器学习模型可将所有患者均识别为负类，从而保证ACC取值极高，但这就忽略了这一模型应该关注的问题。
- **错误率（error rate）**：$errorRate = \frac{FP+FN}{P+N}$，显然有$errorRate = 1 - ACC$。
- **精确率（precision）**：$precision = \frac{TP}{TP+FP}$，也叫查准率，表示被模型预测为正例的样本中实际为正例的比例。
- **召回率（Recall）**：$recall = \frac{TP}{TP+FN}$，也叫查全率，表示所有正例样本中被模型预测为正例的比例

模型度量方法

- 在实际应用中，精确率和召回率之间是相互矛盾的
  - 例：将所有样本分类为正例
    - 召回率：100%
    - 精确率：极低
  - 为了综合考虑精确率和召回率，可采用F1-score这一综合分类率：

$$\mathrm{F1}-score=\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}=2\times\frac{Precision\times Recall}{Precision+Recall}$$

- F1-score是精确率和召回率的调和平均数

参数优化：频率学派与贝叶斯学派

同济大学 TONGJI UNIVERSITY

频率学派

在频率学派中，频率是概率的经验基础，概率表示的是事件发生频率的极限值。当重复试验的次数趋近于无穷大时，事件发生的频率会收敛到真实概率，即“频率依概率收敛于概率”。从频率学派角度而言，对模型参数优化学习的结果就是得到使观测数据发生概率最大的模型参数，又称为最大似然估计（maximum likelihood estimation，MLE）。这里的最大似然可理解为通过调整模型参数使得模型能够最大化样本情况出现的概率。

贝叶斯学派

在贝叶斯学派中，事件发生的频率既与当前观测数据有关，又与对该事件已获得的历史先验知识有关。从贝叶斯学派角度而言，对模型参数优化学习的结果就是似然概率（模型参数产生数据的概率）与先验概率（没有任何实验数据时对模型参数的经验判断）乘积最大，又称为最大后验估计（maximum a posteriori estimation, MAP）。这里的最大后验估计可理解为最大化在给定数据样本的情况下模型参数的后验概率。

参数优化：频率学派与贝叶斯学派

同济大学 TONGJI UNIVERSITY

知识卡片：频率学派与贝叶斯学派

频率学派的体系化理论于20世纪初期由费希尔（Fisher）、皮尔逊（Karl Pearson）、内曼（Neyman）等创立，如费希尔提出最大似然估计方法，皮尔逊提出Pearson卡方检验和Pearson相关系数，内曼提出置信区间（confidence interval）等概念。贝叶斯学派可追溯至在贝叶斯在去世两年后（1763年）才发表的《机遇理论中一个问题的解》一文（Bayes，1763），经过高斯（Gauss）和拉普拉斯（Laplace）的发展而逐渐创立。

原则而言，频率学派认为“事件本身就具有客观的不确定性”，事件在大量独立重复实验中发生的频率趋于事件发生的概率。由于不知道引发事件产生的模型参数具体的取值，因此引入最大似然和置信区间以在参数空间中寻找最优的参数。

贝叶斯学派认为模型参数都是随机变量，服从某个概率分布，具有不确定性，因此需要对模型参数设定一个概率分布（先验概率），通过结合已经观测得到的证据来不断调整模型参数的概率分布，最终得到一个正确的分布（后验概率）。为此，贝叶斯学派引入了先验分布（prior distribution）和后验分布（posterior distribution）来优化最优模型参数。

假设由 n 个数据样本构成的集合 $D=\{x_1, x_2, \cdots, x_n\}$ 从参数为 Θ 的某个模型（如高斯模型等）以一定概率独立采样得到。于是，可以通过最大似然估计算法来求取参数 Θ，使得在参数为 Θ 的模型中数据集 D 出现的可能性最大，即 $\hat{\Theta}=\underset{\Theta}{\operatorname{argmax}} P(D|\Theta)$。

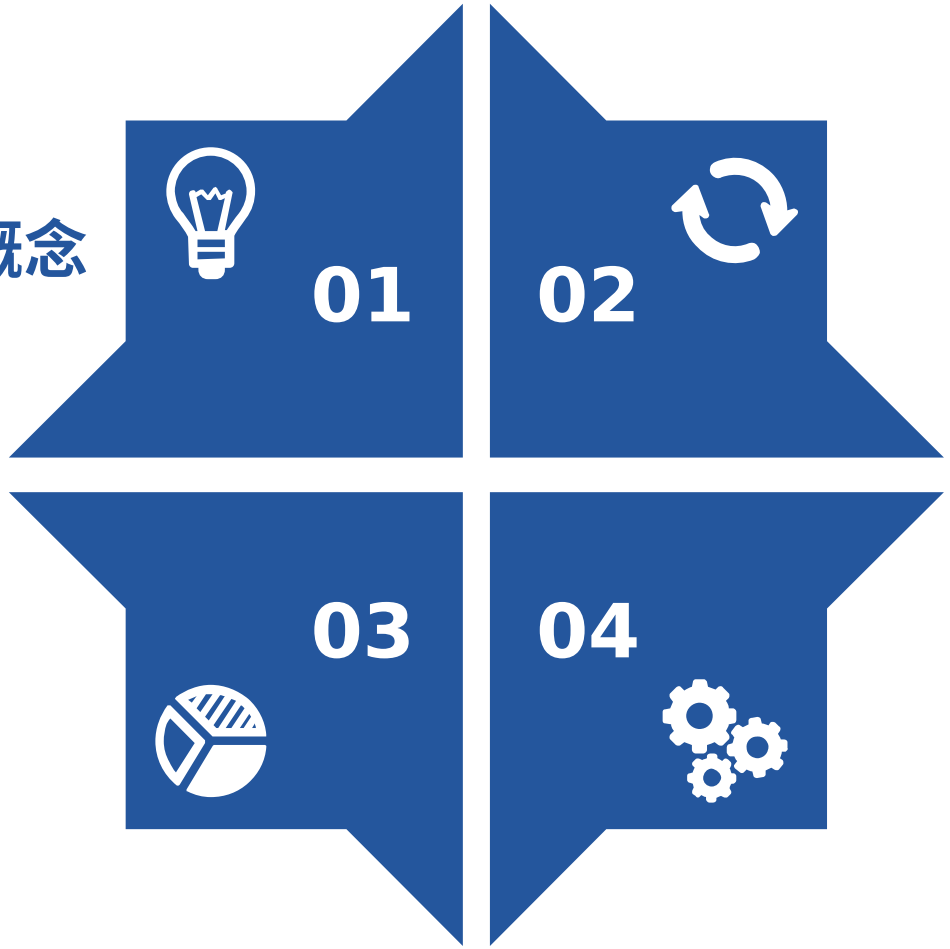
或者也可利用最大后验估计从数据集 D 如下估计参数 Θ：$\hat{\Theta}=\underset{\Theta}{\operatorname{argmax}} P(\Theta|D)=\underset{\Theta}{\operatorname{argmax}} \frac{P(D|\Theta)P(\Theta)}{P(D)}$。由于 $P(D)$ 与 Θ 无关（所以可作为常量省略），则可得 $\underset{\Theta}{\operatorname{argmax}} \frac{P(D|\Theta)P(\Theta)}{P(D)}=\underset{\theta}{\operatorname{argmax}} P(D|\Theta)P(\Theta)$，对这个式子取对数，得到 $\underset{\Theta}{\operatorname{argmax}} \log P(D|\Theta)+\log P(\Theta)$。这里的 $P(\Theta)$ 是先验概率（prior）、$P(D|\Theta)$ 是似然概率（likelihood）、$P(\Theta|D)$ 是后验概率（posterior）。

当然，无论是最大似然估计算法还是最大后验估计算法，都需要对所优化目标中的参数求导，令导数为0，以求取模型的参数值。

频率学派的优化手段可形象比喻为“以史为鉴，可以知兴替”，贝叶斯学派的优化手段可形象比喻为“天行有常（先验概率），不为尧存，不为桀亡”。

同济大学 TONGJI UNIVERSITY

目录



机器学习基本概念

有监督学习

回归分析

决策树

有监督学习

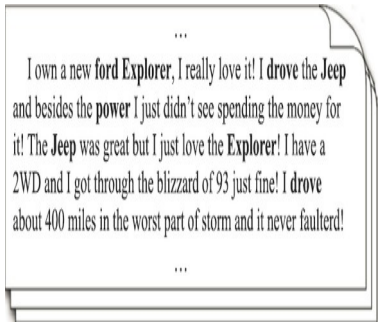
- 监督学习是一种在实践中运用最为广泛的一种机器学习方法，其目标是给定带有标签信息数据的训练集$\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{n}$，学习一个从输入$x_i$到输出$y_i$的映射。$x_i$可是文档、图像、音频或蛋白质基因等数据或者数据的特征表达，y_i为所对应的论文类别、人脸对象、歌曲语音或生命功能等语义内容，其中$\mathcal{D}$被称为训练集，n是训练样例的数量。
- 监督学习算法从假设空间（hypothesis space）学习得到一个**最优映射函数f（又称决策函数）**，映射函数f将输入数据映射到语义标注空间，实现数据的分类和识别。无监督学习则是直接从无标签数据$\{x_i, i=1,...,n\}$出发学习映射函数，而半监督学习在学习映射函数过程中使用的一部分数据有标签、一部分数据没有标签。

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

有监督学习

学习输入和输出之间的映射
典型任务
分类
回归

分类

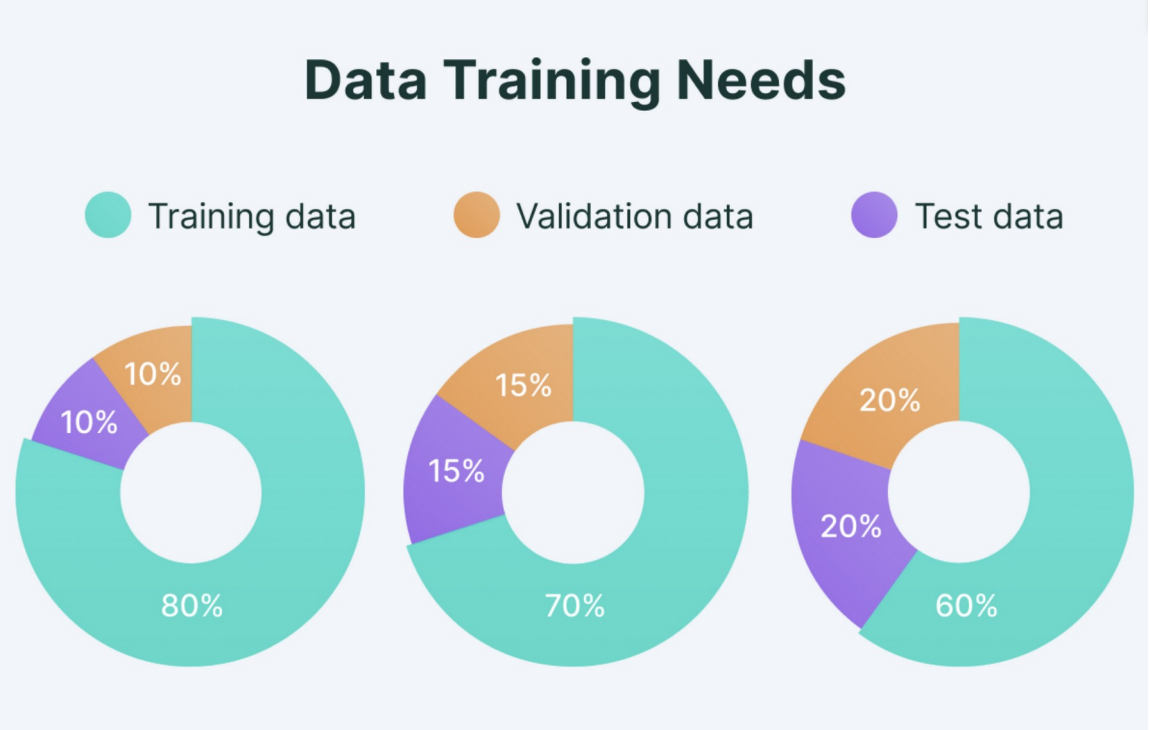
输入 - 城市	输出 - 国家
东京	日本
上海	中国
巴黎	法国
阿姆斯特丹	荷兰
…	…

回归

输入 - 房屋面积	输出 - 价格
85 平米	71 万元
120 平米	140 万元
60 平米	63 万元
150 平米	112.5 万元
…	…

有监督学习：训练集、验证集、测试集

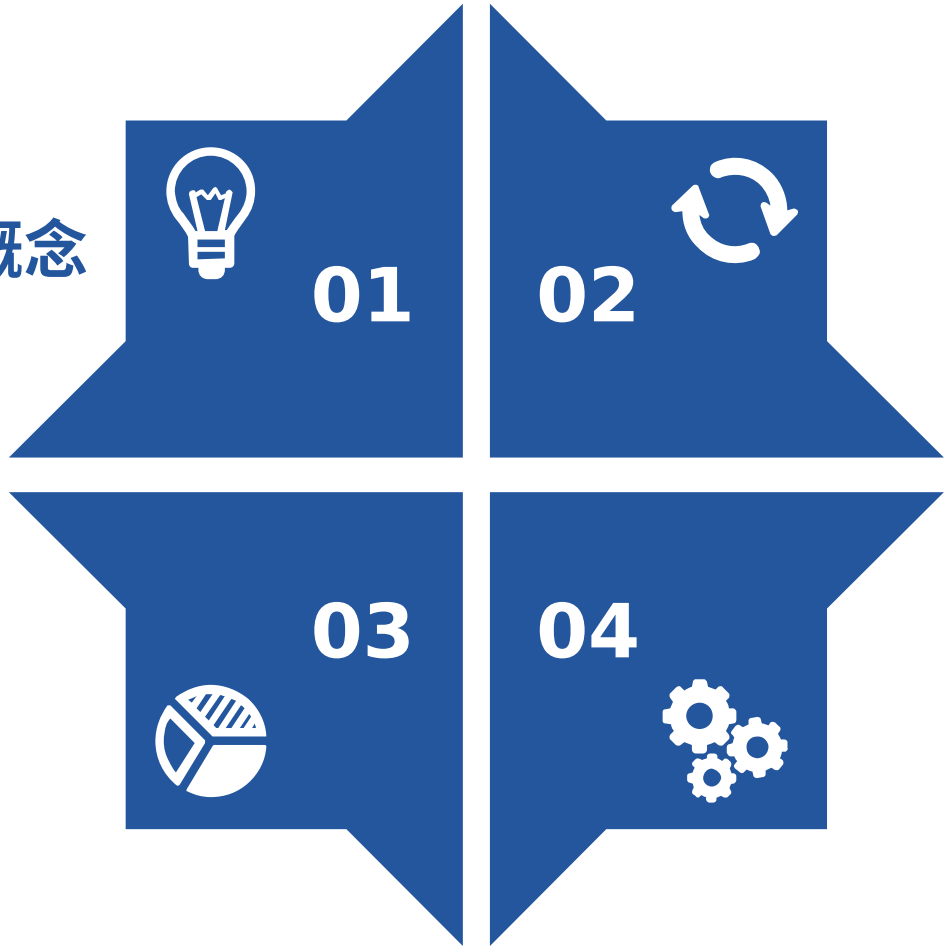
- 一旦在**训练集**上完成了模型参数优化后，需要在测试数据集上对模型性能进行测试。为了在训练优化过程中挑选更好的模型参数，一般可将训练集中一部分数据作为**验证集（validation set）**。在训练集上训练模型的同时会在验证集上对模型进行评估，以便得到最佳参数，最后在**测试集**上进行测试，将测试结果作为模型性能最终结果。
- 要注意的是，训练集、验证集和测试集所包含数据之间没有任何交叉。可以说，训练集用于模型训练（好比学生的练习册）、验证集用于评估模型以调整相应参数（好比学生的模拟考卷或小测验）、测试集用于得到模型的优劣水平（好比真正考试）。



训练集、验证集和测试集三种数据集中数据比例

同济大学 TONGJI UNIVERSITY

目录



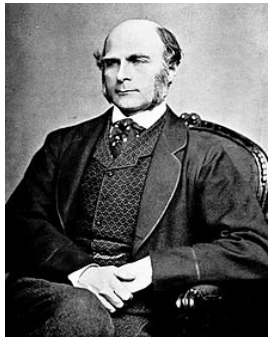
回归分析

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作回归分析，刻画不同变量之间关系的模型称为回归模型。

$$y = 33.73(英寸) + 0.516x$$

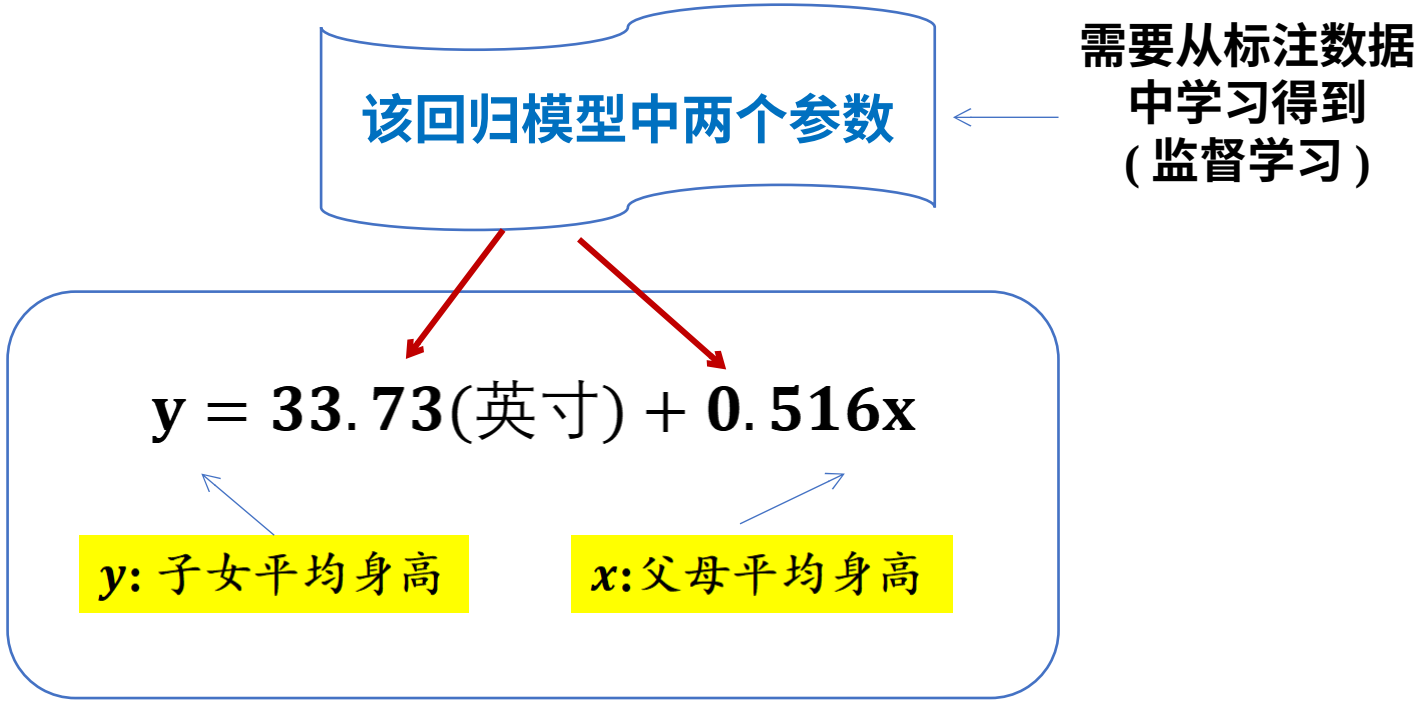
y: 子女平均身高　　x: 父母平均身高

- 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。



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

回归分析

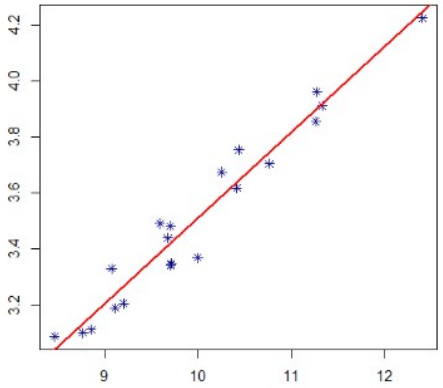


- 给出任意一对父母平均身高，则可根据上述方程，计算得到其子女平均身高
- 从父母平均身高来预测其子女平均身高
- 如何求取上述线性方程（预测方程）的参数?

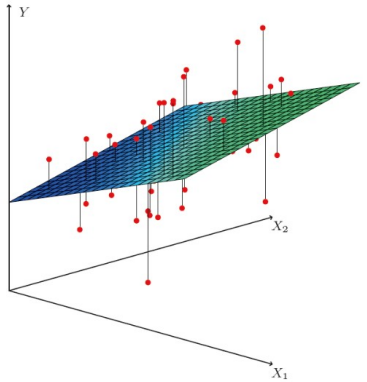
回归的分类

（1）按照涉及的变量多少，分为

一元回归



多元回归



（2）按照自变量和因变量之间的关系类型，可分为

线性回归

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

非线性回归

$$y = \alpha \cdot e^{\beta x} \text{ 或 } y = \alpha \cdot x^{\beta}$$

线性回归作用

- **预测**

  根据输入变量预测输出变量的值。

预测子女身高

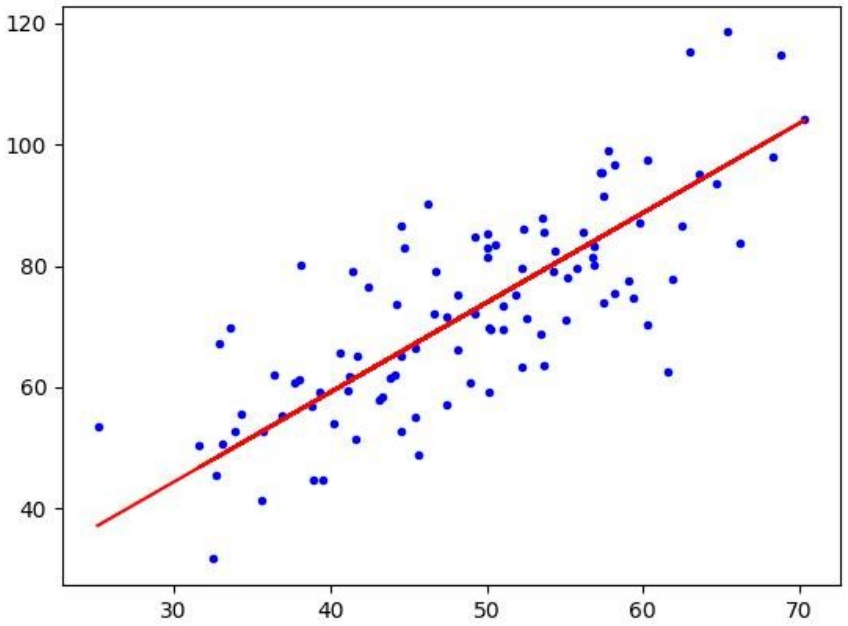
$$y = 33.73(英寸) + 0.516x$$

子女平均身高　　父母平均身高

- **解释关系**

  通过模型参数，解释输入变量与输出变量之间的关系。

亲子身高关系

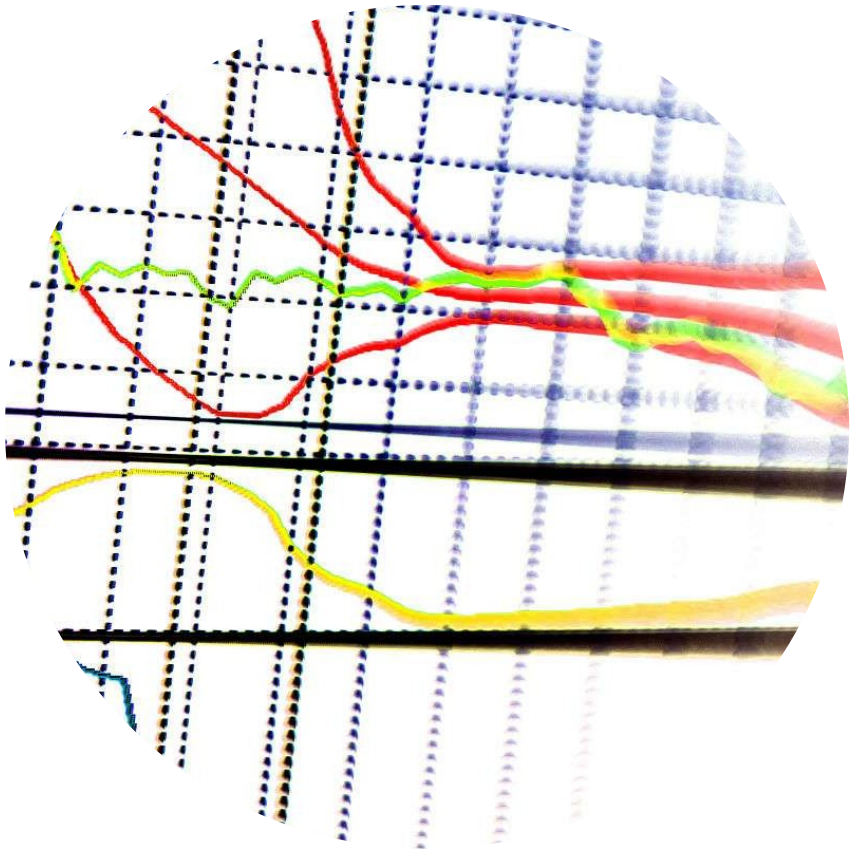


- **数据建模**

  将实际数据拟合到线性模型中，以便更好地理解和分析数据。

代际分布特征

线性回归应用领域



社会科学

线性回归模型用于研究社会现象，如教育水平、收入、健康状况等之间的关系。

医学

在医学研究中，线性回归模型可用于探索疾病与风险因素之间的关系。

经济学

经济学家使用线性回归模型分析经济指标之间的关系，如 GDP 、失业率、通货膨胀等。

回归分析：一元线性回归模型

同济大学 TONGJI UNIVERSITY

表4.3 给出了芒提兹尼欧（Montesinho）地区发生森林火灾的部分历史数据，表中列举了每次发生森林火灾时的气温温度取值x和受到火灾影响的森林面积y。

表 4.3　芒提兹尼欧地区发生森林火灾的部分数据

气温 x	5.1	8.2	11.5	13.9	15.1	16.2	19.6	23.3
火灾影响面积 y	2.14	4.62	8.24	11.24	13.99	16.33	19.23	28.74

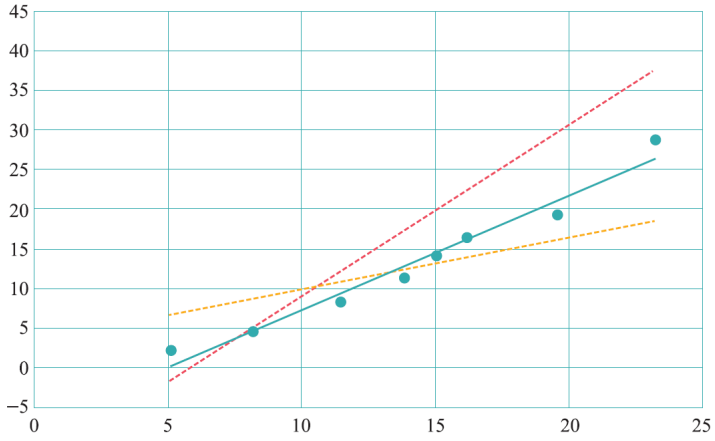


图 4.1　火灾影响的森林面积与气温之间的关系

一元线性回归实际上就是寻找一条用$y = ax + b$表达的直线，使得这条直线尽可能靠近或穿过这8组(x, y)数据，即能够以最小误差来拟合这8组(x, y)数据。

回归分析：一元线性回归模型

最佳回归模型将使得残差平方和的平均值$\frac{1}{N}\sum(y-\tilde{y})^2$最小，残差即预测值和真实值之间的差值。残差平方和的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。

一般而言，回归模型$y_i = ax_i + b\ (1 \le i \le n)$的参数求解过程为：记在当前参数下第$i$个训练样本$x_i$的预测值为$\hat{y}_i$，计算$x_i$的标注值（实际值）$y_i$与预测值$\hat{y}_i$之差的平方$(y_i - \hat{y}_i)^2$，计算训练集中$n$个样本所产生误差总和$L(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$，使用最小二乘法找到误差总和最小，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。

表 4.3　芒提兹尼欧地区发生森林火灾的部分数据

气温 x	5.1	8.2	11.5	13.9	15.1	16.2	19.6	23.3
火灾影响面积 y	2.14	4.62	8.24	11.24	13.99	16.33	19.23	28.74

回归分析：一元线性回归模型参数求解

一般而言，回归模型$y_i = ax_i + b\ (1 \le i \le n)$的参数求解过程为：记在当前参数下第$i$个训练样本$x_i$的预测值为$\hat{y}_i$，计算$x_i$的标注值（实际值）$y_i$与预测值$\hat{y}_i$之差的平方$(y_i - \hat{y}_i)^2$，计算训练集中$n$个样本所产生误差总和$L(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$，使用最小二乘法找到误差总和最小，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。

优化目标：

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

损失函数对b求偏导：

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

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

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

$$\Rightarrow n\bar{y} - na\bar{x} - nb = 0$$

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

这样就得到了参数b的计算公式。

回归分析：一元线性回归模型参数求解

一般而言，回归模型$y_i = ax_i + b\ (1 \le i \le n)$的参数求解过程为：记在当前参数下第$i$个训练样本$x_i$的预测值为$\hat{y}_i$，计算$x_i$的标注值（实际值）$y_i$与预测值$\hat{y}_i$之差的平方$(y_i - \hat{y}_i)^2$，计算训练集中$n$个样本所产生误差总和$L(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$，使用最小二乘法找到误差总和最小，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。

损失函数对a求偏导：

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

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

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

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

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

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

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

回归分析：一元线性回归模型参数求解

可以看出只要给出了训练样本$(x_i, y_i)(i = 1, ..., n)$，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。

这样，对于上面的案例，可以求得参数a和b分别为：

$$a = \frac{x_1 y_1 + x_2 y_2 + ... + x_8 y_8 - 8\bar{x}\bar{y}}{x_1^2 + x_2^2 + ... + x_8^2 - 8\bar{x}^2} = 1.428$$

$$b = \bar{y} - a\bar{x} = -7.09$$

即预测芒提兹尼欧地区火灾所影响森林面积与气温温度之间的一元线性回归模型为“火灾所影响的森林面积 $= 1.428 \times$ 气温温度 $- 7.09$”，即$y = 1.428x - 7.09$

回归分析：从一元线性回归到多元线性回归

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。

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$\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{m}$，其中$\boldsymbol{x_i} = [x_{i,1}, x_{i,2}, \ldots, x_{i,D}] \in \mathbb{R}^D$，$D$为数据特征的维度，线性回归就是要找到一组参数$\boldsymbol{a} = [a_0, a_1, \ldots, a_D]$，使得线性函数：

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

最小化均方误差函数：

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

表 4.4　芒提兹尼欧地区历史森林火灾的部分数据（加入风力因素）

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	4.0	6.3	4.0	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

回归分析：从一元线性回归到多元线性回归

为了方便，我们使用矩阵来表示所有的训练数据和数据标签。

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

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：

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

我们知道在特征维度只有一维的时候线性回归有闭式解（closed form solution），多维情况下同样存在。

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

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

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

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

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

对于上面的例子，转化为矩阵的表示形式为：

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

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

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

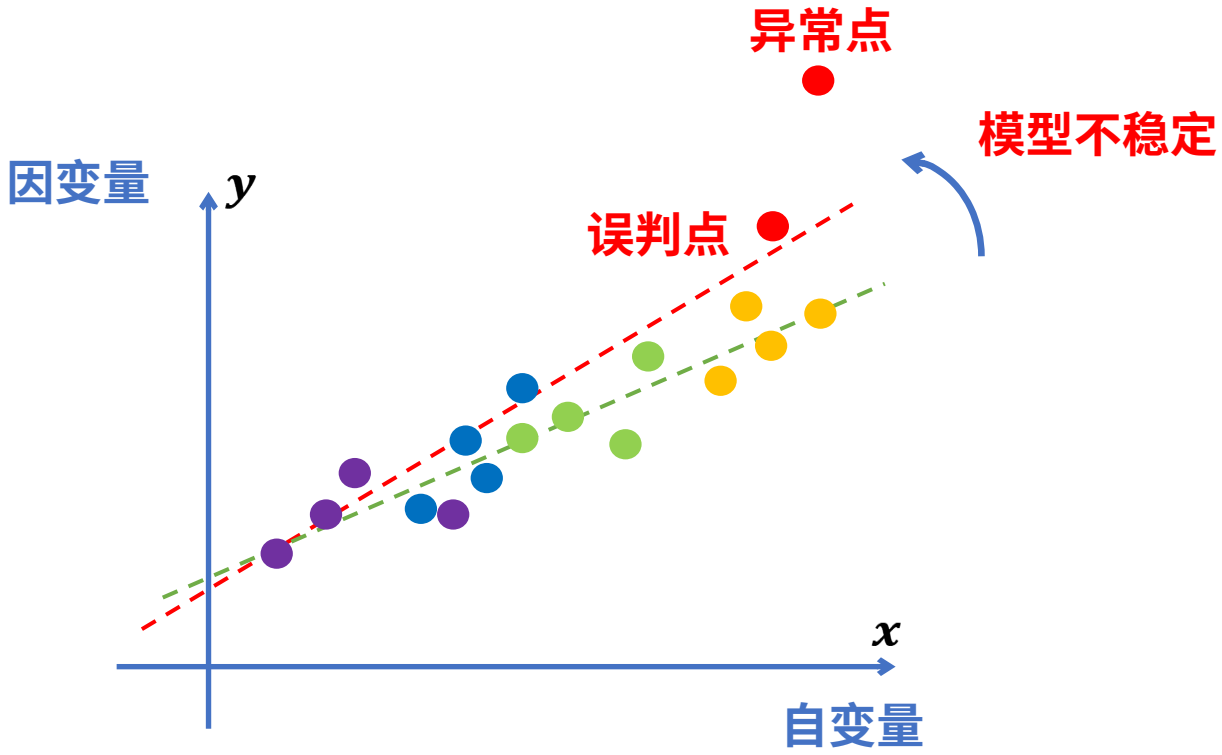
$$\boldsymbol{a} = [1.312 \quad 0.626 \quad -9.103]$$

所以对应的线性函数为：

$$y = -9.103 + 1.312x + 0.626z$$

回归分析：从线性回归到非线性回归

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，可考虑逻辑斯蒂回归 (logistic regression)[Cox 1958] 。



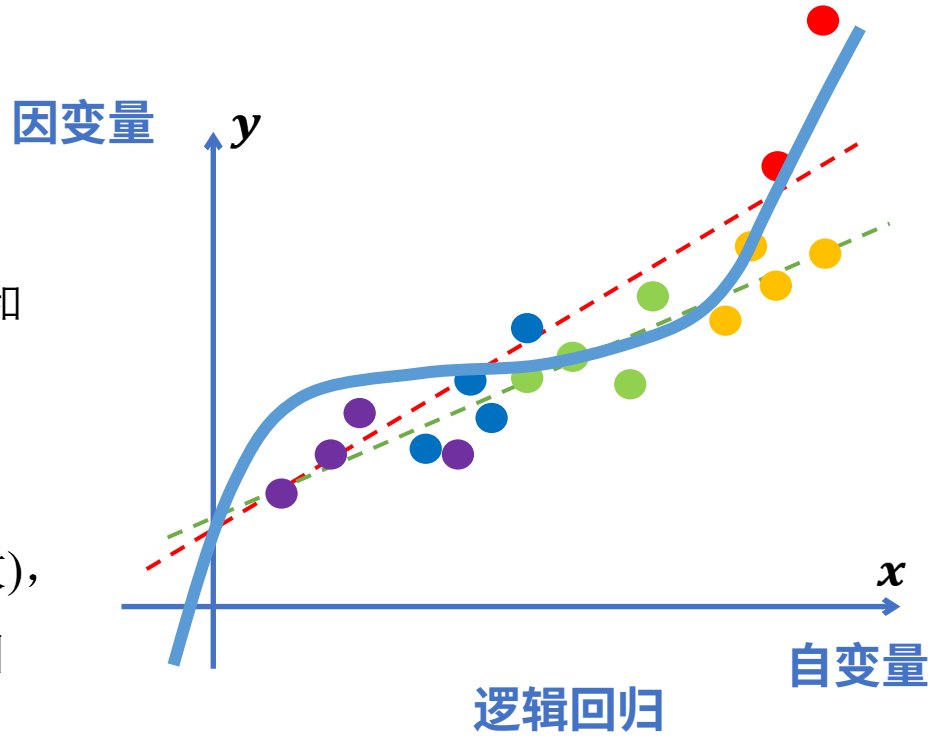
回归分析：从线性回归到非线性回归

逻辑斯蒂回归(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。

logistic回归只能用于解决二分类问题，将它推广为多项逻辑斯蒂回归模型(multi-nominal logistic model，也即softmax函数)，用于处理多类分类问题，可以得到处理多类分类问题的softmax回归。

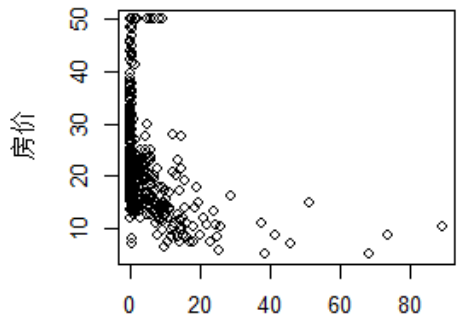


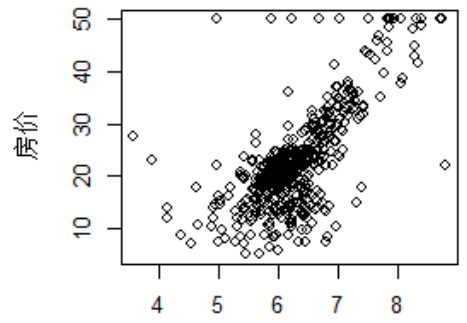
案例一：房价预测

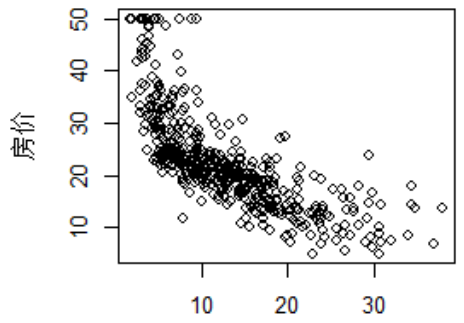
该数据集包含美国人口普查局收集的美国马萨诸塞州波士顿住房价格的有关信息。该数据集共有 506 个样本，13 个属性，其中包括 12 个特征变量和 1 个目标变量（房价中位数），取以下 4 个特征作为自变量。

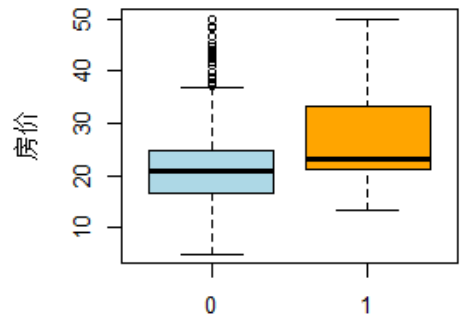
因变量：房价（Y）

自变量：犯罪率（X1）房间数（X2）
低收入人群占比（X3） 是否沿河（X4）









案例一：房价预测

基于最小二乘法计算线性回归模型

最小化残差平方和 $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ ➡ $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{y}}$

⬇

$$\hat{Y}_i = -1.921 - 0.097X_{i1} + 5.076X_{i2} - 0.582X_{i3} + 3.998X_{i4}$$

犯罪率 (X_1)：**控制其他因素不变时**，犯罪率每增加一个百分点，房价**平均降低** 0.097 个单位，即 97 美元

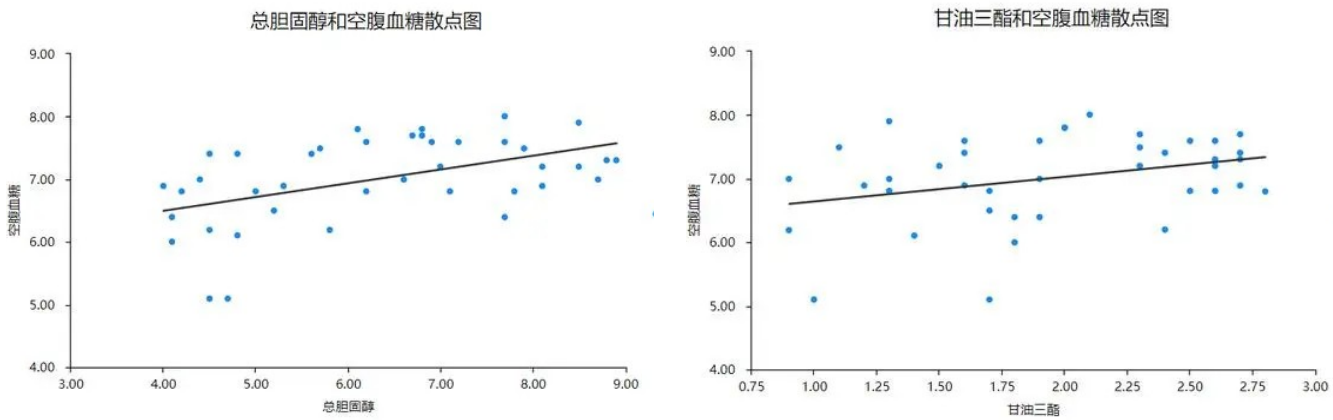
房间数 (X_2)：**控制其他因素不变时**，房间数每增加一个，房价**平均增加** 5.076 个单位，即 5067 美元

低收入人群占比（X3）和是否沿河（X4）如何解读?

案例二：医学分析

某医师研究糖尿病患者的总胆固醇和甘油三酯对空腹血糖的影响，调查 40 名糖尿病患者的总胆固醇、甘油三酯和空腹血糖的测量值如下，试作统计分析。

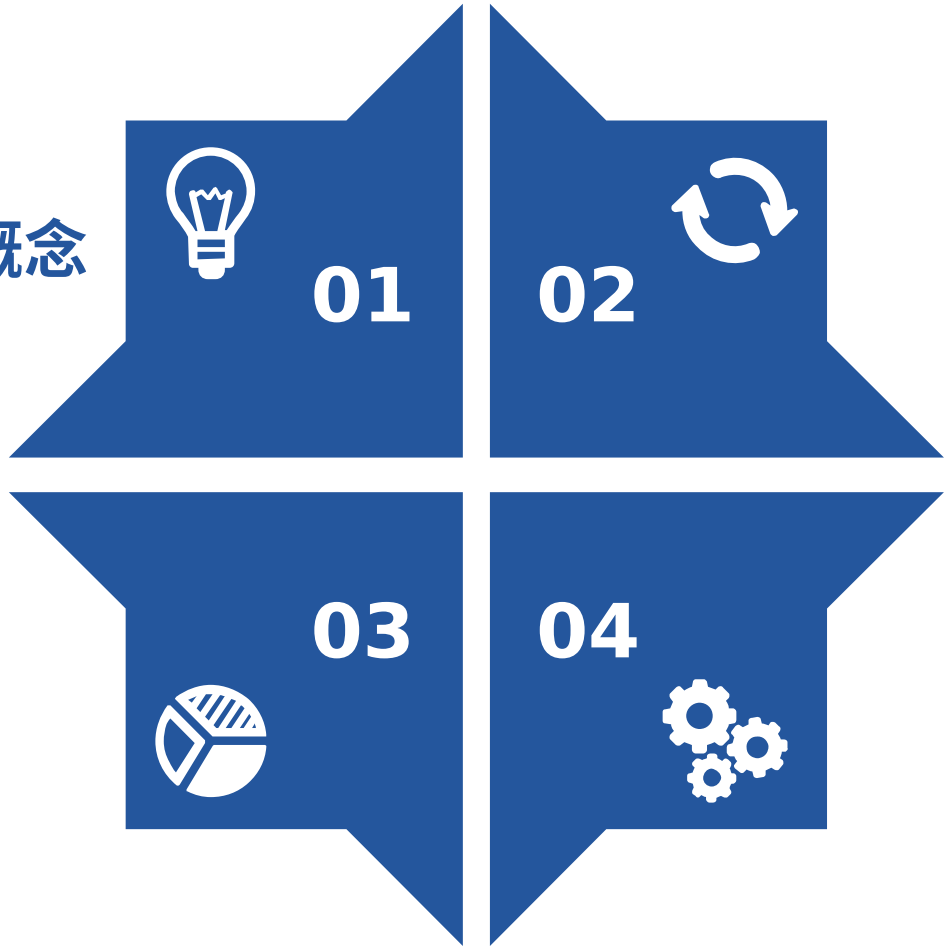
	A	B	C	D
1	编号	总胆固醇	甘油三酯	空腹血糖
2	1	5.70	1.10	7.50
3	2	6.60	0.90	7.00
4	3	7.10	1.30	6.80
5	4	7.00	2.30	7.20
6	5	6.80	2.30	7.70
7	6	6.10	2.00	7.80
8	7	8.90	2.70	7.30
9	8	8.70	1.30	7.00
10	9	8.50	1.50	7.20
11	10	8.80	2.60	7.30
12	11	5.00	2.50	6.80
13	12	5.60	1.60	7.40
14	13	6.90	2.60	7.60
15	14	4.50	1.70	5.10
16	15	4.40	1.90	7.00



空腹血糖 =4.985 + 0.212* 总胆固醇 + 0.351* 甘油三酯

同济大学 TONGJI UNIVERSITY

目录



机器学习基本概念

有监督学习

回归分析

决策树

决策树

- 决策树将分类问题分解为若干基于单个信息的推理任务，采用树状结构来逐步完成决策判断。事实上，人们在逻辑推理过程中经常使用决策树的思想。
- 下面通过一个例子来解释决策树的分类。银行的数据分析师希望通过历史的贷款记录，包括用户的四种特征（年龄，银行流水，婚姻状况，房产状况）以及最终是否给予贷款，来建立分类模型辅助决策者进行决策。
- 银行收集整理的数据如表 4.5 所示。

表 4.5　是否给予贷款与申请人自身状况的关系

序号	年龄 / 岁	银行流水	是否结婚	是否拥有房产	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

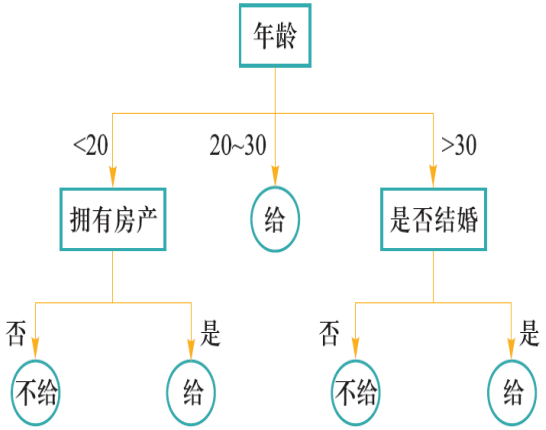


图 4.3　银行贷款决策树

- 第一层是年龄状况，分为小于 20 岁， 20 岁至 30 岁，大于 30 岁三种取值。
- 如果年龄在 20 岁和 30 岁之间，样本子集为 {3, 7, 12, 13} ，这些样本的标签均为“给予贷款”，所以为叶子节点。
- 如果年龄大于 30 ，样本子集为 {1, 2, 8, 9, 11} ，这些样本具有不同的标签，要进一步使用其他属性对这个样本子集进行划分。经观察，通过“是否结婚”这一属性值，可以将该样本子集进一步划分成 {1, 2, 8} （未婚）和 {9, 11} （已婚）两个样本子集。此时这两个样本子集内标签一样，不需要再划分。
- 如果年龄小于 20 岁，样本子集为 {4, 5, 6, 10, 14} ，这些样本具有不同的标签，同样需要继续划分。通过观察，“拥有房产”这个属性值可将该样本子集进一步划分成 {4, 5, 10} （无房产）和 {6, 14} （有房产）两个样本子集。此时这两个样本子集内标签一样，不需要再划分。
- “ 银行流水”这一特点及其属性值在这次决策树构造过程中没有使用。

构建决策树

- 建立决策树的过程，就是：
  - 选择属性值
  - 根据属性值对样本集进行划分
  - 选择下一个属性值
  - 直至每个子集为同一个类别

- 上面的案例数据较少
  - 数据较少：通过观察、穷举的，不断选择属性值对样本集进行划分
  - 数据较多：先选哪个属性？后选哪个属性？

构建决策树

- 顺序
  - 划分属性的顺序选择是重要的
  - 性能好的决策树——越往下信息熵越小

    所包含样本尽可能属于相同类别。
- 信息熵（entropy）
  - 信息熵越大，说明集合的不确定性越大
  - 选择属性划分样本集前后信息熵的减少量被称为信息增益（information gain）

    描述了样本集合复杂度（不确定性）所减少的程度

构建决策树

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。需要指出，所有p_k累加起来的和为1。

构建决策树

现在应用信息熵这个度量标准来构建决策树。表4.5 中14个样本分属于“给予贷款（9 个样本）”和“不给予贷款（5 个样本）”两个类别，即 $K = 2$。

$Ent(D)$

$$=-\sum_{k=1}^{2} p_k \log_2 p_k =-\left(\frac{9}{14}\times\log_2\frac{9}{14}+\frac{5}{14}\times\log_2\frac{5}{14}\right)=0.940$$

表 4.5　是否给予贷款与申请人自身状况的关系

序号	年龄 / 岁	银行流水	是否结婚	是否拥有房产	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

构建决策树

表4.5中有年龄、银行流水、是否结婚、拥有房产四个人物相关的属性特征，下面计算这四个特点所对应的信息熵。

以年龄为例，年龄包含“>30”、“20~30”、“<20”三个属性取值。这三个属性取值对14个样本进行划分，在决策树中产生了三个分支结点，每个分支结点包含一个数据子集，三个数据子集构成了对原数据的划分。如“20~30”这一属性取值包含四个样本{3, 7,12,13}。

对年龄属性划分出的子样本集情况的统计如表4.6所示。这里记属性取值为a_i（即a_i = “年龄 >

表 4.6　按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−)

构建决策树

根据表4.6的统计情况，计算每个属性值划分出的子样本集（即每个分支节点）的信息熵：

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

$$\text{“年龄}20\sim30\text{”：}\quad Ent(D_1) = -\left(\frac{4}{4}\times\log_2\frac{4}{4}+0\right) = 0$$

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

表 4.6　按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−)

构建决策树

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 $A =$ 年龄 代入。于是选择年龄这一属性划分后的信息增益为：

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

按年龄划分**前**的信息熵

按年龄划分**后**的信息熵

减少了多少信息熵

表 4.6　按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−)

构建决策树

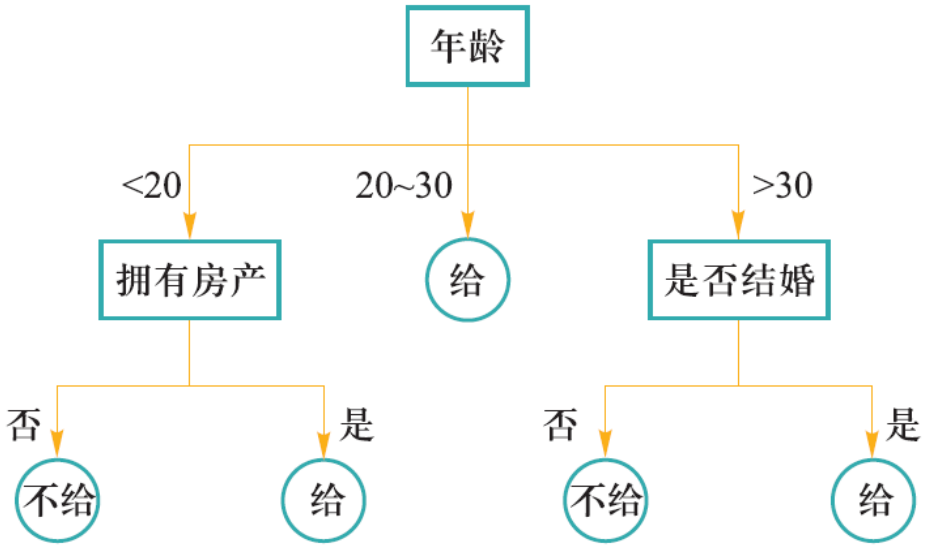
分别计算每种属性划分后带来的信息增益

比较信息增益的高低

选择最佳属性（划分后带来信息增益最大的属性）对原样本集进行划分

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

在该案例中，最终可构建如右图决策树



THANKS !

QUESTION?