BIG DATA

第5章 数值预测

第4章中介绍的分类问题可以看作是一种预测技术，预测的是新样本的类别。另一类常见的预测问题是数值预测。分类问题中预测的是类别，属于类别属性的取值；数值预测则预测的是数值属性的取值。本章将介绍解决数值预测问题的常用技术和方法。

5.1 数值预测的概念

数值预测是一个与分类既相似又有区别的过程。两者都属于有监督学习，解决问题的过程相同，都是先通过训练数据集进行学习得到一个模型，然后利用模型进行预测。不同的是分类问题预测的是类别，而数值预测预测的是数值，通常是连续类型的数值。另一个不同是两者采用的常用技术大多是不同的，数值预测最常用的技术是回归(regression)，因此，有时将回归与数值预测等同。当然，有一些技术是相同的，例如，K 近邻既可以用于分类也可以用于数值预测。有些技术存在相似性，例如，决策树和模型树、回归树的构建过程存在一些共性。

数值预测问题可以这样描述。给定一个样本 x_t，由 $k+1$ 个属性 $A_1,A_2,\cdots,A_k,Y$ 描述，其中属性 Y 为数值属性，称该数值属性为**目标属性**，其他属性称为**描述属性**。假设样本 x_t 前 k 个属性的取值分别为 $x_{t_1},x_{t_2},\cdots,x_{t_k}$，样本 x_t 可以表示为 $x_t=(x_{t_1},x_{t_2},\cdots,x_{t_k})$，要预测其数值属性 Y 的取值，记为 y，需要构建一个样本集，记此样本集为 D，其中每个样本有 $k+1$ 个属性 $A_1,A_2,\cdots,A_k,Y$ 的取值，其中第 i 个样本为 $(x_{i_1},x_{i_2},\cdots,x_{i_k},y_i)$，$i=1,2,\cdots,n$。与分类任务类似，通常将数据集 D 分为两部分：一部分作为学习，用于构建预测模型，称为训练数据集，简称**训练集**；另一部分作为测试模型性能的数据集，称为测试数据集，简称**测试集**。

数值预测有许多实际的应用，例如，预测一个产品的销售量、预测一个客户的消费额度、预测一个客户的月均账户余额、预测一个产品的性能等。这些预测是一个公司或部门的业务运营非常重要的决策依据。下面以一个常用的供研究使用的数据集来进一步解释数值预测问题。该数据集是加州大学欧文分校机器学习数据库(UC Irvine Machine Learning Repository，网址为 http://archive.ics.uci.edu/ml/)中的一个，名为 computer hardware，或者 CPU performance，简称 CPU。该数据完整版包括 209 个样本，每个样本有 10 个属性，其中有 2 个类别属性，8 个数值属性，选择其中的 6 个数值属性作为描述属性，1 个数值属性作为目标属性。表 5.1 是该数据集的一个子集。

表 5.1 数据集 CPU 的子集

MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
125	256	6000	256	16	128	198
29	8000	32 000	32	8	32	269
29	8000	32 000	32	8	32	220
29	8000	32 000	32	8	32	172
26	8000	32 000	64	8	32	318
23	16 000	32 000	64	16	32	367
23	16 000	32 000	64	16	32	489
23	16 000	64 000	64	16	32	636
23	32 000	64 000	128	32	64	1144
400	512	3500	4	1	6	40

表 5.1 中的各属性对应机器周期时间、最小内存、最大内存、缓存、最小信道、最大信道及相对性能，其中最后一个属性为目标属性。给定一台计算机，已知其前 6 个属性的取值，可以预测其相对性能的取值。回归方法是最常用的数值预测方法，下面首先重点介绍回归方法、回归树和模型树方法，最后介绍预测误差的各种度量。

5.2 回归方法

回归(regression)方法是一种历史悠久的统计方法，最早由弗朗西斯·高尔顿(Francis Galton)提出。回归方法中最常用的是线性回归(linear regression)，包括一元线性回归(又称简单线性回归)、多元线性回归(multiple linear regression) 以及非线性回归。线性回归方法不仅可以用于预测，而且可以用作解释模型，以探寻变量之间的关系。另外还有回归树(regression tree)和模型树(model tree)等模型，下面先从简单的一元线性回归开始介绍。

5.2.1 一元线性回归

一元线性回归模型涉及一个因变量(dependent variable，又称响应变量(response variable))y 和一个自变量(independent variable)x，利用如下的线性关系对两者进行建模。

$$y=\beta_0+\beta_1 x+\varepsilon \tag{5.1}$$

其中，β_0、β_1 是系数；ε 是随机变量，服从均值为 0、方差为 σ^2 的正态分布，即 $\varepsilon\sim N(0,\sigma^2)$。构建和使用此模型涉及如下几个步骤：①构建包含因变量和自变量的训练集；②通过散点图，确认因变量和自变量之间的近似线性关系；③计算系数，构建模型；④检验模型；⑤利用模型进行预测。

一元线性回归是用一条直线描述因变量和自变量之间的关系，因此可以通过已观测到的样本的散点图来发现两者之间是否大致符合线性关系。表 5.2 所示的是某公司在 10 个地区的销售额和人口数。

为了建立销售额和人口数之间的数量关系，首先绘制散点图，如图 5.1 所示。从该图中可以看到，二者之间存在线性关系。因此，下面就可以利用线性回归进行建模。

表 5.2　销售额与人口数

销售额/万元	人口数/万人	销售额/万元	人口数/万人
32	33	41	43
28	34	30	36
26	28	22	29
30	30	28	35
29	32	35	36

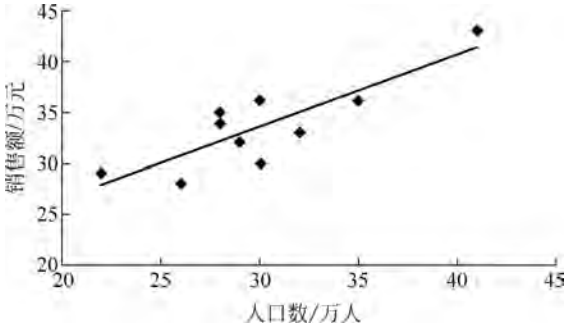

图 5.1　销售额与人口数之间的散点图

为了构建一元线性回归模型，需要根据训练集数据进行参数估计。假设训练集由 n 个样本构成：$\{(x_i,y_i)\mid i=1,2,\cdots,n\}$。用一条直线来拟合这 n 个样本，则原来的 n 个点由直线上的 n 个点来估计，即原来的点 (x_i,y_i) 变为 $(x_i,\hat{y}_i)$，$i=1,2,\cdots,n$，其中的误差 $\varepsilon_i=y_i-\hat{y}_i$ 称为残差（$\hat{y}_i=a+bx_i$，其中，a、b 为系数 β_0、β_1 的估计值，b 又称为斜率系数）。为使拟合的误差最小，采用最小二乘法，即使 n 个样本的拟合误差的平方和最小：

$$\mathrm{SS}_E=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5.2}$$

SS_E 称为**残差平方和**。为使 SS_E 最小，通过将该式对系数 a 和 b 求偏导，使其等于 0，可解得 a 和 b 的值如下：

$$\begin{aligned} b&=\frac{S_{xy}}{S_{xx}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\ a&=\bar{y}-b\bar{x}=\frac{1}{n}\sum_{i=1}^{n}y_i-b\,\frac{1}{n}\sum_{i=1}^{n}x_i \end{aligned} \tag{5.3}$$

式(5.3)中 $\bar{x}$ 和 $\bar{y}$ 分别是 n 个样本的自变量和因变量的均值，S_{xx} 称为 ***x* 的校正平方和**，S_{xy} 称为**校正交叉乘积和**。同理，$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$ 称为 ***y* 的校正平方和**。

无论因变量和自变量之间是否存在真实的线性关系，利用公式(5.3)都能得到该线性模型。因此，在构建完此模型之后，需要对模型进行各种检验。下面介绍**拟合优度检验**、**回归关系的显著性检验**和**回归系数的显著性检验**，常用的检验方法有 R 检验、F 检验和 t 检验。

首先介绍回归模型的**方差分析**。前面已经给出了残差平方和 SS_E，下面给出回归平方和 SS_R 的定义。

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{5.4}$$

总离差平方和 SS_T 是将 y 的均值作为总体估计值时的误差，定义如下。

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5.5}$$

可以证明，$SS_T = SS_E + SS_R$。可以理解为，总离差平方和中被回归模型解释的部分为回归平方和，未被回归模型解释的部分是残差平方和。那么，在总离差平方和中回归平方和占的比例越大，说明模型的拟合效果越好。因此，定义样本决定系数(determination coefficient)R^2(R square)和修正样本决定系数 $\bar{R}^2$(adjusted R square)如下。

$$\begin{aligned} R^2 &= \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \\ \bar{R}^2 &= 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2) \end{aligned} \tag{5.6}$$

公式(5.6)中 n 为样本个数，k 为自变量个数，修正样本决定系数 $\bar{R}^2$ 用于自变量个数增多的时候，它将自变量的个数加以考虑，以便不同自变量个数的回归方程可以进行比较。显然，样本决定系数 R^2 越接近 1，说明模型的拟合程度越高。

为了检验线性回归关系的显著性，需要检验假设 $H_0: b=0$ 和 $H_1: b\neq 0$。可以证明在 H_0 成立的情况下由公式(5.7)定义的 F 符合 $F(1,n-2)$分布。

$$F = \frac{SS_R}{SS_E/(n-2)} \tag{5.7}$$

给定显著性水平 α，查自由度为$(1,n-2)$的 F 分布临界值表，可得临界值 $F_\alpha(1,n-2)$ 使得概率 $P(F>F_\alpha(1,n-2))=\alpha$。然后，通过样本计算公式(5.7)得 F 值，设为 F_0，若 $F_0>F_\alpha(1,n-2)$则因变量和自变量之间的线性关系显著，假设 H_0 被拒绝。

最后介绍**回归系数的显著性检验**。为了检验回归模型中每个回归系数的显著性，可以推导出系数 a 和 b 的样本方差如下。

$$S_b^2 = \frac{SS_E/(n-2)}{S_{xx}} \tag{5.8}$$

$$S_a^2 = \frac{SS_E}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \tag{5.9}$$

可以证明 $t_b = b/S_b$ 和 $t_a = a/S_a$ 均符合自由度为$(n-2)$的 t 分布。其中重要的是检验系数 b 是否为 0。因此需要检验假设 $H_0: b=0$ 和 $H_1: b\neq 0$ 是否成立。给定显著性水平 α，查自由度为$(n-2)$的 t 分布表，得到 $t_\alpha(n-2)$，若 $t_b>t_\alpha(n-2)$，则拒绝假设 H_0，即回归系数 b 显著。通常 $t_b>2$ 时说明该系数显著，即对模型的贡献大。同时可以计算出 P 值(P value)，一般以 $P<0.05$ 为显著，$P<0.01$ 为非常显著。

5.2.2 多元线性回归

多元线性回归是对一个因变量和多个自变量之间的回归分析。设因变量 y 和 k 个自

变量 $x_1,x_2,\cdots,x_k$ 之间满足如下线性关系：

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon \tag{5.10}$$

公式(5.10)中 $\beta_i,i=1,2,\cdots,n$，称为回归系数；ε 称为残差，服从均值为 0、方差为 σ^2 的正态分布，即 $\varepsilon\sim N(0,\sigma^2)$。该式称为多元线性回归模型。

假设因变量和自变量有 n 个观测样本，$(x_{i1},x_{i2},\cdots,x_{ik},y_i),i=1,2,\cdots,n$。根据这些样本观测值，可以得到估计的回归方程为：

$$\hat{y}=b_0+b_1x_1+\cdots+b_kx_k \tag{5.11}$$

其中，$b_0,b_1,\cdots,b_k$ 是 $\beta_0,\beta_1,\cdots,\beta_k$ 的最小二乘估计，即使得公式(5.12)中残差平方和 SS_E 最小。

$$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\cdots-b_kx_{ik})^2 \tag{5.12}$$

与一元线性回归类似，为使 SS_E 最小，利用多元函数求极值的方法$\left(令\frac{\partial SS_E}{\partial b_i}=0,i=0,1,\cdots,k\right)$可以解得 $b_0,b_1,\cdots,b_k$ 的取值。求解过程这里不再赘述，下面给出结果。

为方便表述，n 个观测样本$(x_{i1},x_{i2},\cdots,x_{ik},y_i)$及回归系数用如下矩阵表示。

$$\boldsymbol{Y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}\quad\boldsymbol{X}=\begin{bmatrix}1&x_{11}&x_{21}&\cdots&x_{k1}\\1&x_{12}&x_{22}&\cdots&x_{k2}\\\vdots&\vdots&\vdots&&\vdots\\1&x_{1n}&x_{2n}&\cdots&x_{kn}\end{bmatrix}\quad\boldsymbol{B}=\begin{bmatrix}b_0\\b_1\\\vdots\\b_k\end{bmatrix} \tag{5.13}$$

则回归系数可以通过矩阵运算 $\boldsymbol{B}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 得到。

建立了多元回归模型后，同样要进行模型的检验，包括拟合优度检验、回归模型的显著性检验和回归系数的显著性检验。

拟合优度检验仍然使用样本决定系数 R^2 和修正样本决定系数 $\bar{R}^2$，其公式同公式(5.6)，其中，残差平方和见公式(5.12)，回归平方和与总离差平方和的公式同公式(5.4)、(5.5)，$SS_T=SS_E+SS_R$。因变量 y 的自由度(degree of freedom)df_T 也可以分解为回归自由度 df_R 和残差自由度 df_E，即 $df_T=df_R+df_E$，其中 $df_T=n-1,df_R=k,df_E=n-k-1$，$n$ 为样本个数，k 为自变量个数。R^2 越接近 1，模型的拟合效果越好。将三种方差(即 SS_T、SS_E、SS_R)分别除以相应的自由度，得到的是相应的均方差，即 $MS_T=SS_T/df_T$，$MS_E=SS_E/df_E$，$MS_R=SS_R/df_R$。

回归模型的显著性检验用于检验因变量与自变量整体之间是否存在线性关系，仍然采用 F 检验。显著性检验的无效假设和备择假设分别为 $H_0:b_1=b_2=\cdots=b_k=0$ 和 $H_1:b_1,b_2,\cdots,b_k$ 不全为零。可以证明，在 H_0 成立的情况下，由公式(5.14)定义的 F 变量符合 $F(k,n-k-1)$分布。

$$F=\frac{MS_R}{MS_E} \tag{5.14}$$

给定显著性水平 α，查自由度为$(k,n-k-1)$的 F 分布临界值表，可得临界值 $F_\alpha(k,n-k-1)$使得概率 $P(F>F_\alpha(k,n-k-1))=\alpha$。然后，通过样本计算公式(5.14)得 F 值，设为 F_0，若 $F_0>F_\alpha(n-k-1))$，则因变量和自变量之间的线性关系显著，假设 H_0 被拒绝。

表 5.3 是回归分析结果中通常返回的方差分析表的构成示例。

表 5.3　回归分析结果中通常返回的方差分析表的构成示例

方差类型	自由度	平方和	均方差	F
回归	k	SS_R	MS_R	MS_R/MS_E
残差	$n-k-1$	SS_E	MS_E	
总离差	$n-1$	SS_T		

回归系数的显著性检验。前面所述的 F 检验如果说明回归关系是显著的，并不能说明每个回归系数是显著的，有可能存在某些回归系数不显著的情况。因此，需要接着对单个回归系数分别进行检验，不显著的系数可以去掉，重新建立回归模型。

回归系数的显著性检验可以采用 t 检验。对于每个回归系数 $b_i(i=1,2,\cdots,k)$，显著性检验的两个假设分别为 $H_0: b_i=0$ 和 $H_1: b_i\neq 0$。若 $b_i=0$ 说明自变量 x_i 的变化对因变量没有线性影响，即变量 x_i 对因变量的影响不显著。为每个回归系数 b_i 构造变量 t_{b_i}：

$$t_{b_i}=\frac{b_i}{S_{b_i}}=\frac{b_i}{\sqrt{c_{ii}}\sqrt{MS_E}} \tag{5.15}$$

其中，c_{ii} 是矩阵 $\boldsymbol{C}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 的对角线上的第 i 个值。

给定显著性水平 α，查自由度为 $(n-k-1)$ 的 t 分布表，得到 $t_\alpha(n-k-1)$，若 $t_{b_i}>t_\alpha(n-k-1)$，则拒绝假设 H_0，即回归系数 b_i 显著。

在通过以上显著性检验之后，给定一个因变量未知的样本 $(x_{t1},x_{t2},\cdots,x_{tk})$ 将其带入公式(5.11)中的各个自变量取值中就可以得到因变量的一个预测值。

5.2.3　非线性回归

实际应用中并不是所有的因变量和自变量之间都存在线性关系，有时存在非线性关系，如图 5.2 所示。

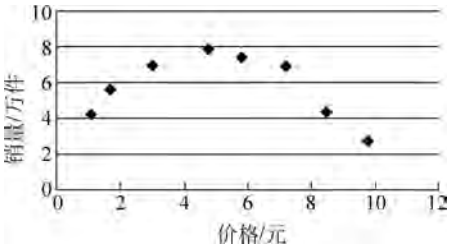

图 5.2　销售量与商品价格之间的散点图

图 5.2 中所示的是某商品在某段时间内的价格 x 和销量 y 之间的关系图，显然这两个变量之间存在的不是线性关系。有些非线性关系通过一定的变换可以转换为线性回归问题。例如，图 5.2 中所示的两个变量之间的关系为 $y=a+bx^2$。此时，可以假设 $x_1=x^2$，则原来的非线性关系变为 $y=a+bx_1$。因此，对于形如 $y=a_0+a_1x+a_2x^2+\cdots+a_px^p$ 的因变量和自变量之间的关系，假设 $x_1=x, x_2=x^2,\cdots,x_p=x^p$，则有 $y=a_0+a_1x_1+a_2x_2+\cdots+a_px_p$。经过这样的变换之后，就可以利用多元线性回归进行建模了。

一些常用的非线性函数除了前面提到的多项式函数外，还包括幂函数、指数函数、对数函数及双曲函数等。

对于幂函数 $y=ax^b$，可以通过两边取对数变换为 $\lg y=\lg a+b\lg x$，设 $Y=\lg y, a_0=\lg a, a_1=b, X=\lg x$，则有 $Y=a_0+a_1X$。

对于指数函数 $y=a\mathrm{e}^{bx}$，可以通过两边取对数变换为 $\ln y=\ln a+bx$，设 $Y=\ln y, a_0=\ln a, a_1=b$，则有 $Y=a_0+a_1x$。

对于对数函数 $y=a+b\lg x$，设 $X=\lg x$，则有 $y=a+bX$。

对于双曲函数 $y=x/(ax+b)$，可以通过两边取倒数变换为 $1/y=a+b/x$，设 $Y=1/y$，$X=1/x$，则有 $Y=a+bX$。

另外，非线性关系也可以通过构建回归树（regression tree）或模型树（model tree）的方法进行建模。

5.3 回归树与模型树

回归树和模型树与第 4 章中介绍的决策树存在许多相似之处，都是通过自顶向下分而治之的思想，将训练集不断分割成子数据集来不断扩展树枝，当满足一定条件时停止树的生长。不同之处包括数据集的分割条件不同和叶子结点的内容不同。决策树的叶子结点对应某个类别，而回归树的叶子结点对应一个数值，模型树的叶子结点对应一个线性回归方程。

图 5.3 所示的模型树是利用 Weka 中的 M5P 模型树算法对 CPU 数据集（此数据集是 UCI 机器学习数据库中的一个，网址为 http://archive.ics.uci.edu/ml/datasets/Computer+Hardware）构建的。表 5.1 是该数据集的部分数据。

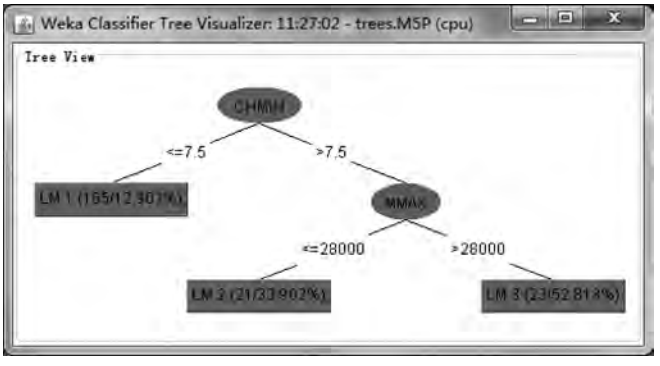

图 5.3 CPU 数据集的模型树

此模型树中有 2 个内部结点，3 个叶子结点。叶子结点对应的回归方程以及相应的条件如下：

```
If CHMIN <= 7.5, then
PRP = -0.0055 * MYCT + 0.0013 * MMIN + 0.0029 * MMAX + 0.8007 * CACH + 0.4015 * CHMAX +
11.0971
if CHMIN > 7.5 and MMAX <= 28000, then
PRP = -1.1492 * MYCT + 0.0086 * MMIN + 0.0031 * MMAX + 0.8422 * CACH - 4.0839 * CHMIN +
1.1597 * CHMAX + 101.3434
if CHMIN > 7.5 and MMAX > 28000, then
PRP = -0.4882 * MYCT + 0.0218 * MMIN + 0.003 * MMAX + 0.3865 * CACH + 3.2333 * CHMAX -
67.9242
```

通常模型树比回归树小，更适合处理大规模数据，且回归树可以看作模型树的一个特例。下面重点介绍模型树的构建方法。为了避免过分拟合问题，模型树构建之后也需要进行剪枝。

5.3.1 模型树的构建

给定训练数据集，构建模型树的过程与构建决策树的过程非常类似，也是一个将训练集不断分裂的过程。假设训练数据集用 D 表示，涉及的类别由 $C=\{c_1,c_2,\cdots,c_k\}$ 表示。构建决策树 T 的主要步骤如下。

(1) 创建一个结点 t，与结点 t 关联的数据集记为 D_t。初始情况下训练数据集中的所有样本与根结点关联，即 $D_t=D$。将 t 设为当前结点。

(2) 如果当前结点 t 所关联的数据集 D_t 中样本个数小于给定阈值或者 D_t 中样本的目标属性取值的标准差小于给定阈值(例如初始数据集 D 的标准差的 5%)，则将该结点标记为叶子结点，停止对该结点所关联的数据集的进一步分裂，对数据集 D_t 运用多元线性回归建模方法构建回归模型。否则，进入下一步。

(3) 为数据集 D_t 选择分裂属性和分裂条件。根据分裂条件将数据集 D_t 分裂为两个子数据集，为结点 t 创建两个子女结点，将这两个子数据集分别与之关联。依次将每个结点设为当前结点，转至步骤(2)进行处理，直至所有结点都标记为叶子结点。

上述过程中的一个关键点在于步骤(2)中的分裂属性和分裂条件的选择。分裂属性的选择以分裂后的各个子数据集中目标属性取值的标准差为依据，将标准差作为一种误差度量，将分裂前后标准差的减少量作为误差的期望减少，称为 SDR(standard deviation reduction)。假设数据集 D 按照属性 A 的取值分裂为两个子数据集 D_1 和 D_2，此次分裂的 SDR 值的计算公式如下：

$$\mathrm{SDR}(D,A)=\mathrm{sd}(D)-\sum_{i=1}^{2}\frac{|D_i|}{|D|}\times\mathrm{sd}(D_i) \tag{5.16}$$

其中，$\mathrm{sd}(D)$代表数据集 D 中目标属性取值的标准差，$|D|$代表数据集 D 中包含的样本个数。

在选择分裂属性时，选取使 SDR 值最大的属性。

按照属性 A 的取值分裂数据集的方法取决于属性 A 的类型。如果 A 是连续取值的属性，则将 A 的所有取值升序排列，每两个相邻的取值的中点可以作为一个候选的分裂点，中点假设用 v_m 表示，分裂条件则为 $A\leqslant v_m$ 和 $A>v_m$。计算每个候选分裂点的 SDR 值，选取具有最大值的分裂点作为该属性的分裂条件，与其他属性进行比较。

如果 A 是离散变量或定性属性，假设 A 属性有 k 个不同取值$\{v_1,v_2,\cdots,v_k\}$，则可以将此属性进行如下处理：对于 A 的每个取值 v_i，求出对应的目标属性的平均值 $\mu(v_i)$，然后将这 k 个不同取值按照目标属性的平均值进行升序排序，设其顺序为 $v_{l_1},v_{l_2},\cdots,v_{l_k}$，其中 $l_i\in\{1,2,\cdots,k\}$。则有$(k-1)$组不同的分裂条件，分别为 $A\in\{v_{l_1}\}$和 $A\in\{v_{l_2},\cdots,v_{l_k}\}$，$A\in\{v_{l_1},v_{l_2}\}$和 $A\in\{v_{l_3},\cdots,v_{l_k}\}$，$A\in\{v_{l_1},v_{l_2},v_{l_3}\}$和 $A\in\{v_{l_4},\cdots,v_{l_k}\}$，$\cdots$，$A\in\{v_{l_1},v_{l_2},\cdots,v_{l_{k-1}}\}$和 $A\in\{v_{l_k}\}$。分别计算这 $k-1$ 组不同的分裂条件对应的 SDR 值，选择 SDR 值最大的作为该属性的分裂条件。

例如，假设银行储户的婚姻状况是属性 A，目标属性 Y 是每月平均账户余额，表 5.4 是一个有关这两个属性的示例数据。

表 5.4　定性属性数据集 D 示例

婚姻状况	账户余额/万元	婚姻状况	账户余额/万元
单身	20	已婚	200
单身	40	已婚	130
单身	90	离异	60
已婚	30	离异	100

属性婚姻状况有 3 个不同取值，其中单身对应的账户余额的平均值 μ（单身）=（20+40+90）/3=50，同理，μ（已婚）=120，μ（离异）=80。因此，3 个取值的排序为单身、离异、已婚。对应两组分裂条件，分裂条件 A_1：婚姻状况∈{单身}和婚姻状况∈{离异，已婚}，分裂条件 A_2：婚姻状况∈{单身，离异}和婚姻状况∈{已婚}。

根据分裂条件 A_1，表 5.4 可以分裂为 2 个子表，前 3 行婚姻状况为单身的构成一个数据集 D_1，剩下的 5 行为另一数据集 D_2。$\mathrm{sd}(D_1)=29.4$，$\mathrm{sd}(D_2)=58.9$，$\mathrm{sd}(D)=56.3$。则：

$$\mathrm{SDR}(D,A_1)=\mathrm{sd}(D)-\frac{3}{8}\mathrm{sd}(D_1)-\frac{5}{8}\times\mathrm{sd}(D_2)=8.47$$

$$\mathrm{SDR}(D,A_2)=11.4$$

因此按照分裂条件 A_2 进行分裂更优，故 $\mathrm{SDR}(D,A)=11.4$。

5.3.2　模型树的剪枝

模型树构建之后，为了避免过度拟合，需要对模型树进行剪枝。剪枝通过对树深度优先遍历从叶子结点向根结点进行。以图 5.3 中的模型树为例，首先查看内部结点 MMAX 对应的子树是否需要用一个叶子结点代替。方法是计算该结点以及其下的两个叶子结点的期望误差。给定结点 t 及所关联的数据集 D_t，设样本个数为 n，数据集 D_t 对应的多元线性回归模型为 M_t，M_t 中涉及的自变量的个数为 v，设利用该模型，D_t 中每个样本的目标属性的预测值为 p_i、真值为 a_i，其期望误差 error(t)计算如下：

$$\mathrm{error}(t)=\frac{n+v}{n-v}\frac{1}{n}\sum_{i=1}^{n}|p_i-a_i| \tag{5.17}$$

两个叶子结点的期望误差通过加权求和结合在一起作为子树误差，权值是叶子结点包含样本占其父结点样本个数的比例。若当前结点含有 n 个样本，两个叶子结点含有样本分别为 n_1 和 $n-n_1$，则其权重分别为 n_1/n 和 $(n-n_1)/n$。若当前结点的期望误差小于子树误差，则将该结点设为叶子结点，即此子树被一个叶子结点代替。

5.3.3　算法

构造模型树的主要步骤如下所示。

算法 5.1：模型树构建算法 gen_modelTree(D)
输入：训练数据集 D
输出：模型树

主要步骤：

(1) **if** dataset D meets stopping criteria **then**
(2) create node t;
(3) t.type=leaf;
(4) t.data=D
(5) t.model=linearRegression(t);
(6) **else**
(7) create node t;
(8) t.type=interior;
(9) t.split_condition=find_split_condition(D);
(10) split dataset D into two subsets,D_1 and D_2
(11) t.leftChild=gen_modelTree(D_1),t.leftChild.data=D_1;
(12) t.rightChild=gen_modelTree(D_2),t.rightChild.data=D_2;
(13) **end if**
(14) return t;

模型树的剪枝的重要步骤如下：

算法 5.2：模型树剪枝算法 prune (node t)
输入：模型树
输出：剪枝后的模型树
主要步骤：

(1) **if** t.type=interior **then**
(2) prune(t.leftChild);
(3) prune(t.rightChild);
(4) t.model=linearRegression(t);
(5) **if** treeError(t)>error(t) then
(6) t.type=leaf;
(7) **end if**
(8) **end if**

过程 treeError(node t)
主要步骤：

(1) **if** t.type=interior **then**
(2) l=t.leftChild;
(3) r=t.rightChild;

(4)　return (|l.data|×treeError(l)+|r.data|×treeError(r))/|t.data|;
(5) **else** return error(t);
(6) **end if**

在过程 treeError(node t)中，|l.data|指的是与结点 l 关联的数据集包含的样本个数。error(t)是根据公式(5.17)计算的结点 t 的期望误差。调用该过程时利用 prune(root)，即从根结点调用即可。

5.4 K 近邻数值预测

与第 4 章中介绍的 K 近邻分类类似，可以利用一个样本的 K 个最相似的邻居的目标属性的取值来进行预测。

假设训练集 D 由 n 个观测样本构成：$\{o_i=(x_{i1},x_{i2},\cdots,x_{iK},y_i),i=1,2,\cdots,n\}$，其中 y_i 是目标属性 Y 的取值，$(x_{i1},x_{i2},\cdots,x_{iK})$是 K 个描述属性的取值。对于测试集 T 中的一个测试样本 $t_j=(x_{j1},x_{j2},\cdots,x_{jK},y_j)$，$j>n$，可以利用相似度衡量方法计算此样本与 D 中每个观测样本的相似度，选取与测试样本最相似的 K 个观测样本。例如，可以通过欧氏距离，计算此测试样本与第 i 个观测样本的距离，$d(o_i,t_j)=\sqrt{\sum_{l=1}^{k}(x_{il}-x_{jl})^2}$，找到距离最近的 K 个样本。设 $N(t_j)$ 是这 K 个观测样本的集合，则测试样本 t_j 的目标属性的预测值 p_j 计算如下：

$$p_j=\frac{\sum_{o\in N(t_j)}\mathrm{sim}(o,t_j)\times y(o)}{\sum_{o\in N(t_j)}\mathrm{sim}(o,t_j)} \tag{5.18}$$

式中，$y(o)$代表观测样本 o 的目标属性取值，$\mathrm{sim}(o,t_j)$代表观测样本 o 和测试样本 t_j 直接的相似度，相似度的衡量方法在第 6 章详述，此处可以利用欧氏距离的倒数进行衡量，即 $\mathrm{sim}(o,t_j)=1/d(o,t_j)$。

5.5 预测误差的度量

预测性能的优劣需要一定的度量来衡量。常用的度量是平均绝对误差(mean absolute error，MAE)、均方误差(mean square error，MSE)、均方根误差(root mean square error，RMSE)、相对平方误差(relative square error，RSE)和相对绝对误差(relative absolute error，RAE)。

假设训练集 D 由 n 个观测样本构成：$\{o_i=(x_{i1},x_{i2},\cdots,x_{ik},y_i),i=1,2,\cdots,n\}$，其中 y_i 是目标属性 Y 的取值，$(x_{i1},x_{i2},\cdots,x_{ik})$是 k 个描述属性的取值。假设测试集 T 包含 m 个样本，对于测试集 T 中的每个测试样本 $t_j=(x_{j1},x_{j2},\cdots,x_{jk},y_j)$，$n<j<n+m$，利用预测模型得出的目标属性的预测值为 p_j，则平均绝对误差 MAE 的计算公式如下：

$$\mathrm{MAE}=\frac{1}{m}\sum_{j=n+1}^{n+m}|p_j-y_j| \tag{5.19}$$

均方误差、均方根误差、相对平方误差和相对绝对误差的计算公式分别如下：

$$\mathrm{MSE}=\frac{1}{m}\sum_{j=n+1}^{n+m}(p_j-y_j)^2 \tag{5.20}$$

$$\mathrm{RMAE}=\sum_{j=n+1}^{n+m}\sqrt{\frac{(p_j-y_j)^2}{m}} \tag{5.21}$$

$$\mathrm{RSE}=\frac{\sum_{j=n+1}^{n+m}(p_j-y_j)^2}{\sum_{j=n+1}^{n+m}(y_j-\bar{y})^2},\text{其中}\bar{y}=\frac{1}{m}\sum_{j=n+1}^{n+m}y_j \tag{5.22}$$

$$\mathrm{RAE}=\frac{\sum_{j=n+1}^{n+m}|p_j-y_j|}{\sum_{j=n+1}^{n+m}|y_j-\bar{y}|},\text{其中}\bar{y}=\frac{1}{m}\sum_{j=n+1}^{n+m}y_j \tag{5.23}$$

练习题 5

1. 数值预测和分类的异同点有哪些？
2. 用线性回归建模的基本步骤有哪些？
3. 在线性回归中样本决定系数 R^2 说明了什么？
4. 如何进行回归关系的显著性检验？
5. 如何进行回归系数的显著性检验？
6. 表 5.5 中给出了 10 个学生的身高和体重的数据。

(1) 绘制散点图。

(2) 求出回归方程。

(3) 对回归模型进行统计检验。

(4) 一个身高 1.66m 的人的体重可预测为多少？

表 5.5　习题 6 数据

身高/m	体重/kg	身高/m	体重/kg
1.62	55	1.68	62
1.65	57	1.75	60
1.60	45	1.80	90
1.72	65	1.76	70
1.73	70	1.82	75

7. 根据 10 年的年度统计资料，利用多元回归对同一因变量构建了两个回归方程。第一个方程中 $k=5$，$R^2=0.83$；第二方程中 $k=1$，$R^2=0.80$。试对这两个回归方程的拟合程度做出评价。

8. 表 5.6 是来自中国统计年鉴的 1990—2003 年我国的城镇居民家庭人均可支配收入与城市人均住宅建筑面积的数据。

表 5.6　习题 8 数据

年　　度	城镇居民家庭人均可支配收入/元	城市人均住宅建筑面积/m^2
1990	1510.2	13.65
1991	1700.6	14.17
1992	2026.6	14.79
1993	2577.4	15.23
1994	3496.2	15.69
1995	4283	16.29
1996	4838.9	17.03
1997	5160.3	17.78
1998	5425.1	18.66
1999	5854.02	19.42
2000	6280	20.25

（1）做出散点图，建立住宅建筑面积为因变量的一元线性回归模型，并解释斜率系数的经济意义。

（2）对回归模型进行统计检验。

9. 图 5.4 是利用 Excel 进行多元回归建模的部分输出结果。

（1）计算①②③④⑤中的值。

（2）判断回归方程以及回归系数的显著性。

回归统计

R^2	①
调整的t^2	②
标准误差	2.010050279
观测值	10

方差分析

	df	SS	MS	F	显著性水平F
回归分析	2	423.0178851	⑤	52.34978375	6.16117E-05
残差	③	④	4.040302126		
总计	9	451.3			

	系数	标准误差	t值	P值	下 95%	上 95%
截距	-38.82516938	8.478591118	-4.579200582	0.00254617	-58.87383722	-18.77650155
X_1	1.340693618	0.143315893	9.354814676	3.31495E-05	1.001805625	1.679581612
X_2	0.022802293	0.004754224	4.796217239	0.001974896	0.011560347	0.03404424

图 5.4　习题 9 回归模型结果