第二章 统计学习

2018年11月29日 15:42

2.1基本概念

1.X:输入变量(input variable)、预测(predictor)变量、自(independent) 变量、属性(feature) 变量，变量(variable)
Y:输出变量(output variable)、响应(response)变量、因变量(dependent variable)

2.一个定量的响应变量Y 和p 个不同的预测变量，记为X_1，X_2,.....X_p。假设这个Y 和X=(X_1，X_2,.....X_p）有一定的关系，可以表达成一个比较一般的形式:
Y=f(X)+ε
这里的f是X_1，X_2,.....X_p的函数，它是固定的但未知，ε是随机误差项(error term)，与X独立，且均值为0 。

3.训练集（training data）：x_{ij}表示观测点i 的第j 个预测变量，i = 1 , 2,... , n 和j = 1 , 2 ， ... ， p。令y_i 表示第i 观测点的响应变量值，训练集记做{(x_{11},x_{12},x_{13},...,x_{1p},y_1),(x_{21},x_{22},x_{23},...x_{2p},y_2),...,(x_{n1},x_{n2},x_{n3},...,x_{np},y_n)}， $x_i=(x_{i1},x_{i2},x_{i3},...,x_{ip})^T$

4.Y的预测精确性依赖于可约误差（reducible error）和不可约误差（irreducible error）。
（1）不可约误差来源（ε）：对预测Y 起作用但却不可直接观测的变量信息。例如，某个病人不良反应的风险也许会在一天内很不一样，比如药物的药效本身在设计上随着一天内服药环境的温度和湿度的不同而不同，或者风险与病人当天服药的情绪状态有关等。
（2）主要关注使可约误差最小，不可约误差是预测精度的上届，这个上界是未知的。

5.统计学习问题分类——监督学习、无监督学习、半监督学习
半监督学习：假设有n 个观测。其中m（m< n）个观测点，可以
同时观测到预测变量和响应变量。而对于其余n-m 个观测点，只能观测到预测变量但无法观测到响应变量。比如对预测变量的采集相对简单，而相应的响应变量却比较难采集，我们称这种问题为半指导学习(semi supervised learning)问题。在这种情形里，我们希望能够有一种统计学习方法既用到m 个观测点的预测变量和响应变量的信息，同时又包含了n-m 个不能获取响应变量观测值的信息。

6.变量分类——定性和定量。
当响应变量是定量时，通常选用线性回归模型，当响应变量是定性变量时，用逻辑斯谛回归。而预测变量是定性的还是定量的，通常对选择模型并不十分重要。如果在分析之前，所有定性变量的取值都已正确编码，无论预测变量是什么类型，本书讨论的大部分统计学习方法大都能够应用。

2.1.1分析的目标（为什么要估计f(X)，需求）——预测（prediction）和推断（inference）

1.区别
预测：在现有数据基础上预测结果，对每个特征对结果的影响不关系。
推断：关系每个特征变化时对结果会有什么影响。

2.简单模型比如线性模型适用于简单和解释性的推断，但预测的精确性可能不够；复杂模型比如高度非线性方法能提供更加精确的预测，但解释性可能不够清晰。

2.1.2分析的方法（估计f(X)的方法）——参数方法和非参数方法

1.参数方法：假设f具有一定的形式或形状，比如假设为线性模型。缺点：选择的模型未必与真正的f一致，选择复杂模型则容易出现过拟合（overfitting）（拟合了错误或噪声（noise））
2.非参数方法：不限定函数的具体形式，追求接近数据点，比如KNN方法。缺点：无法将估计f的问题简化到仅仅对

少数参数进行估计的问题，往往需要大量的观测点(远远超出参数方法所需要的点)。

2.2评价模型精度

2.2.1回归模型的精度

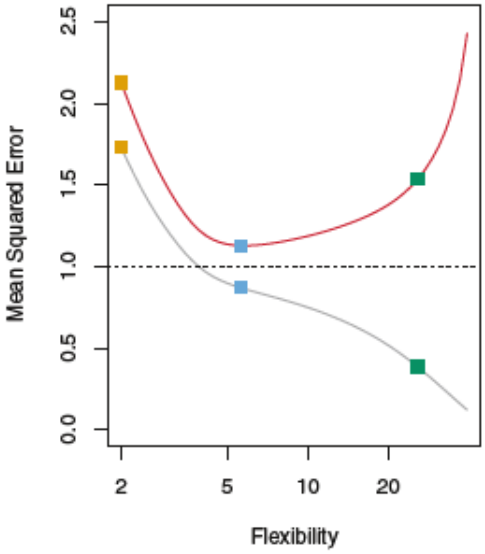
1.拟合效果检验

在回归中，最常用的评价准则是均方误差(mean squared error , MSE)，

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

其中$\hat{f}(x_i)$是第i个观测点上应用$\hat{f}$的预测值。

注意训练均方误差（Training MSE）和测试均方误差（test MSE）：训练均方误差降低时测试均方误差不一定降低，较小的训练均方误差可能会有较大的测试均方误差（过拟合）。



红色：测试均方误差

蓝色：训练均方误差

虚线：不可约误差ε

2.偏差-方差的权衡

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

$$E\left(y_0 - \hat{f}(x_0)\right)^2$$

为模型的期望测试均方误差（excepted test MSE）

一般而言，光滑度更高（更精密复杂）的方法，偏差减小，方差增加。

2.2.2分类模型的精度

1.计算训练错误率（error rate）

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

其中$\hat{y}_i$是使用$\hat{f}$预测数据的第i个值。$I(y_i \neq \hat{y}_i)$表示一个**示性变量**（indicator variable），当$y_i \neq \hat{y}_i$时，值等于1，当$y_i = \hat{y}_i$时，值等于0。如果$I(y_i \neq \hat{y}_i) = 0$，那么第$i$个观测值用分类模型实现了正确的分类，否则，它被误分了。因此，等式（2.8）计算了误分类的比例。

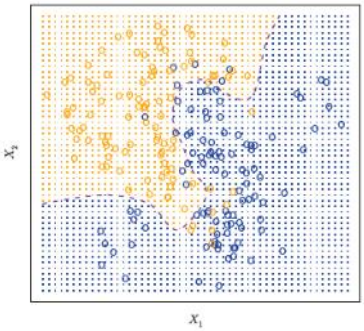
与回归同理，有测试错误率（test error rate）

2.3常见分类模型——贝叶斯和KNN

2.3.1贝叶斯分类器——知道所有数据集的真实情况，即知道这个数据集所有数据的真实可能性，画出来的是真实情况。

1.思路

比如在一个二分类问题中，一个称为类别1，另一个称为类别2. 如果$Pr(Y = 1 \mid X = X_0) > 0.5$，贝叶斯分类器将该观测的类别预测为1，否则预测为类别2。概率为50%的线称为贝叶斯决策边界。



贝叶斯分类器产生的是最低的测试错误率，称为贝叶斯错误率。类似于不可约误差。

2.局限

实际很难知道给定X后Y的真实分布，所以贝叶斯是一个难以达到的黄金标准。（KNN方法）

2.3.2KNN方法——利用测试点最近的点的状态判定该测试点的状态。

类似的，有KNN决策边界。

K较小，偏差低方差高，光滑度高，柔性强，训练错误率小，测试错误率高（过拟合）；
K较大，偏差高方差低，光滑度低，决策边界接近线性，方差小偏差高。K要适当。

第三章 线性回归

2018年12月1日 14:52

1.Var（方差，variance）：将各个误差平方相加再除以总数（真实的方差，总体方差）

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

2.SD（标准差，standard error）：方差开算术平方根，σ

3.DF（自由度，Degrees of Freedom）：在多元回归模型分析中，指观测值的个数减去待估参数的个数。 统计量的自由度是指可自由变化的样本观测值的个数，它等于样本观测值的个数减去对样本观测值的约束个数。参见https://zh.wikipedia.org/wiki/%E8%87%AA%E7%94%B1%E5%BA%A6_(%E7%BB%9F%E8%AE%A1%E5%AD%A6)

4.RSS（残差平方和/误差平方和，residual sum of squares）

$$e_i = y_i - \hat{y}_i$$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

5.RSE（残差标准误，就是样本方差，redusial standard error），对总体方差（真实方差）σ²的估计。

$$\mathrm{RSE} = \sqrt{\frac{RSS}{DF}}$$

6.SE（标准误差/标准误，standard error）

标准误和标准差的区别：

标准差 = **一次抽样**中**个体分数**间的离散程度，反映了**个体分数对样本均值的代表性**，用于描述统计

标准误 = **多次抽样**中**样本均值**间的离散程度，反映了**样本均值对总体均值的代表性**，用于推论统计

来自 <https://www.zhihu.com/question/22864111>

标准差是反映个体变量对其统计量的代表性，标准误是反映计算的样本统计量对真实统计量（不可能知道的）的代表性。

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

均跟σ²（RSE）有关

7.TSS（总平方和，total sum of squares）

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

3.1简单线性回归

$$Y \approx \beta_0 + \beta_1 X$$

β_0、β_1被称为模型的系数(coefficient) 或参数(parameter) 。

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

来预测未来的销量。其中，$\hat{y}$ 表示在 $X = x$ 的基础上对 Y 的预测。这里我们用帽子符号“^”表示对一个未知的参数或系数的估计值，或表示响应变量的预测值。

假设要用简单线性回归模型

3.1.1首先，判断X=(X_1，X_2,……X_p）与Y有没有关系——假设检验：

对零假设(null hypothesis) :H_0：X 和Y 之间没有关系。

备择假设(alternative hypothesis)： Hα :X 和Y 之间有一定的关系。

对于简单线性回归，相当于检验

$$H_0 : \beta_1 = 0 \text{ 和 } H_a : \beta_1 \neq 0$$

通过计算t统计量确定，它测量了β_1^偏离0的标准偏差。

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

相应的有p值，p 值足够小，便可拒绝零假设(reject the null hypothesis)，也就是声明X 和Y 之间存在关系。典型的拒绝零假设的临界p 值是5% 或1% 。

对表3-1的解释：标准误差小，t统计量大，说明β离0很远，所以β为0的概率（p值）很小，说明H_0为假，X和Y存在关系。

3.1.2确定X、Y存在关系后，评价简单线性回归模型准确性：

判断线性回归的拟合质量常用残差标准误(residual standard error , RSE) 和R^2 统计量。两个值是多少可接受视具体情况而定。

RSE：对不可约误差ε的标准偏差的估计，是响应值会偏离真正的回归直线的平均量（即使模型和参数都正确）。

R^2：测量的是Y 的误差中能被X 解释的部分所占的比例。

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

越接近1说明回归可以解释Y的大部分误差，接近0说明回归没有解释太多Y的误差，这可能因为线性模型是错误的，或者不可约误差较大，抑或两者兼有。

3.1.3确定所选模型是正确的后，估计其参数：最小二乘法

方法是残差平方和最小化原则

通过微积分运算，使RSS 最小的参数估计值为:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

（代价函数为RSS，通过正规方程法确定β_0、β_1）

3.1.4确定参数后，评估估计值的准确性：

真实总体回归直线：

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

由同一真实模型产生的不同的数据集对应的最小二乘线略有不同，但是未观察到的总体回归直线是不变的。

样本均值和总体均值的含义是不同的，但一般来说，样本均值能提供对总体均值的良好估计。样本β_0、β_1值与总体β_0、β_1值同理。

无偏性：样本估计值与总体真实值相差在可以接受的范围内就可以认为样本估计值等于总体真实值，即具有无偏性。无偏性通俗解释可见 https://www.zhihu.com/question/22983179/answer/404391738马同学回答。

衡量无偏性（衡量估计值距离真值有多远）：计算标准误差（standard error，SE）

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

其中，σ 是变量 Y 的每个实现值 y_i 的标准差[⊖]。粗略地说，标准误差告诉我们估计 $\hat{\mu}$ 偏离 μ 的实际值的平均量。公式（3.7）也告诉我们这种偏差随着 n 的增大而减小——我们有的观测越多，$\hat{\mu}$ 的标准误差越小。同样，我们可以探究 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 与真实值 β_0 和 β_1 的接近程度。要计算 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准误差，可以使用下列公式：

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (3.8)$$

有了标准误差，就可以得到置信区间。对于简单线性回归模型，β_1的95% 置信区间（表示这个区间约有95% 的可能会包含β_1的真实值）约为:

$$\left[\hat{\beta}_1 - 2 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_1)\right]$$

至此，得到结论。

3.2多元线性回归

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

假设要用多元回归线性模型

3.2.1首先，判断X=(X_1，X_2,……X_p）与Y有没有关系——假设检验：

对零假设：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

备择假设：

$$H_a : \text{至少有一个} \beta_j \text{不为} 0$$

通过计算F统计量确定：

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)},$$

F>(>>)1结合计算出的p值，由此拒绝H_0，判定Hα为真，即X与Y有关系。

既已得到各个变量的p 值，为什么还需要看整体的F统计量呢?
即使预测变量与响应变量之间没有任何关联，仍能看到个别小的 p 值，会产生误导，所以要结合F统计量。

注意：当变量数非常大，p>n时，不能用最小二乘法拟合多元模型，F统计量也无法使用。

3.2.2确定X与Y存在关系后，确定是哪些重要X与Y存在关系（简单线性回归直接看p值，多元线性回归中如果预测变量数p很大，看p值可能就会得出错误的结论）：
通过变量选择：向前选择、向后选择、混合选择
注意：p>n时，不能使用向后选择，而向前选择在各种情况下都能使用。向前选择可能在前期将后来变得多余的变量纳入模型。混合选择可以修正这个问题。

3.2.3选定重要变量后，评价多元线性回归模型准确性：
采用RSE和R^2，解释同简单线性回归模型。

注意：添加变量必然会更准确地拟合训练数据(尽管对测试数据未必如此)。因此，根据训练数据计算出的R^2统计量也必然增加。但某个变量加入后如果只有极小的增加，说明这一变量加入改善不大，可以删除。从本质上讲，它没有真正地改善模型对训练样本的拟合，将其纳入模型很
可能导致模型出现过拟合，从而在独立测试样本上效果不佳。

3.2.4确定所选模型是正确的后，估计其参数：最小二乘法，软件包都可以计算。

3.2.5确定参数后，评估估计值的准确性：
与简单线性回归同理，最小二次平面

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

只是对真实总体回归平面

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

的模拟，有可约误差和不可约误差，可用预测区间。

预测区间总是比置信区间宽，因为预测区间既包含f(X) 的估计误差(可约误差)也包含单个点偏离总体回归平面程度的不确定性(不可约误差)。

此外线性模型假设是可约误差的一种来源，称为模型误差，此处忽略，假设线性模型是正确的。

3.3回归模型中的其他注意事项

3.3.1定性预测变量

方法：对定性变量编码哑变量（dummy variable），比如：

哑变量。第一个哑变量是：

$$x_{i1} = \begin{cases} 1 & \text{亚洲人} \\ 0 & \text{非亚洲人} \end{cases} \tag{3.28}$$

第二个哑变量是：

$$x_{i2} = \begin{cases} 1 & \text{白种人} \\ 0 & \text{非白种人} \end{cases} \tag{3.29}$$

这两个变量都可以用于回归方程中，得如下模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{亚洲人} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{白种人} \\ \beta_0 + \varepsilon_i & \text{非裔美国人} \end{cases} \tag{3.30}$$

现在 β_0 可以解释为非裔美国人的平均信用卡债务，β_1 可以解释为亚洲人和非裔美国人的平均信用卡债务差异，β_2 可解释为白种人和非裔美国人的平均信用卡债务差异。哑变量个数总是比水平数少 1。没有相对应的哑变量的水平——本例中的非裔美国人——被称为**基准水平**（baseline）。

无论使用哪种编码方式，最后的预测结果是不变的。不同编码方式的唯一区别在于对系数的解释不同。

3.3.2对线性模型假设的扩展

线性模型两个重要假设：可加性和线性

1.去除可加性（预测变量X的变化对响应变量Y 产生的影响与其他预测变量的取值无关）——考虑交互作用（interaction）

方法：加入交互项（interaction term），交互项由变量之间的乘积组成。比如原模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

若X_1和X_2好像有关系，可以加入$\beta_3 X_1 X_2$的项，变成

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

即

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon \end{aligned} \tag{3.32}$$

其中 $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$。因为 $\tilde{\beta}_1$ 随 X_2 变化，所以 X_1 对 Y 的效应不再是常数；调整 X_2 的值将改变 X_1 对 Y 的影响。

然后可以求交互项的p值，p值低则证明$\beta_3 \neq 0$，存在协同效应。

注意：实验分层原则(hierarchical principle) 规定，如果模型中含有交互项，那么即使主效应的系数的p 值不显著，也应包含在模型中。换句话说，如果X_1和X_2之间的交互作用是重要的，那么即使X_1和X_2的系数估计的p 值较大，这两个变量也应该被包含在模型中。

交互项的方法同样适用于有定性变量的情况。详见书P61-62。

2.非线性关系

方法：用多项式回归。

注意：

（1）要与实际情况符合。比如下图像二次函数

$$h_\theta(x) = \theta_0 + \theta_1(size) + \theta_2(size)^2$$

但是二次函数后期图像会下降，房价不会随面积的增大而下降，所以用

$$h_\theta(x) = \theta_0 + \theta_1(size) + \theta_2\sqrt{size}$$

更合适。

（2）它仍然是一个线性模型，比如

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \varepsilon$$

X_1 = horsepower
X_2 = horsepower^2
即可以用标准线性回归软件来估计β_0、β_1和β_2以得到非线性拟合。

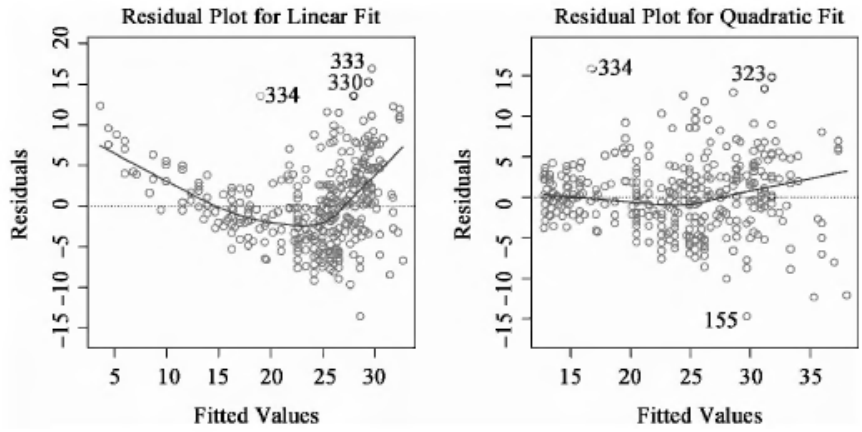
3.3.3其他可能会碰到的问题

（1）非线性的响应-预测关系（nonlinearity of response-predictor relationship）。
（2）误差项自相关（correlation of error term）。
（3）误差项方差非恒定（non-constant variance of error term）。
（4）离群点（outlier）。
（5）高杠杆点（high-leverage point）。
（6）共线性（collinearity）。

1.数据的非线性

判别方法：画残差图，识别非线性。

理想情况下，残差图显示不出明显的规律。若存在明显规律，则表示线性模型的某些方面可能有问题。比如下面左图，残差有规律，显示明显的U形，说明采用的线性模型有问题，数据非线性。右图残差没什么规律，说明采用的模型没问题。



如果残差图表明数据中存在非线性关系，那么一种简单的方法是在模型中使用预测变量的非线性变换，例如 log$X$，$\sqrt{X}$和 X^2。在本书后面的章节中，我们将讨论用其他更先进的非线性方法解决这个问题。

2.误差项之间有相关性（违背了假设，导致系数的真实区间比置信区间和预测区间宽）

判别方法：画残差图。如果误差项不相关，那么图中应该没有明显的规律；如果误差项是正相关的，那么我们可能在残差中看到跟踪(tracking) 现象——相邻的残差可能有类似的值。例子图详见书P65。

3.误差项方差非恒定（违背假设）

判别方法：画残差图，图上残差有变化就有问题，基本恒定就没问题。

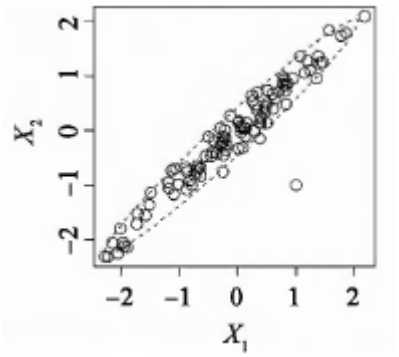
缓解办法：给观测值赋予权重。

4.离群点——y异常

判别方法：绘制学生化残差图（studentized residual），绝对值大于3的可能是离群点。

注意：一个离群点可能不是由失误导致的，而是暗示模型存在缺陷，比如缺少预测变量x。

5.高杠杆点——x异常



X_1正常，但对应X_2不正常，即X_i不正常，所以是高杠杆点。

判别方法：计算杠杆统计量（leverage statistic）。所有观测的平均杠杆值总是等于(p+1)/n。 如果给定观测的杠杆统计量大大超过该值，就有理由怀疑对应点有较高的杠杆作用。

6.共线性——两个或更多的预测变量高度相关，减小了假设检验的效力。

判别方法：计算方差膨胀因子（variance inflation factor，VIF），VIF超过5或10就表示有共线问题。

解决方案：

（1）剔除一个问题变量。

（2）把共线变量组合成一个新的预测变量。例如，对limit和rating求平均创建一个新的变量。

3.4线性回归于K最近邻法比较

当真实关系为线性时， KNN 略逊于线性回归，但在非线性情况下， KNN 大大优于线性回归。但在现实中，即使真实关系是高度非线性的， KNN 的结果仍有可能比线性回归更差。在更高维的情况下， KNN 的表现往往不如线性回归。预测效果随着维数的增加而恶化是KNN的一个普遍问题，这是因为在高维中样本量大大减少，出现维数灾难（curse of dimensionality）——高维情况下，出现给定的观测距离（K）附近没有邻点的情况。

数据中，有 100 个训练观测。当 $p = 1$ 时，这些点提供了足够的信息来准确估计 $f(X)$。然而，当这 100 个观测值分布在 $p = 20$ 个维度上时，将使得给定的观测附近没有**邻点**（nearby neighbours）——这就是所谓的**维数灾难**（curse of dimensionality）。当 p 很大时，与观测 x_0 最接近的 K 个观测可能在 p 维空间中距 x_0 很远，导致对 $f(x_0)$ 的预测非常差，从而产生一个很差的 KNN 拟合。一般规则是，若每个预测变量仅有少量观测，参数化方法往往优于非参数方法。

即使在低维问题上，从可解释性的角度来看，与 KNN 相比我们也会更倾向于线性回归。如果 KNN 的测试集 MSE 仅略低于线性回归，我们可能放弃一些预测精度，转而建立能被几个系数描述，且这些系数的 p 值都可知的简单模型。

第四章 分类

2018年12月3日 19:13

1.先验概率与后验概率：解释见https://zhuanlan.zhihu.com/p/26464206

2.大部分分类方法：估计变量属于不同类别的概率，将分类问题作为概率估计的结果。

3.为什么线性回归不适用了？
对于二元定性响应变量仍然是适用的，对有自然的程度顺序的多元响应变量也适用例如温和、中等和剧烈，其中温和和中等间的程度差距与中等和剧烈间的程度差距是相近的，那么Y=1,2,3的编码就是合理的。下面这个编码就是不合理的，因为它们之间的程度差距不一样。

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

4.1Logistic回归——对Y属于某一类的概率建模而不是对Y建模

响应变量为两类（Y=0,1）

4.1.1 简单logistic回归（p=1）

1.logistic模型
采用线性回归模型表示概率的话会出现（0,1）以外的概率值，不符合实际，所以要找到一个函数，其建立的模型输出结果都在0和1之间，有很多函数满足这个要求，在logistic回归中使用logistic函数

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

整理可得

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

2.步骤
（1）首先，判断X与Y有没有关系——假设检验：
类似于线性回归，这里用z统计量，p值小拒绝H_0
（2）确定X、Y存在关系后，评价简单logistic回归模型准确性：
（3）确定所选模型是正确的后，估计系数：极大似然法
基本思想：寻找β_0、β_1的估计值，使得算出来的p(X)与实际情况接近。就是将求出来的β_0、β_1估计值代入p(X)公式使所有违约人的值接近1，未违约人的值接近0。
即所估计的β_0、β_1使似然函数最大

$$\ell(\beta_0,\beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'})).$$

（4）确定参数后，评估估计值的准确性：
用标准误差（SE）衡量

注：哑变量方法分析定性预测变量同样适用于logistic回归，详见P94

4.1.2多元logistic回归（p≥2）

1.logistic模型：

X=(X_1，X_2,.....X_p)

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

2.步骤：

（1）首先，判断X与Y有没有关系——假设检验：

同上

（2）确定X、Y存在关系后，评价简单logistic回归模型准确性：

（3）确定所选模型是正确的后，估计系数：极大似然法

同上

（4）确定参数后，评估估计值的准确性：

同上

注意：与线性回归类似，只用一个预测变量得到的结果可能与多个预测变量得到的结果完全不一样，尤其是当这些因素之间存在相关性时。这种现象称为混淆现象(confounding) 。（混淆矩阵，改变阈值，ROC曲线，AUC大小）

响应变量超过两类（Y=1,2,3，……）

两类的logistic回归模型可以推广到多类，但不常用，更多用判别分析方法。

4.2线性判别分析（LDA，linear discriminant analysis）

当类别的区分度高，或者样本量n比较小，而且每一类中预测变量X近似服从正态分布以及相应分类多于两类时，用LDA。

响应变量两类与响应变量多类一样，都可以用LDA，故此处不分了

4.2.1简单LDA（p=1）

1.假设

每一类中的样本都服从均值不同，方差相同的正态分布。

2.思路

LDA分类器将

$$\hat{\pi}_k = n_k/n.$$

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2$$

代入

$$\hat{\delta}_k(x) = x\cdot\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

然后将样本分入使其值最大的一类中。

其中：

n：样本总量

n_k：第k类的样本量

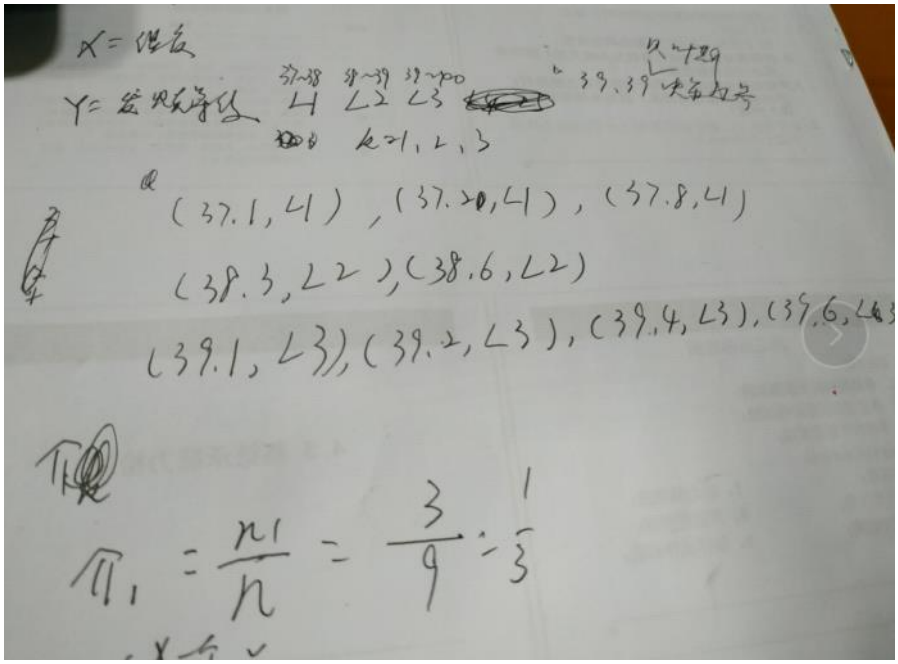
μ_k：第k类的样本均值

σ^{2}^：可视为K类样本方差的加权平均

π_k：一个随机选择的样本来自第k类的先验概率，已知直接用，信息不全时用上面那个公式

LDA决策边界：令不同类之间的$\delta(x)$互相相等解得x即为决策边界。

例子



4.2.2多元LDA（p≥2）——把简单LDA中的量变成向量、矩阵形式

1.假设

在预测变量的维度 $p>1$ 的情况下，LDA 分类器假设第 k 类观测服从一个多元高斯分布 $N(\mu_k, \Sigma)$，其中 μ_k 是一个均值向量，Σ 是所有 K 类共同的协方差矩阵。将第 k 类的密度函数

2.思路

多元LDA分类器将

$$\mu_1, \ldots, \mu_K,$$

$$\pi_1, \ldots, \pi_K, \text{ and } \Sigma;$$

代入

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k$$

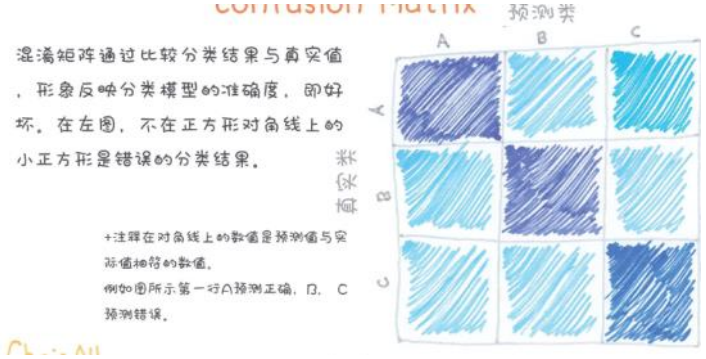
然后将样本分入使其值最大的一类中。

参数含义及决策边界同上。

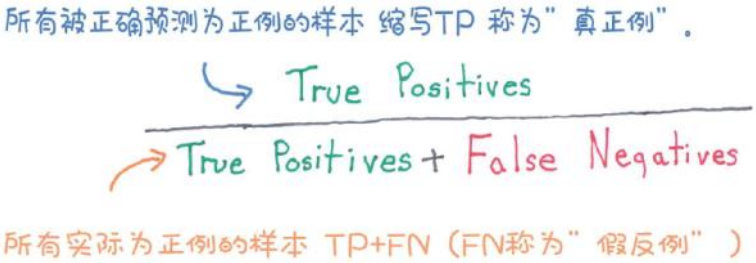
3.注意点

（1）训练错误率往往比测试错误率要低。参数p与样本总数n的比值越高（样本量不够），越容易过拟合。

（2）模型会错误归类。用混淆矩阵展示判别失误信息。



a. 灵敏度（也称召回率，sensitivity）



灵敏度太低的解决办法：降低阈值。比如原来概率大于0.5才判入那一类，降低到大于0.2,0.3等小于0.5的值就判入那一类。

降低阈值后灵敏度提高，总错误率也会提高——确定最优阈值取决于各自专业领域的知识。

b. ROC曲线：分类器的性能通过ROC曲线下的面积（AUC）衡量，AUC越大，分类器越好。

c. 混淆矩阵

针对一个二分类问题，将实例分成正类(postive)或者负类(negative)。但是实际中分类时，会出现四种情况.

(1)若一个实例是正类并且被预测为正类，即为真正类(True Postive, TP)

(2)若一个实例是正类，但是被预测成为负类，即为假负类(False Negative, FN)

(3)若一个实例是负类，但是被预测成为正类，即为假正类(False Postive, FP)

(4)若一个实例是负类，但是被预测成为负类，即为真负类(True Negative, TN)

TP:正确的肯定数目

FN:漏报，没有找到正确匹配的数目

FP:误报，没有的匹配不正确

TN:正确拒绝的非匹配数目

混淆矩阵如下，1代表正类，0代表负类：

		预测		
		1	0	合计
实际	1	True Postive **TP**	Frue Negative **FN**	Actual Postive(**TP+FN**)
	0	False Postive **FP**	True Negative **TN**	Actual Negative(**FP+TN**)
合计		Predicted Postive (**TP+FP**)	Predicted Negative (**FN+TN**)	**TP+FN+FP+TN**

由上表可得出横，纵轴的计算公式：

(1)真正类率(True Postive Rate)TPR: TP/(TP+FN),代表分类器预测的正类中实际正实例占所有正实例的比例。Sensitivity

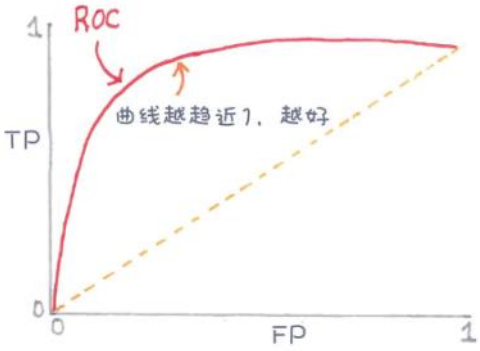
(2)负正类率(False Postive Rate)FPR: FP/(FP+TN)，代表分类器预测的正类中实际负实例占所有负实例的比例。1-Specificity

(3)真负类率(True Negative Rate)TNR: TN/(FP+TN),代表分类器预测的负类中实际负实例占所有负实例的比例，TNR=1-FPR。Specificity

假设采用逻辑回归分类器，其给出针对每个实例为正类的概率，那么通过设定一个阈值如0.6，概率大于等于0.6

的为正类，小于0.6的为负类。对应的就可以算出一组(FPR,TPR),在平面中得到对应坐标点。随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样也掺杂着真正的负实例，即TPR和FPR会同时增大。阈值最大时，对应坐标点为(0,0),阈值最小时，对应坐标点(1,1)。

如下面这幅图为ROC曲线，线上每个点对应一个阈值。



(详见https://www.cnblogs.com/dlml/p/4403482.html)

流程：指定阈值，按前面步骤归类后，画出混淆矩阵，得到灵敏度情况，改变阈值，得到不同的混淆矩阵，从而绘制ROC曲线。不同分类器重复此步骤可进行性能比较。

4.3二次判别分析（QDA，quadratic discriminant analysis）

4.3.1思路

假设每一类观测都有自己的协方差主矩阵，即第k类的观测服从

$$X \sim N(\mu_k, \Sigma_k),$$

这时候公式变成

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k\end{aligned}$$

是关于x的二次函数

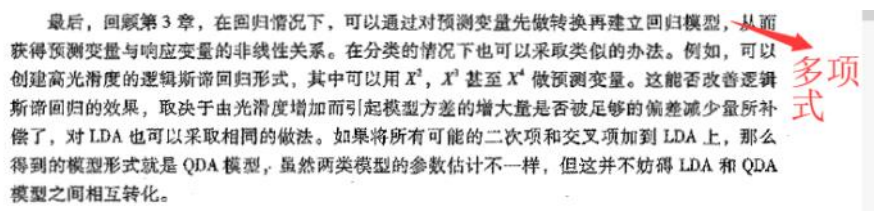
4.3.2LDA和QDA的选择是一个偏差-方差权衡的问题。

需要估计。所以，LDA 没有 QDA 分类器光滑，于是拥有更低的方差，该模型有改善预测效果的潜力，但这里也需要权衡考虑：如果 LDA 假设 K 类有相同的方差是一个非常糟糕的假设，那么 LDA 会产生很大的偏差。一般而言，如果训练观测数据量相对较少，LDA 是一个比 QDA 更好的决策，降低模型的方差很有必要。相反地，如果训练集非常大，则更倾向于使用 QDA，这时分类器的方差不再是一个主要关心的问题，或者说 K 类的协方差矩阵相同的假设是站不住脚的。

即看假设和训练样本量。

4.4分类方法的比较——logistic、LDA、QDA、KNN

当真实决策边界是线性时，logistic和LDA较好；边界是一般非线性时，QDA较好；边界更复杂时，非参数方法比如KNN较好，但是要谨慎选择非参数方法的光滑水平（K）。同时注意考虑假设成立与否和样本量的大小。


最后，回顾第 3 章，在回归情况下，可以通过对预测变量先做转换再建立回归模型，从而获得预测变量与响应变量的非线性关系。在分类的情况下也可以采取类似的办法。例如，可以创建高光滑度的逻辑斯谛回归形式，其中可以用 X^2，X^3 甚至 X^4 做预测变量。这能否改善逻辑斯谛回归的效果，取决于由光滑度增加而引起模型方差的增大量是否被足够的偏差减少量所补偿了，对 LDA 也可以采取相同的做法。如果将所有可能的二次项和交叉项加到 LDA 上，那么得到的模型形式就是 QDA 模型，虽然两类模型的参数估计不一样，但这并不妨碍 LDA 和 QDA 模型之间相互转化。

多项式

第五章 重抽样方法

2018年12月4日 15:23

1.常用方法：交叉验证法（cross-validation）和自助法（bootstrap）。

5.1交叉验证法——在拟合过程中，保留训练集的一个子集，然后将保留的子集作为测试集，运用统计学习方法，估计测试错误率。

5.1.1由来

一般缺少测试数据集

5.1.2验证集方法（validation set approach）——k=2的k折交叉验证法

1.思路

随机地把数据集分为两部分:一个训练集(training set)和一个验证集（validation set）/保留集（hold-out set），是在训练集上拟合统计学习方法，然后在验证集上评价其表现，用均方误差MSE衡量验证集误差。

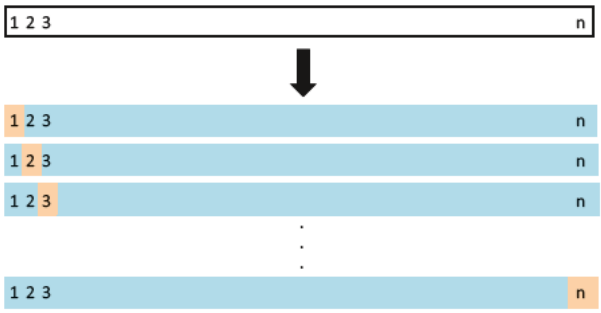
2.缺点

（1）波动大。

（2）被训练的样本量少，验证集错误率可能高估测试错误率。

5.1.3留一交叉验证法（leave-one-out cross-validation，LOOCV）——k=n的k折交叉验证法

1.思路



LOOCV 的原理图。一个有 n 个数据点的集合被反复地分割为一个训练集，包含除了一个观测之外的全部观测（图中蓝色部分），以及一个验证集，只包含剩下的那个观测（图中米黄色部分）。测试误差通过对 n 个所得到的均方误差求平均来估计。第一个训练集包含除了观测 1 之外的全部观测，第二个训练集包含除了观测 2 之外的全部观测，以此类推。

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

2.缺点

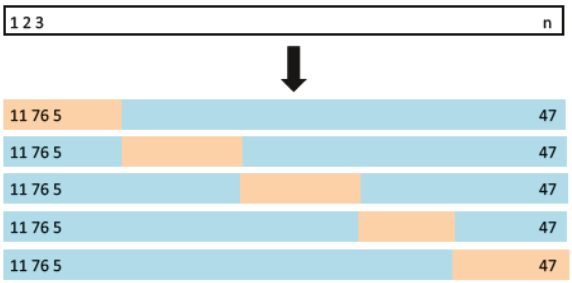
计算量大，只有在用最小二乘法拟合模型时可以用公式

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-h_i}\right)^2,$$

其他模型要拟合n次。

5.1.4k折交叉验证法（k-fold CV)

1.思路



5 折 CV 方法的原理图。一个有 n 个观测的集合被随机地分为 5 个不重叠的组。每一个组轮流作为验证集（图中米黄色部分），剩下的组作为训练集（图中蓝色部分）。测试误差通过对 5 个所得到的均方误差估计求平均来估计。

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$

5.1.5偏差-方差权衡

LOOCV法比k折CV法偏差小，方差大；

使用k折CV法时一般k取5或10。

5.1.6交叉验证法在分类问题中的应用

交叉验证法应用在分类问题中与应用在回归问题中的区别仅仅是用被误分类的样本的数量，而不是均方误差MSE来作为衡量测试误差的指标。

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i,$$

$$\mathrm{Err}_i = I(y_i \neq \hat{y}_i).$$

5.1.7交叉验证法的用途

1.估计某一种指定的统计模型的测试误差，进而评价其表现。

2.对几种统计模型比较选择最优的一种；或者为某一种统计模型选择最佳光滑度。——找测试均方误差曲线的最小值点。（横坐标为光滑度，纵坐标为交叉验证法的均方误差）

5.2自助法

5.2.1思路

从原始数据集（一共n个样本）有放回地抽取n个样本组成一个自助法数据集，抽B次（B很大），由B个自助法数据集估计出B个相应要检查的系数，然后用公式算出其估计值的标准误差。例子见书。

5.2.2用途

常用于衡量某个参数估计值的准确度或统计模型中某个值/预测（习题5.4.4）的准确度。线性回归的参数估计值准确度可以用软件包自动输出标准误差那些来衡量，自助法几乎可以用于所有情况，特别是测量指标波动性大和软件包无法自动输出衡量参数，很难甚至无法直接计算变量标准差的情况。

第六章 线性模型选择和正则化

2018年12月8日 10:24

1.为什么需要扩展?

与最小二乘法相比，其他拟合方法具有更高的预测准确度(prediction accuracy) 和更好的模型解释力(model interpretability) 。

预测准确率：若不满足n>>p，最小二乘法得出的结果方差大，过拟合。

模型解释力：在多元回归模型中，常常存在与响应变量无关的变量，它们增加了模型的复杂性，却与模型无关。而最小二乘法很难将其系数缩减至0，无法实现对无关变量的筛选。

2.主要有子集选择、压缩估计、降维法。

6.1子集选择——从p 个变量中挑选出与响应变量最相关的变量形成子集，再对子集使用最小二乘方法。包括最优子集选择和逐步模型选择。

6.1.1最优子集选择（best subset selection）

1.思路

1. 记不含预测变量的零模型为$\mathcal{M}_0$，只用于估计各观测的样本均值。
2. 对于$k=1, 2, \cdots, p$:
   (a) 拟合$\binom{p}{k}$个包含k个预测变量的模型;
   (b) 在$\binom{p}{k}$个模型中选择 RSS 最小或 R^2 最大的作为最优模型，记为 $\mathcal{M}_k$。
3. 根据交叉验证预测误差、C_p(AIC)、BIC 或者调整 R^2 从 $\mathcal{M}_0, \cdots, \mathcal{M}_p$ 个模型中选择出一个最优模型。

注意：如果只看RSS和R^2最终会选出包含所有变量的模型，过拟合，系数估计方差高。因为低RSS和高R^2表示训练误差低。所以要交叉验证。

2.缺点

计算效率低

6.1.2逐步选择

1.向前逐步选择（forward stepwise selection）

（1）思路

1. 记不含预测变量的零模型为$\mathcal{M}_0$。
2. 对于$k=0, 1, 2, \cdots, p-1$:
   (a) 从$p-k$个模型中进行选择，每个模型都在模型$\mathcal{M}_k$的基础上增加一个变量;
   (b) 在$p-k$个模型中选择 RSS 最小或 R^2 最高的模型作为最优模型，记为$\mathcal{M}_{k+1}$。
3. 根据交叉验证预测误差、C_p (AIC)、BIC 或者调整 R^2 从 $\mathcal{M}_0, \cdots, \mathcal{M}_p$ 个模型中选择出一个最优模型。

特点在于，每次只将能够最大限度地提升模型效果的变量加入模型中。

（2）缺点

可能在前期将后来多余的变量纳入模型，无法保证找到的模型是所有2^p个模型中最优的。

2.向后逐步选择（backward stepwise selection）

（1）思路

1. 记包含全部p个预测变量的全模型为$\mathcal{M}_p$。
2. 对于$k=p, p-1, \cdots, 1$:
   (a) 在k个模型中进行选择，在模型$\mathcal{M}_k$的基础上减少一个变量，则模型只含$k-1$个变量;
   (b) 在k个模型中选择 RSS 最小或 R^2 最高的模型作为最优模型，记为 $\mathcal{M}_{k-1}$。
3. 根据交叉验证预测误差、C_p(AIC)、BIC 或者调整 R^2 从 $\mathcal{M}_0, \cdots, \mathcal{M}_p$ 个模型中选择出一个最优模型。

特点在与，每次剔除一个对当前模型拟合效果最不利的变量。

（2）缺点

可能在前期将对后面来说的最优变量剔除了。

向后选择方法需满足样本量 n 大于变量个数 p（保证全模型可以被拟合）的条件。相反，向前逐步选择即使在 $n<p$ 的情况下也可以使用，因此当 p 非常大的时候，向前逐步选择是唯一可行的方法。

3.混合方法

逐次将变量加入模型中，在加入新变量的同时，也移除不能提升模型拟合效果的变量。

6.1.3选择最佳模型

RSS和R^2并不适用于对包含不同个数预测变量的模型进行选择，只看它们只会选出包含所有变量的训练误差小测试误差大的模型。因此引入C_p、AIC（赤池信息量准则，Akaike information criterion）、BIC（贝叶斯信息准则，Bayesian information criterion）、调整R^2（adjusted R^2）、交叉验证

1.C_p——越低测试误差越小

RSS 的基础上增加惩罚项 $2d\hat{\sigma}^2$，用于调整训练误差倾向于低估测试误差的这一现象。

2.AIC——适用于许多使用极大似然法拟合的模型，越低测试误差越小

$$\mathrm{AIC}=\frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS}+2d\hat{\sigma}^2\right),$$

3.BIC——越低测试误差越小

因为替换成了log，所以BIC对包含多个变量的模型施以更重的惩罚，与C_p相比，所得模型规模较小。

4.调整R^2——越高测试误差越小

$$\text{Adjusted } R^2=1-\frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}.$$

原理：当模型包含所有正确的变盘，再增加其他冗余变量只会导致RSS小幅度的减小。但加入这些冗余变量的同时d值增加，因此这些冗余变量的加入会导致调整R^2的值减小。

5.交叉验证

与前四种方法相比，交叉验证适用范围更广，在难以确定模型自由度和误差方法的情况下仍可使用。

试误差估计值相差不大。此外，如果对不同的训练集和验证集重复使用验证集方法，或者对于不同的交叉验证折数重复使用交叉验证方法，会得到不同的具有最低测试误差的精确模型。针对这种情况，可以使用一倍标准误差准则（one-standard-error rule）进行模型选择。首先计算不同规模下模型测试均方误差估计值的标准误差，然后选择测试样本集误差估计值在曲线最低点一倍标准误差之内且规模最小的模型。这样做的原因是在一系列效果近似相同的模型中，总是倾向于选择最简单的模型，也就是说，具有最少预测变量的模型。在

6.2压缩估计方法——将系数估计值往0的方向压缩，常用的有岭回归和lasso回归

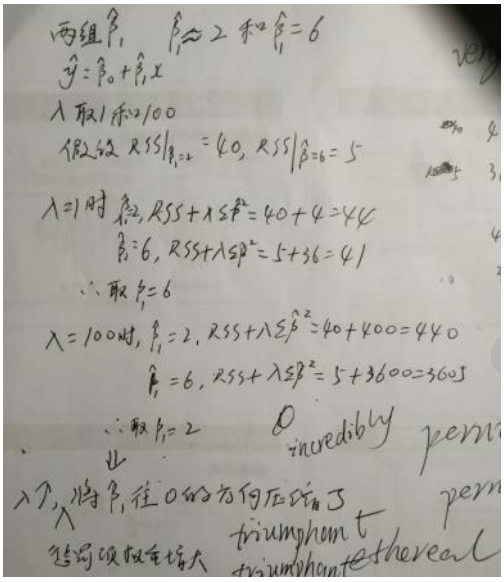
6.2.1岭回归（ridge regression）

1.思路

用岭回归方法得到的系数估计值就是指使下列式子最小的β_j值。$\lambda\geq 0$，称为调节参数。

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

实质上是加入了后面那项惩罚项，λ越大，惩罚项所占权重越大，要使总式子最小的话，在λ那么大的情况下，β_j能取的值越小。为什么的例子：



注意：在使用岭回归之前，要对预测变量进行标准化（因为岭回归公式中有系数平方和项，变量尺度变化有影响，这与最小二乘估计不同）：

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}},$$

2.优缺点

（1）优点：与最小二乘相比，岭回归的优势在于它综合权衡了误差与方差。随着λ 的增加，岭回归拟合结果的光滑度降低，方差降低，偏差增加。

（2）缺点：增加λ 的值能减小系数绝对值，但仍然无法剔除任何变量。

6.2.2lasso回归——改进版岭回归，当λ 足够大时能将某些系数的估计值强制设为0，得到的是稀疏模型（sparse model）

1.思路

用lasso回归方法得到的系数估计值就是指使下列式子最小的β_j值

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

其余和岭回归类似。

6.2.3岭回归与lasso回归

1.岭回归和lasso的其他形式

可以证明，两者等价于求解以下问题（线性规划，求交点）：

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s \tag{6.8}$$

and

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \le s, \tag{6.9}$$

λ与s是对应关系

2.为什么lasso可以将系数估计完全压缩至零，而岭回归不可以?

如图6-7，因为岭回归的条件区域是没有尖点的圆形，所以这个相交点一般不会出现在坐标轴上，所以岭回归系数估计不为零。而lasso的条件区域在每个坐标轴上都有拐角，所以椭圆经常在坐标轴上与条件区域相交。这种情况下，其中一个系数就会为零。而在多维情况下（图形将变成球体、超球面（岭回归）和多面体（lasso回归）），就会有一些系数为0。

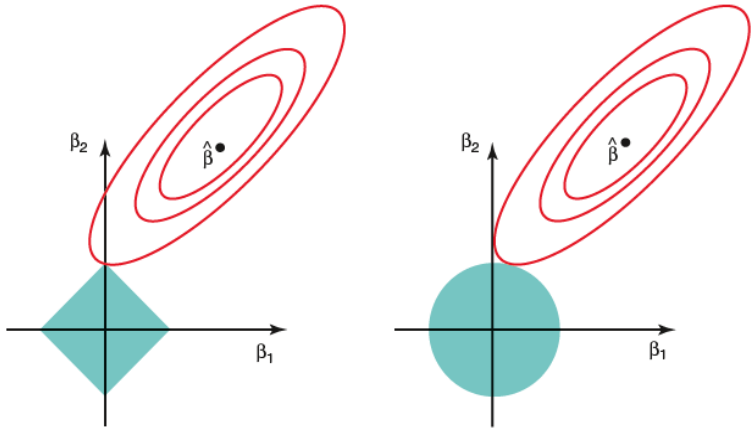


图 6-7　误差等高线和限制条件区域（左：lasso；右：岭回归）。实心区域是限制条件 $|\beta_1|+|\beta_2|\leqslant s$ 和 $\beta_1^2+\beta_2^2\leqslant s$，椭圆是 RSS 等高线。

3.选择合适的λ：交叉验证法。选择一系列λ 的值，计算每个λ 的交叉验证误差。然后选择使得交叉验证误差最小的λ。最后，用所有可用变量和选择的λ拟合模型。

4.岭回归和lasso比较

（1）相同点：随着λ增大，方差减小，偏差增大。

（2）不同点：当只有一小部分预测变量是真实有效的而其他预测变量系数非常小或者等于零时， lasso 要更为出色（它能筛选掉那些无效的变量，解释性更强），当响应变量是很多预测变量的函数并旦这些变量系数大致相等时，岭回归较为出色。

然而，对于一个真实的数据集，与响应变量有关的变量个数无法事先知道。因此需要交叉验证。

存在可以同时拟合岭回归和lasso回归的高效算法，每个算法的系数估计运算与一个最小二乘拟合运算的计算量基本一致。

5.岭回归和lasso的例子

一个最简单的例子（压缩机理）：

为了更好地理解岭回归和 lasso 的原理，考虑一个简单而特殊的例子，即 $n=p$，$\mathbf{X}$ 是对角线都为 1、非对角线位置都为零的对角矩阵。进一步简化问题，假设考虑的是没有截距的回归。在这些假设之下，普通的最小二乘问题就简化为寻找 β_1，…，β_p 来最小化

$$\sum_{j=1}^{p}(y_j-\beta_j)^2 \tag{6.11}$$

此时，最小二乘的解是 $\hat{\beta}_j=y_j$。

在这种情况下，岭回归等同于寻找 β_1，…，β_p，使得

$$\sum_{j=1}^{p}(y_j-\beta_j)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{6.12}$$

达到最小；lasso 则相当于寻找使得

$$\sum_{j=1}^{p}(y_j-\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j| \tag{6.13}$$

达到最小的系数。可以证明在这种情况下，岭回归估计有如下形式

$$\hat{\beta}_j^R=y_j/(1+\lambda) \tag{6.14}$$

而 lasso 估计形式如下

$$\hat{\beta}_j^L=\begin{cases}y_j-\frac{\lambda}{2}, & y_j>\frac{\lambda}{2}\\ y_j+\frac{\lambda}{2}, & y_j<-\frac{\lambda}{2}\\ 0, & |y_j|\leqslant\frac{\lambda}{2}\end{cases} \tag{6.15}$$

图 6-10 表示了这种情况。可以看到，岭回归和 lasso 表现出两种不同的系数压缩方式。岭回归中每个最小二乘系数以相同比例压缩。相比而言，lasso 中每个最小二乘系数以 $\lambda/2$ 为阈值压缩至零，即那些绝对值小于 $\lambda/2$ 的系数被完全压缩至零。在简单设置（6.15）下，lasso岭回归以相同比例压缩每个维度，然而lasso回归以绝对数值压缩系数，小于一定值的系数被设为0。高维情况一样。

6.3降维方法（dimension reduction）

6.3.1降维法思路

$Z_1,Z_2,\ldots,Z_M$表示M个所有（p个）原始预测变量的线性组合（M < p），即

$$Z_m=\sum_{j=1}^{p}\phi_{jm}X_j$$

然后用最小二乘法对转换后的变量$Z_1,Z_2,\ldots,Z_M$拟合：

$$y_i=\theta_0+\sum_{m=1}^{M}\theta_m z_{im}+\epsilon_i,\quad i=1,\ldots,n,$$

“降维”这个术语是指某种方法可以使估计 $p+1$ 个系数 β_0，β_1，…，β_p 的问题简化为估计 $M+1$ 个系数 θ_0，θ_1，…，θ_M 的问题，这里 $M<p$。也就是说，问题的维度从 $p+1$ 降至 $M+1$。

关键是选择合适的$Z_1,Z_2,\ldots,Z_M$，即选择合适的φ_{jm}，主要有主成分分析和偏最小二乘法。

6.3.2主成分分析法（principal components analysis，PCA）

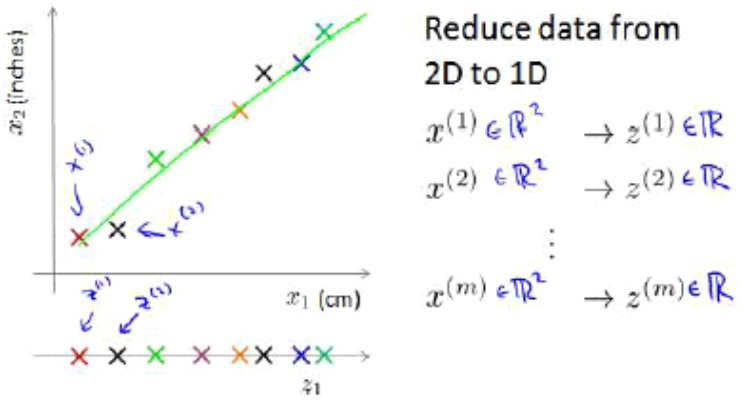
1.助于理解的储备知识

详见https://blog.csdn.net/Murray_/article/details/79945148

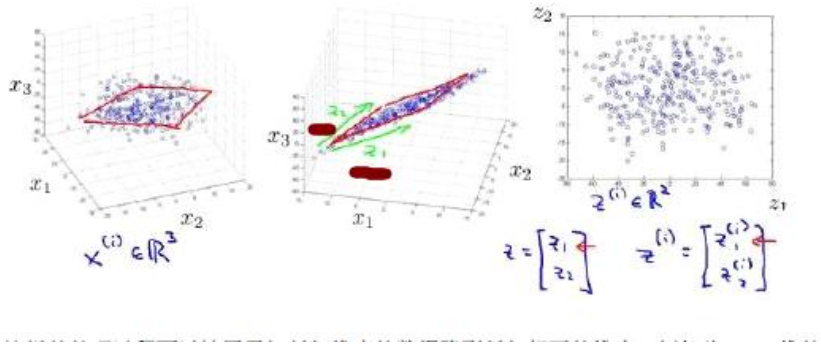
矩阵相乘的意义是将右边矩阵中每一列列向量变换到左边矩阵中每一行行向量为基所表示的空间中去，所得结果就是右边矩阵在新空间中的表示（右边矩阵每一列列向量在新空间中的坐标表示）

2.PCA思路

二维降为一维：



三维降为二维：



三维也可降为一维，如果三维空间中点的分布实际是一个面的话。

其他高维情况类似。Z即所称的主成分。

$$Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}}). \quad (6.19)$$

变量标准化

（0.839, 0.544）：新空间的基向量

Z：横轴上的坐标

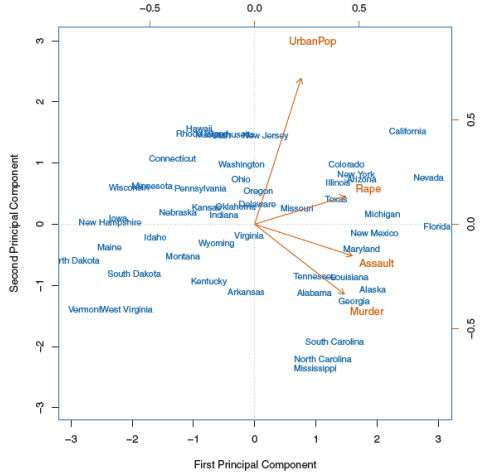
所以，主成分可以理解为新空间的维度。主成分分析就是找到一些维度/主成分（一个新空间，维度比现空间小），将原始预测变量都变换到新空间里面，用新的维度/主成分来拟合模型。

（1）主成分得分（score）：表示原模型在新空间中各个维度的值。

（2）判断主成分（新维度）代表什么指标（新变量是什么）：看载荷向量（即新空间的基向量）在原空间各个x_i上的权重，哪些x_i权重比较大就意味着这个维度主要表示哪个意思。比如下面美国50个州犯罪数据集，n=50,p=4：

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

第一载荷向量在Assault 、Murder 和Rape 这3个变量上的权重大致相等，而在UrbanPop上的权重相对较小，因此这个主成分大致反映了严重罪行的总体犯罪率。第二主成分向量在UrbanPop上有较大权重，在其他3 个变量上权重较小，因此这个主成分大致反映了每个州的城市化水平。像下图印第安纳州这样在两个主成分上的值都趋于0 的州，表示它在犯罪率水平和城市化水平上大致处于中等水平。



（3）投影要点

a. 投影到新空间的点要尽可能分散，也就是方差最大化的原则。

b. 维度/主成分之间要零相关，即相互正交。

c. 新空间的基（主成分Z的系数）如何找？（一系列数学运算）

协方差矩阵可以很好地表示各个特征的方差和不同特征之间的协方差；按目标要求需要将协方差矩阵除去对角线元素化为0，对角线元素按从大到小排列。协方差矩阵对角化；对角化后按特征值大小排列，选定前k个最大的特征值对应的特征向量作为降维后的新基。

3.PCA注意点

（1）PCA不是特征提取方法，因为用于回归的M个主成分是全部p个原始变量的线性组合。

（2）构造主成分之前要对变量进行标准化处理

第一步是均值归一化。我们需要计算出所有特征的均值，然后令 $x_j = x_j - \mu_j$。如果特征是在不同的数量级上，我们还需要将其除以标准差 σ^2。

（3）主成分数量M的确定

a. PCA用于回归问题中M的确定：交叉验证法。

b. PCA用于无指导学习中M的确定：看主成分的PVE和累计PVE（proportion of variance explained，方差解释比例，反映了将原始样本投影到前几个少数主成分损失了多少信息）、碎石图（scree plot），满足要求即可。

在实践中，往往通过看前几个主成分来寻找数据中有价值的模式。如果在前几个主成分中都找不到有价值的模式，那更多的主成分也不太可能会有价值。相反，如果前几个主成分有价值，那通常会继续观察随后的主成分，直到找不到更多有价值的模式为止。

使用几个少数主成分的好处是可以得到一个噪声较小的结果，因为数据集中的主要信号（而不是噪声）通常集中在少数几个主成分中。

4.PCA缺点

是无指导的方法，构造主成分的过程中没有用到响应变量Y，无法保证那些很好地解释预测变量的方向同样可以很好地预测响应变量。

6.3.3偏最小二乘（partial least squares，PLS）——解决PCA的缺点，是有指导的方法（使新空间的选择更准确一点）

1.PLS思路

同样的，它将原始变量的线性组合Z作为新的变量集，然后用这M 个新变量拟合最小二乘模型。与PCA不同，偏最小二乘利用了响应变量Y 的信息筛选新变量。即计算Z时将最大权重赋给与响应变量相关性最强的变量。例子见书P162-163。

2.PLS注意点

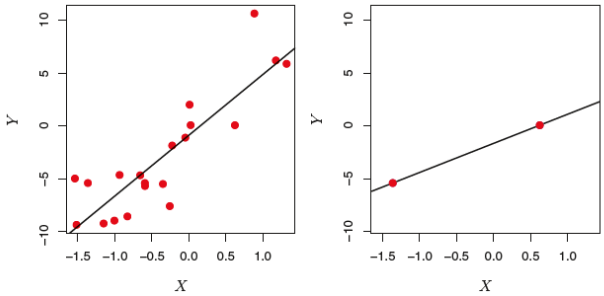
（1）个数M通过交叉验证选择。

（2）应用前应对预测变量和响应变量做标准化处理。

6.3.4高维数据——p > n或p≈n或p只是略小于n的数据

1.高维数据引起的问题

（1）最小二乘回归线非常光滑，导致过拟合。



（2）引起维数灾难：除非给模型新增的预测变量（维度）与响应变量确实相关，否则测试误差随着预测变量（维度）的增大而增大。比如噪声的引入，为了减少训练误差，就将噪声拟合了进去，导致测试误差增大，过拟合。

2.运用高维数据的注意点

因此，不能在训练集上用RSS、p值、R^2统计量或其他传统的拟合效果度量方法来度量高维情况下的拟合效果，应当在独立测试集上验证或者进行交叉验证。比如**独立测试集**的均方误差或R^2就是对模型拟合效果的有效度量，而训练均方误差则不是。

第七章 线性模型拓展/非线性模型

2018年12月10日 16:07

- 多项式回归（polynomial regression）对线性模型的推广思路是以预测变量的幂作为新的预测变量以替代原始变量。举例来说，一个三次回归模型有三个预测变量 X，X^2，X^3。这是一种简单实用的表达数据非线性关系的模型。
- 阶梯函数（step function）拟合是将某个预测变量的取值空间切割成 K 个不同区域，以此来生成一个新的定性变量，分段拟合一个常量函数。
- 回归样条（regression spline）方法在形式上比多项式回归和阶梯拟合方法更灵活，实际上回归样条可以看做是这两类方法的推广。首先将 X 的取值范围切割成 K 个区域，在每个区域分别独立拟合一个多项式函数。回归样条的多项式一般有一些限制以保证在区域边界或称为结点的位置，这些多项式得到光滑的连接。只要 X 被切割成尽可能多的区域，这个方法就能产生非常光滑的拟合效果。
- 光滑样条（smoothing spline）与回归样条类似，但是产生机理略有不同，一般是通过最小化一个带光滑惩罚项的残差平方和的式子来得到光滑样条的结果。
- 局部回归（local regression）与样条结果比较相近，最大的差别在于局部回归中的区域之间是可以重叠的，这个条件保证了局部回归整体光滑的拟合结果。
- 广义可加模型（generalized additive model）实际上是将上述模型推广到有多个预测变量的情形。

1.本章中前五种都是单预测变量的模型。

2.基函数（basis functions）

用b(X)代替X进行建模，即

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i.$$

$b_K(X)$称为基函数，建模之前形式就已选定。例如对于多项式回归，基函数就是$b_j(x_i)=x_i^j$，对于阶梯函数基函数就是$b_j(x_i)=I(c_j \le x_i < c_{j+1})$。然后可以用最小二乘法估计系数。有关线性模型的标准误、F统计量等指标都可以用。

上图前三种就是用基函数代替X然后用最小二乘法，广义可加模型可以看成是广义的基函数代替法。

7.1多项式回归（polynomial regression）——基函数是幂次方

7.1.1定义

1.定量回归

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,$$

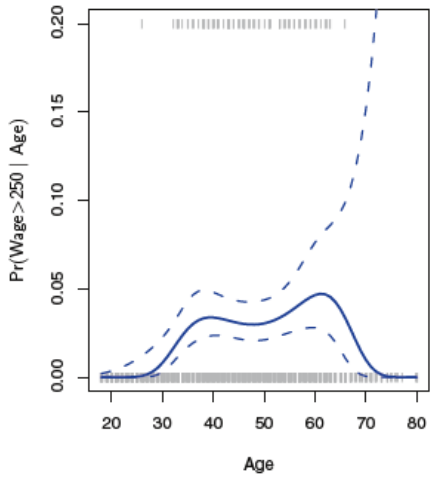
d一般不大于3或4

2.定性回归

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}.$$

7.1.2注意点

对于定性变量的回归，正确的样本数要足够，否则估计系数会有较大方差，置信区间会较宽。比如下图不同年龄段针对二元变量wage是否大于250，尽管建模总样本量（n=3000）足够，但>250的高收入人群只有79个，所以估计误差较大。（类似的书上P199类似问题）



7.2阶梯函数（step function）即分段函数——基函数是分段函数

7.2.1定义

1.定量回归

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \varepsilon_i$$

对于 X 的一个给定值，$C_1(X)$，$C_2(X)$，…，$C_K(X)$ 中至多只有一项系数非零。
$X < c_1$ 时，式（7.5）每个预测变量都为零，所以 β_0 即为 $X < c_1$ 时的 Y 的平均值。

2.定性回归

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}$$

7.3回归样条（regression spline）基函数为截断幂基函数

1.本质：设定结点和基函数，然后用最小二乘法估计样条函数（系数）。

2.自由度：K个结点，d阶样条，产生K+d+1个自由度。(d+1)(K+1)-Kd=K+d+1

3.样条：具有d-1阶连续导数的d阶分段多项式。一般用三阶样条，阶数过大就趋于多项式就没意义了。

4.自然样条：边界区域线性处理（变成一阶），这里边界区域指的是X的值比最小的结点值小或比最大的结点值大。则自由度减去2(d-1)。非自然样条在边缘处易过拟合。

7.3.1定义——分段多项式

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \qquad (7.8)$$

其中不同区域内的系数 β_0，β_1，β_2 和 β_3 都不相同。系数发生变化的临界点称为结点（knot）。

例如

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i, & x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i, & x_i \geqslant c. \end{cases}$$

7.3.2约束条件

1.要有连续的约束条件，否则分段处会跳跃、不连续。

2.d阶样条的约束条件：在每个结点处直到d-1阶都是连续的（函数连续性（0阶导连续）、一阶导连续、二阶导连续.....d-1阶导连续）

7.3.3确定结点个数和位置

对不同的结点数K采用交叉验证法，或者根据实践经验（令结点在数据上均匀分布）

7.3.4与多项式回归对比

样条函数通过增加结点个数但保持自由度固定的方法来使结果变得光滑。

7.4光滑样条（smoothing spline）

7.4.1定义

使下列式子最小化的函数g就是光滑样条。λ是非负的调节参数（tuning parameter）。

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

式子本质是采用“损失函数+惩罚项”的形式，惩罚项是对函数g的波动性进行惩罚。一阶导表示斜率，二阶导表示斜率的变化程度（函数粗糙度，roughness），积分表示求和，所以式子$\int g''(t)^2 dt$表示的是函数g的斜率变化程度之和。函数g波动性大，则积分项大，在λ存在下，因为要使式子最小化，所以惩罚项会使系统选取积分项小，即波动性小（光滑）的g函数。

当 $\lambda = 0$ 时，式（7.11）中的惩罚项不起作用，因此函数 g 会很跳跃并且会在每个训练数据点上做插值。当 $\lambda \to \infty$ 时，g 会变得非常平稳，也就是说会变成一条尽可能接近所有训练点的直线。实际上，在这种情况下，g 是最小二乘直线，因为式（7.11）的损失函数就是最小化残差平方和。对于一个比较适中的 λ 值，g 会尽可能地接近训练点但同时也比较光滑。λ 在这里就是控制光滑样条的偏差－方差的权衡。

7.4.2光滑样条与回归样条

光滑样条就是将每一个不同的xi都设为结点的**自然**样条（边界线性处理）。但是将每个数据点都作为一个结点会使得光滑样条的自由度太高，从而变动剧烈，所以引入惩罚项控制自由度。

光滑样条将每个数据点作为结点，回归样条粗略划分结点；光滑样条用惩罚项得到样条函数（的系数），回归样条用最小二乘法得到样条函数（ 的系数）。

7.4.3确定λ

使用交叉验证，利用下面式子计算光滑样条的LOOV，代价与计算一个拟合模型一样。

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_{\lambda}^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - \{\mathbf{S}_{\lambda}\}_{ii}}\right]^2 .$$

除非有足够理由支持一个更复杂的模型，否则以简为佳（选自由度较低的）

7.5局部回归（local regression）

7.5.1定义

详见http://lib.csdn.net/article/machinelearning/36684

如果目标假设不是线性模型，比如一个忽上忽下的函数，这时用线性模型就拟合的很差。为了解决这个问题，在预测一个点的值时，选择和这个点相近的点而不是全部的点做线性回归拟合。即在预测一个X所对应的Y值的时候，只采用在X周围的数据进行数据拟合。

算法：

（1）选取合适s比例的靠近x_0的数据x_i。

（2）对选出的数据点赋予不同权重$w^{(i)}$。离x_0最远及没被选中的数据点权重为0，最近的点权重最高。一个比较好的权重函数是

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

（3）用分配好的权重在x_i处进行加权最小二乘回归，使下式最小

$$\sum_i w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2.$$

（4）根据得到的回归函数带入x_0得到预测值。

7.5.2关键点

1.选择合适的s。s过小拟合效果会剧烈起伏，过大则相当于全局拟合，没有意义。交叉验证法确定或根据经验直接给定。

2.选择合适的权重函数$w^{(i)}$。

7.5.3缺点

对于每一个要查询的点，都要重新依据整个数据集（新的数据集、新的权重值）计算一个线性回归模型出来，计算代价极高。

7.5.4推广

变系数模型（varying coefficient model）：将模型扩展到最近收集数据上的有效方式。对某些变量使用全局回归，另一些变量比如时间使用局部回归。

局部回归于KNN法类似，都是取一定范围的最近的数据点，所以局部回归也有维数灾难问题。

7.6广义可加模型（generalized additive model，GAM）——既可用于定量变量也可用于定性变量

7.6.1定义

可以这么理解：用“广义的不同形式的基函数”代替各个X_i组合在一起。

1.回归问题的GAM：

将

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

变成

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned}$$

2.分类问题的GAM：

将

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

变成

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$

7.6.2缺点

被限定为可加形式，在多变量情况下，会忽略有意义的交互项。不过类似线性回归，也可以增加形式为Xj×Xk的交互项或其他低维交互项，然后用二维光滑方法如局部回归或二维样条来拟合。

若想摆脱形式上的限定，需要用随机森林和提升法之类更加光滑的方法。

GAM可以视为介于线性模型和完全非参数模型之间的一类折中建模方法。

第八章 基于树的方法

2018年12月11日 17:33

8.1决策树基本原理

8.1.1基本概念

根节点、内部节点、终端节点/树叶、分支

8.1.2建立树——包括特征选择、树的生成和剪枝

1.储备知识

（1）回归树中：

分割点将预测变量Xj的空间分为两个区域

$$R_1(j,s)=\{X|X_j<s\} \text{ and } R_2(j,s)=\{X|X_j\geq s\},$$

RSS就是

$$\sum_{i:\, x_i\in R_1(j,s)}(y_i-\hat{y}_{R_1})^2+\sum_{i:\, x_i\in R_2(j,s)}(y_i-\hat{y}_{R_2})^2,$$

（2）分类树中：

a. 分类错误率（classification error rate）：此区域训练集中除了最常见类别以外的其余类别所占比例。

$$E=1-\max_k(\hat{p}_{mk}).$$

p_{mk}表示第m个区域的训练集中第k类所占比例。

b. 基尼系数（Gini index）：衡量节点纯度（purity），越小越表示该节点包含的样本几乎来自同一类别。

$$G=\sum_{k=1}^{K}\hat{p}_{mk}(1-\hat{p}_{mk}),$$

c. 互熵（cross-entropy）：同基尼系数

$$D=-\sum_{k=1}^{K}\hat{p}_{mk}\log\hat{p}_{mk}.$$

（3）RSS、分类错误率、基尼系数、互熵都是拟合树产生的拟合误差。

2.建立完整树——递归二叉分裂（recursive binary splitting）

在所有变量X_1，X_2，.....X_p中遍历变量X_j（j从1取到p），在每个X_j中扫描且分点s，最后选出使RSS（回归树）、分类错误率/基尼系数/互熵（分类树）这些拟合误差最小的(j,s)对，作为分裂点，然后对划分后的区域循环这个过程，直到满足某个阈值时停止，即生成了决策树。

3.剪枝

上一步产生的完整树可能会过拟合（将噪声包含进去等等），因此需要减掉一些分支。分为预剪枝和后剪枝。

（1）预剪枝

a. 思路：在构建决策树的同时进行剪枝，设定一个阈值，当节点分裂后性能提升小于这个阈值就停止分裂。

b. 缺点：一些起初看来不值得的分裂可能会在后面某一步产生很好的分裂效果。因此预剪枝实际应用效果不好。

（2）后剪枝

a. 思路：在用训练集构建出一个完整的决策树（尽量大）后，再进行剪枝。

b. CART剪枝法：先生成一系列子树，再用交叉验证法对子树进行选择。

任一棵树的损失函数：

$$C_\alpha(T)=C(T)+\alpha|T| \tag{5.26}$$

其中，T 为任意子树，$C(T)$ 为对训练数据的预测误差（如基尼指数），$|T|$ 为子树的叶结点个数，$\alpha\geqslant 0$ 为参数，$C_\alpha(T)$ 为参数是 α 时的子树 T 的整体损失．参对固定的 α，一定存在使损失函数 $C_\alpha(T)$ 最小的子树，将其表示为 T_α．T_α 在损失函数 $C_\alpha(T)$ 最小的意义下是最优的．容易验证这样的最优子树是唯一的．当 α 大的时候，最优子树 T_α 偏小；当 α 小的时候，最优子树 T_α 偏大．极端情况，当 $\alpha=0$ 时，整体树是最优的．当 $\alpha\to\infty$ 时，根结点组成的单结点树是最优的．

如何选择合适的α？α在这里表示一个区间（依次增大的g(t)顺次组成的区间）的临界值，在这个区间内不剪比剪掉好，临界时（g(t)）不剪和剪掉一样，就剪掉（取简单形式），并将临界值取为α。于是一系列α分别对应各个区间内最优子树。再对这些子树用交叉验证法得到整体最优子树。

具体地，从整体树 T_0 开始剪枝．对 T_0 的任意内部结点 t，以 t 为单结点树的损失函数是

剪后的局部损失函数大小 $C_\alpha(t)=C(t)+\alpha$ 单结点树：$|T|=1$ (5.27)

以 t 为根结点的子树 T_t 的损失函数是

未剪的局部损失函数大小 $C_\alpha(T_t)=C(T_t)+\alpha|T_t|$ (5.28)

当 $\alpha=0$ 及 α 充分小时，有不等式

$$C_\alpha(T_t)<C_\alpha(t) \tag{5.29}$$

当 α 增大时，在某一 α 有

$$C_\alpha(T_t)=C_\alpha(t) \tag{5.30}$$

当 α 再增大时，不等式 (5.29) 反向．只要 $\alpha=\dfrac{C(t)-C(T_t)}{|T_t|-1}$，$T_t$ 与 t 有相同的损失函数值，而 t 的结点少，因此 t 比 T_t 更可取，对 T_t 进行剪枝．

为此，对 T_0 中每一内部结点 t，计算

$$g(t)=\frac{C(t)-C(T_t)}{|T_t|-1} \tag{5.31}$$

它表示剪枝后整体损失函数减少的程度．在 T_0 中剪去 $g(t)$ 最小的 T_t，将得到的子树作为 T_1，同时将最小的 $g(t)$ 设为 α_1．T_1 为区间 $[\alpha_1,\alpha_2)$ 的最优子树．

如此剪枝下去，直至得到根结点．在这一过程中，不断地增加 α 的值，产生新的区间．

（比较**每一个**内部结点，剪去g（t）最小的树）

为什么要选择最小的g(t)呢？以两个点为例，结点1和结点2，g(t)2大于g(t)1，假设在所有结点中g(t)1最小，g(t)2最大，当选择最大值g(t)2，即结点2进行剪枝，但此时结点1的不修剪的误差大于修剪之后的误差，即如果不修剪的话，误差变大，依次类推，对其它所有的结点的g(t)都是如此，从而造成整体的累计误差更大。反之，如果选择最小值g(t)1，即结点1进行剪枝，则其余结点不剪的误差要小于剪后的误差，不修剪为好，且整体的误差最小。

算法 5.7（CART 剪枝算法）

输入：CART 算法生成的决策树 T_0；

输出：最优决策树 T_α.

（1）设 $k=0$，$T=T_0$.

（2）设 $\alpha=+\infty$.

（3）自下而上地对各内部结点 t 计算 $C(T_t)$，$|T_t|$ 以及

$$g(t)=\frac{C(t)-C(T_t)}{|T_t|-1}$$

$$\alpha=\min(\alpha,g(t))$$

这里，T_t 表示以 t 为根结点的子树，$C(T_t)$ 是对训练数据的预测误差，$|T_t|$ 是 T_t 的叶结点个数.

（4）自上而下地访问内部结点 t，如果有 $g(t)=\alpha$，进行剪枝，并对叶结点 t 以多数表决法决定其类，得到树 T.

（5）设 $k=k+1$，$\alpha_k=\alpha$，$T_k=T$.

（6）如果 T 不是由根结点单独构成的树，则回到步骤 (4).

（7）采用交叉验证法在子树序列 $T_0,T_1,\cdots,T_n$ 中选取最优子树 T_α. ■

参考解释

PRUNING TREE

修剪(*pruning*) 是为了获得大小合适、误分率（*misclsssification*）低，评价准确率高的子树。*CART*采用的代价复杂性剪枝（*cost-complexity pruning*）算法是后剪枝方法的一个实例.

利用该算法生成一系列T_{max}的修剪子树$T_k: T_1 > T_2 > T_3 ... > t_k$（$t_k$为一棵以根和左右子树为叶节点的树）。修剪过程主要完成生成有序树序列和确定叶节点的所属类两步骤.

当树T在节点t被剪枝时，它的表面误差率增加$R(t)-R(T_t)$，而树叶的数量减少$|N_{T_t}|-1$，则

$$\alpha=\frac{R(t)-R(T_t)}{|N_{T_t}|-1}$$

度量每个被剪树叶的表面误差率的增加。算法对每棵子树计算α，并选择具有最小α值的子树进行剪枝.

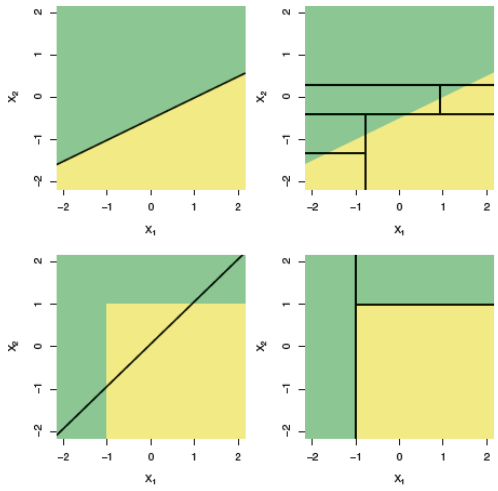
详见https://www.zhihu.com/question/22697086

https://blog.csdn.net/zhengzhenxian/article/details/79083643

4.优缺点

（1）优点：解释性强，可视化强，能直接处理定性变量而不用设哑变量。

（2）缺点：但预测准确性一般无法达到其他回归或分类方法的水平。



8.2改进树的预测效果

8.2.1装袋法（bagging），也叫自助法聚集（bootstrap aggregation）

1.思路——自助法

从训练集中重复取样，生成B个自助抽样训练集，生成B棵“根深叶茂”未经剪枝的回归树，分别对应预测值，最后对所有预测值求平均得到整体模型预测值

$$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

对于分类树：则对于一个给定的预测变量x_i，记录B棵树各自对它给出的预测类别，用多数投票（majority vote）的方式将B个预测中出现频率最高的类别作为预测结果。

注意：B很大也不会产生过拟合，足够大的B值能使误差稳定下来。

2.装袋法误差估计——袋外误差估计
可以证明，平均每棵树抽取到的样本是原始数据集的2/3，剩余1/3没有被抽到的样本称为这棵树的袋外样本（out-of-bag，OOB）。因此，对每一棵树，将OOB样本带入，得到每棵树的误差，由此计算总体模型误差。

3.度量变量重要性
当大量的树被装袋后，就无法仅用一棵树展现相应的统计学习过程，也不清楚哪些变量在分类过程中最为重要 （解释性差），因此可在装袋建模过程中记录下任一给定预测变量分裂后RSS/基尼系数的减小量，对每个预测变量分裂导致的减小总量在所有B棵树上取平均，值越大说明该预测变量越重要。

8.2.2随机森林（random forest）
1.思路
在建立决策树时，每此要进行分裂时，都要从全部的p个预测变量中选出包含m个预测变量的随机样本作为候选分裂点。这个分裂点的预测变量只能从这m个变量中选择。在每个分裂点处都重新进行抽样，通常$m\approx\sqrt{p}$。

也就是说，在建立随机森林的过程中，在每一个分裂点处，算法将大部分预测变量排除在外。为什么要这么做？因为在装袋法中，几乎所有树都会将最强的预测变量用于顶部分裂点，这样装袋法的树都很相似，这就意味着装袋法树中的预测变量是高度相关的，它与单棵树相比无法大幅降低方差。而随机森林法强迫每个分裂点只考虑预测变量一个子集，就让其他变量有了更多机会（对树去相关，decorrelate），降低方差。

随机森林和装袋法一样，不会因为B的增大而造成过拟合，所以在实践中应取足够大的B，使分类错误率降低到稳定的水平。

8.2.3提升法（boosting）
1.思路
先生成一个决策树，之后用该树的残差生成响应值为残差的新树加入到上一个树中，即利用残差改进决策树。也就是说，如果一个点的值被预测错误，那么在下一个回归树里面的模型的权值会变大。通过这个方式，来提高模型的效果。

算法 8.3（回归问题的提升树算法）

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i\in\mathcal{X}\subseteq\mathbf{R}^n$，$y_i\in\mathcal{Y}\subseteq\mathbf{R}$；

输出：提升树 $f_M(x)$.

（1）初始化 $f_0(x)=0$

（2）对 $m=1,2,\cdots,M$

（a）按式 (8.27) 计算残差

$$r_{mi}=y_i-f_{m-1}(x_i),\quad i=1,2,\cdots,N$$

（b）拟合残差 r_{mi} 学习一个回归树，得到 $T(x;\Theta_m)$

（c）更新 $f_m(x)=f_{m-1}(x)+T(x;\Theta_m)$

（3）得到回归问题提升树

$$f_M(x)=\sum_{m=1}^{M}T(x;\Theta_m)$$ ■

算法 8.2 对回归树应用提升法

1. 对训练集中的所有的 i，令 $\hat{f}(x)=0$，$r_i=y_i$。
2. 对 $b=1,2,\cdots,B$ 重复以下过程：
   （a）对训练数据 (X,r) 建立一棵有 d 个分裂点（$d+1$ 个终端结点）的树 $\hat{f}^b$。
   （b）将压缩后的新树加入模型以更新 $\hat{f}$：

$$\hat{f}(x)\leftarrow\hat{f}(x)+\lambda\hat{f}^b(x) \tag{8.10}$$

   （c）更新残差：

$$r_i\leftarrow r_i-\lambda\hat{f}^b(x_i) \tag{8.11}$$

3. 输出经过提升的模型：

$$\hat{f}(x)=\sum_{b=1}^{B}\lambda\hat{f}^b(x) \tag{8.12}$$

例子（https://www.jianshu.com/p/7902b2eb5f21）：

例 8.2 已知如表 8.2 所示的训练数据，x 的取值范围为区间[0.5,10.5]，y 的取值范围为区间[5.0,10.0]，学习这个回归问题的提升树模型，考虑只用树桩作为基函数.

表 8.2 训练数据表

x_i	1	2	3	4	5	6	7	8	9	10
y_i	5.56	5.70	5.91	6.40	6.80	7.05	8.90	8.70	9.00	9.05

训练提升树的步骤：

- step1 构建第一个回归树T1(x)

由表 8.3 可知，当$s=6.5$时$m(s)$达到最小值，此时$R_1=\{1,2,\cdots,6\}$，$R_2=\{7,8,9,10\}$，$c_1=6.24$，$c_2=8.91$，所以回归树$T_1(x)$为

$$T_1(x)=\begin{cases}6.24, & x<6.5\\ 8.91, & x\geqslant 6.5\end{cases}$$
$$f_1(x)=T_1(x)$$

也就是说，我们现在得到了第一颗回归树，T1（x）。对于小于6.5的数据，我们把他预测成6.24，对于大于等于6.5的数据，我们把它预测成8.91。

然后，**就到了最重要的一步，将残差数据放入下一个回归树进行训练。**

用$f_1(x)$拟合训练数据的残差见表 8.4，表中$r_{2i}=y_i-f_1(x_i)$，$i=1,2,\cdots,10$.

表 8.4 残差表

x_i	1	2	3	4	5	6	7	8	9	10
r_{2i}	−0.68	−0.54	−0.33	0.16	0.56	0.81	−0.01	−0.21	0.09	0.14

下面去训练下一个学习器：

第 2 步求$T_2(x)$. 方法与求$T_1(x)$一样，只是拟合的数据是表 8.4 的残差. 可以得到：

$$T_2(x)=\begin{cases}-0.52, & x<3.5\\ 0.22, & x\geqslant 3.5\end{cases}$$

$$f_2(x)=f_1(x)+T_2(x)=\begin{cases}5.72, & x<3.5\\ 6.46, & 3.5\leqslant x<6.5\\ 9.13, & x\geqslant 6.5\end{cases}$$

以此类推

最后：

$$f_6(x)=f_5(x)+T_6(x)=T_1(x)+\cdots+T_5(x)+T_6(x)$$
$$=\begin{cases}5.63, & x<2.5\\ 5.82, & 2.5\leqslant x<3.5\\ 6.56, & 3.5\leqslant x<4.5\\ 6.83, & 4.5\leqslant x<6.5\\ 8.95, & x\geqslant 6.5\end{cases}$$

用$f_6(x)$拟合训练数据的平方损失误差是

$$L(y,f_6(x))=\sum_{i=1}^{10}(y_i-f_6(x_i))^2=0.17$$

假设此时已满足误差要求，那么$f(x)=f_6(x)$即为所求提升树.

2.关键参数

（1）树的总数B。生成的树总数过大可能会过拟合。交叉验证选择。

（2）压缩参数λ：控制提升法的学习速度，一般去0.01或0.001。若λ很小，则需要很大的B来实现良好的预测效果。

（3）每棵树的分裂点数d：控制提升模型的复杂性。d=1通常效果好。

第九章 支持向量机

2018年12月13日 19:06

1.超平面（hyperplane）在p维空间中p-1维的平面仿射子空间。例如在二维空间中超平面是直线，三维空间中超平面是平面。

方程：

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

X代入式子 > 0或 < 0表示X位于超平面的两侧。

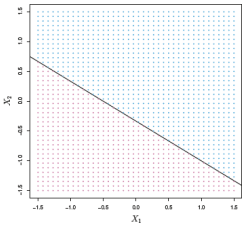
2.用超平面分割：

假设 $\boldsymbol{X}$ 为 $n \times p$ 的数据矩阵，由 p 维空间中的 n 个训练观测组成，

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \cdots, x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix} \qquad (9.5)$$

这些观测分为两个类别，即 $y_1, \cdots, y_n \in \{-1, 1\}$，

将下图蓝色的样本标记为y=1的类别，紫色的样本标记为y=-1的类别。



则

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \text{ if } y_i = 1,$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \text{ if } y_i = -1.$$

即

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$$

所以对于预测样本，

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*.$$

的值越大，则表示样本距离分割超平面很远，分类准确的把握很大。

3.间隔（margin）：所有训练样本到一个特定分割超平面的垂直距离的最小值。

4.最大间隔超平面（Maximal margin hyperplane）：也叫最优分离超平面（optimal separating hyperplane），离训练样本最远的分割超平面，即训练样本到分割超平面的间隔最大。

5.支持向量（support vector）：落在间隔上或间隔间的样本。分割超平面只与支持向量有关，与其他观测无关。即落在间隔以外的正确样本变化不会影响到分割超平面的位置，分割超平面的判定只由训练样本的一个小子集（支持向量）确定。

这些点“支持”着分割超平面的生成，所以叫“支持向量”。

9.1最大间隔分类器（Maximal Margin Classifier）

9.1.1思路

用最大间隔超平面来分割。

9.1.2构建方法——怎么找到最大间隔超平面

构建最大间隔分类器就是下面三个问题的优化解

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{maximize}} \; M \tag{9.9}$$

$$\text{subject to} \; \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.10}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n. \tag{9.11}$$

其中M为正数。M就是间隔。

约束条件9.10推出第i个样本到超平面的距离为

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}).$$

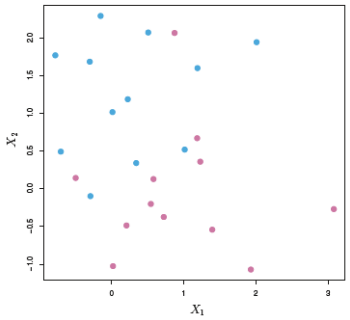
约束条件9.11假设M为正数就是要求每个样本个超平面保持一定距离。

因此，9.10和9.11保证了每个样本都落在超平面的正确一侧，并且与超平面的距离至少为M。优化问题就是找出最大化M（间隔）的参数

$$\beta_0, \beta_1, \ldots, \beta_p$$

9.1.3局限

最大间隔分类器对单个样本变化及其敏感，说明它可能过拟合了训练数据。而且在大多数情况下并不存在分割超平面，这种情况下那三个优化问题无解。比如下图（线性不可分）。

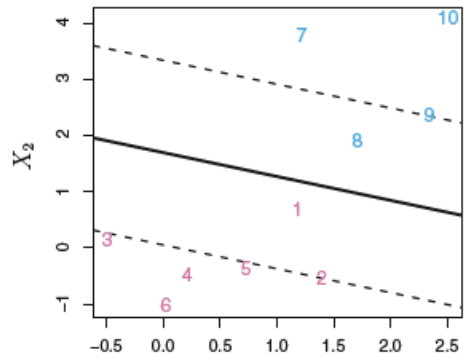


9.2支持向量分类器（support vector classifiers）——d=1的多项式核函数

9.2.1思路

采用软间隔（soft margin）解决最大间隔分类器中线性不可分的情况。即允许某些训练样本越过间隔落在间隔错误的一侧甚至是超平面错误的一侧。

注：间隔错误的一侧：下图的点8和点1。



9.2.2构建方法——怎么找到最大间隔超平面并允许软间隔

构建支持向量分类器就是下面四个问题的优化解

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \quad M \tag{9.12}$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.13}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i), \tag{9.14}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C, \tag{9.15}$$

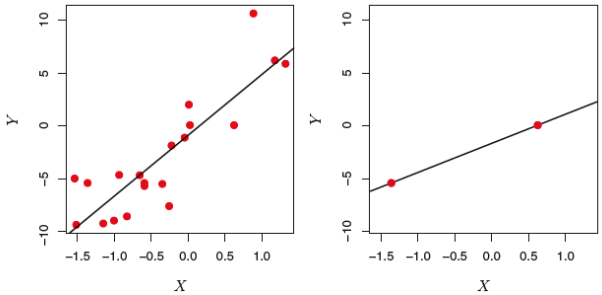
用ε_i放松限制，用C调节放松的程度。

1.ε_i：松弛变量（slack variable），表示第i个样本的位置。$\varepsilon_i=0$则第i个样本落在间隔正确的一侧；$0 < \varepsilon_i < 1$则落在间隔错误的一侧（穿过了间隔）；$\varepsilon_i > 1$则落在超平面错误的一侧（穿过了超平面）。通过式9.14反映。

2.C：非负的调节参数，表示可以多大程度地容忍样本穿过间隔和超平面的数目和严重程度。C=0则不允许任何样本穿过间隔，也就是最大间隔分类器；C > 0则只有不超过C个样本可以落在超平面错误的一侧，因为如果第i个样本落在超平面错误的一侧，就会有$\varepsilon_i > 1$，而优化要求

$$\sum_{i=1}^{n} \epsilon_i \leq C.$$

C通过交叉验证选择，表示偏差-方差的平衡。C较大（ε_i大，1-ε_i小，则M大，间隔大）允许穿过间隔的样本比较多，即支持向量多，这种情况下超平面的确定涉及较多样本，低方差高偏差。而如果C较小，允许穿过间隔的样本比较少，支持向量就会比较少，低偏差高方差。这与线性规划相似，用到的支持向量多，可以认为是拟合用到的有影响数据多，比较精确；用到支持向量少，可以认为是拟合用到的有影响数据少，高度拟合个别点，偏差小方差大。（跟下图类似）



9.3支持向量机（support vector machines）

9.3.1SVM综述

当决策边界是非线性时，最大间隔分类器和支持向量分类器效果差。在线性规划的问题中，可以使用预测变量的函数扩大特征空间，比如多项式，这里类似，用核函数扩大空间，然后优化。比如可以用2p个特征X_1，X_1^2，X_2，X_2^2，.....X_p，X_p^2或X的其他函数形式或者引入交互项得到支持向量分类器，这样优化问题变成了下面这几个，就能得到非线性的决策边界了。

$$\underset{\beta_0,\beta_{11},\beta_{12}\ldots,\beta_{p1},\beta_{p2},\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} M \tag{9.16}$$

$$\text{subject to } y_i\left(\beta_0+\sum_{j=1}^{p}\beta_{j1}x_{ij}+\sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right)\geq M(1-\epsilon_i),$$

$$\sum_{i=1}^{n}\epsilon_i\leq C,\ \ \epsilon_i\geq 0,\ \ \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2=1.$$

在扩大的特征空间中，计算的决策边界实际上是线性的，在原始特征空间中是非线性的。对于X_1，X_1^2，X_2，X_2^2，.....X_p，X_p^2或其他形式的X函数来说是线性的，对于原始的X_1，X_2，.....X_p来说是非线性的。见作业2.（d）。所以核函数就是把原始特征空间扩大，将本来非线性解在扩大的空间中变成线性的（趋于简单化的思路）

9.3.2思路

如上所述，用核函数扩大特征空间，然后优化得到边界。

通过各种证明，可以得到支持向量机的函数形式为

$$f(x)=\beta_0+\sum_{i\in\mathcal{S}}\alpha_i K(x,x_i).$$

αi：每个训练样本对应一个α，估计参数α_i和β_0只需n个训练样本的所有n(n-1)/2个成对组合，而又可以证明，只有支持向量对应的α非零，支持向量样本的集合用S表示

K：代替内积，称为核函数（Kernel），衡量x_i之间相似性的函数，有多种形式。比如简单的核函数

$$K(x_i,x_{i'})=\sum_{j=1}^{p}x_{ij}x_{i'j},$$

就是支持向量分类器，多项式核函数

$$K(x_i,x_{i'})=(1+\sum_{j=1}^{p}x_{ij}x_{i'j})^d.$$

就是用d阶多项式扩展特征空间。还有其他比如常用的径向核函数（radial kernel）（径向核函数如何起作用：参考书P244)

$$K(x_i,x_{i'})=\exp(-\gamma\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2).$$

注：用到的证明：支持向量分类器的优化问题的解只涉及变量的内积，而不是变量本身。

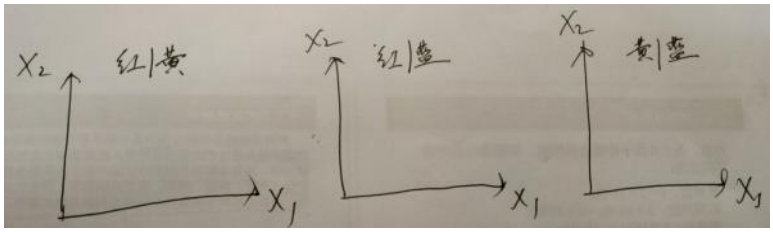
注意：核函数的参数取值（比如径向核函数的γ）可能会有过拟合问题。

9.4多分类的SVM

9.4.1一类对一类（one-versus-one）

1.思路

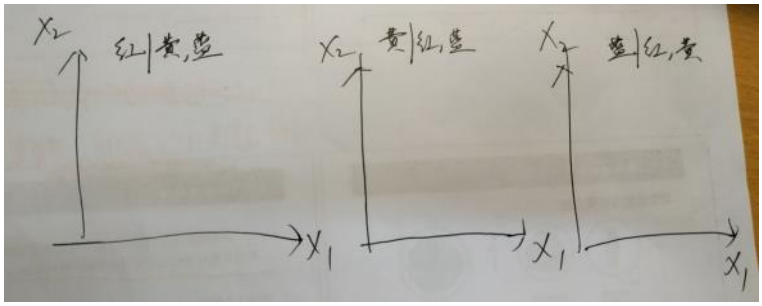
一共K类响应变量，构建K(K-1)/2个SVM，每个SVM分隔两个类别。使用所有K(K-1)/2个SVM对一个测试变量进行分类，记录这个测试样本被分到每个类别的次数，最终预测类别就是预测次数最多的那一类。比如：



9.4.2一类对其余（one-versus-all）

1.思路

一共K类响应变量，构建K个SVM，每个SVM分隔K个类别中的一个类别和其余K-1个类别。分类时这K个SVM依次对相应单独的那个类别回答是或不是，最后得到的类别即为所属类别。比如：



9.5SVM与logistic回归的关系

9.5.1SVM代价函数形式

1.可以证明，拟合支持向量分类器

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

可以改写成

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

可见代价函数依旧采取了**“损失函数+惩罚项”**的形式：

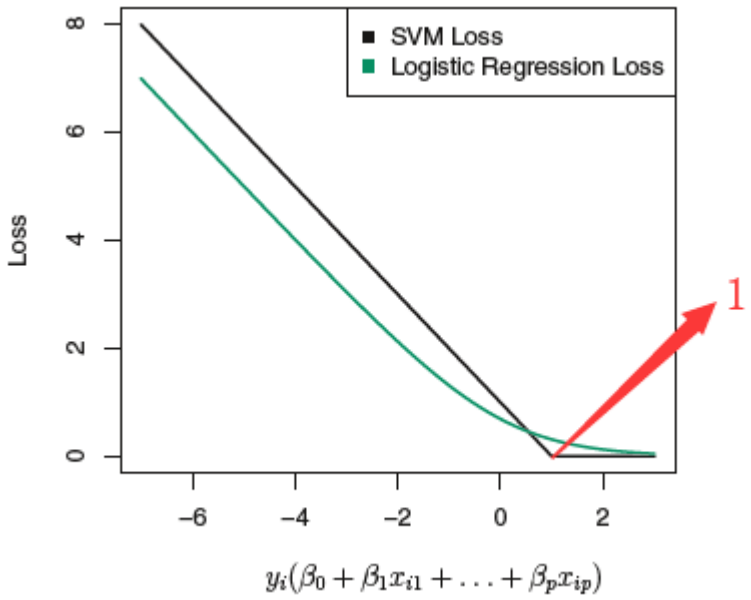
$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ L(\mathbf{X}, \mathbf{y}, \beta) + \lambda P(\beta) \right\}.$$

在式（9.26）中，$L(\mathbf{X}, \mathbf{y}, \beta)$ 是损失函数，是模型对数据的拟合程度的某种量化，其参数为 β，待拟合的数据为（$\mathbf{X}$，$\mathbf{y}$）。$P(\beta)$ 为惩罚函数，其参数向量为 β，β 的效应由非负的调节参数 λ 控制。例如，岭回归和 lasso 采用的损失函数都是

$$L(\mathbf{X}, \mathbf{y}, \beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

岭回归的惩罚函数为 $P(\beta) = \sum_{j=1}^{p} \beta_j^2$，而 lasso 的惩罚函数为 $P(\beta) = \sum_{j=1}^{p} |\beta_j|$。如果写成式

2.SVM的损失函数称为铰链损失（hinge loss）



对于铰链损失+惩罚项这种形式，间隔（M)对应的值是1。间隔的宽度取决于
$\sum \beta_j^2$.
这里1后面的损失为0是因为对于满足
$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq 1;$$
的样本来说，是落在间隔以外得到正确分类的样本。也正因为1后面的损失函数为0，所以有了间隔以外被正确分类的样本并不影响分类器这个特点。

9.5.2SVM与logistic的比较

1.当不同类别的样本可以很好地被分离时， SVM 的表现比logistic回归好。但如果不同类别存在较多重叠，选择logistic回归比较合适。

2.SVM 还可以扩展到回归中(响应变量是定量变量而不是定性变量)，称为支持向量回归（support vector regression）

第十章 无指导学习

2018年12月14日 17:02

1.主成分分析（principle components analysis，PCA）：用于数据可视化以及在有指导学习方法之前对数据进行预处理（之前章节的作为派生变量，降维）。

2.聚类分析（clustering）：探索数据，在数据集中划分子类。注意这需要定义不同样本间差异多少划分为同一类，而这个问题需结合具体要解决的问题背景经数据分析后方能得到答案。

3.本章讨论以特征为基础对样本的聚类。只需将数据矩阵转置就可以实现以样本为基础对特征的聚类。

10.1主成分分析：见第六章降维法

10.2聚类分析

10.2.1K均值聚类（K-means clustering）——自上而下的方法

1.思路

首先要确定想要得到的类数K。

用C_1，C_2，……C_k表示在每个类中的样本集合，其满足：

(1) $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \cdots, n\}$。即每个观测属于$K$个类中至少一个类。

(2) $C_k \cap C_{k'} = \emptyset$ 对每个 $k \neq k'$都成立。即类与类之间是无重叠的：没有一个观测同时属于两个类或更多类。

思想是让同一类的类内差异尽量小，用$W(C_k)$度量差异，则问题转化为最小化式子

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}.$$

$W(C_k)$比较经常采用平方欧式距离，则上式可表示为

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \qquad (10.11)$$

$|C_k|$表示第k类中样本的数量。

解决这个最小化问题用下面这个算法：

1. 为每个观测随机分配一个从 1 到 K 的数字。这些数字可以看做对这些观测的初始类。
2. 重复下列操作，直到类的分配停止为止：
   (a) 分别计算 K 个类的类中心。第 k 个类中心是第 k 个类中的 p 维观测向量的均值向量。
   (b) 将每个观测分配到距离其最近的类中心所在的类中（用欧式距离定义“最近”）。

最后得到的是局部最优解（一开始分配到的初始类下的最优情况，所以称为“局部”）比如下图：

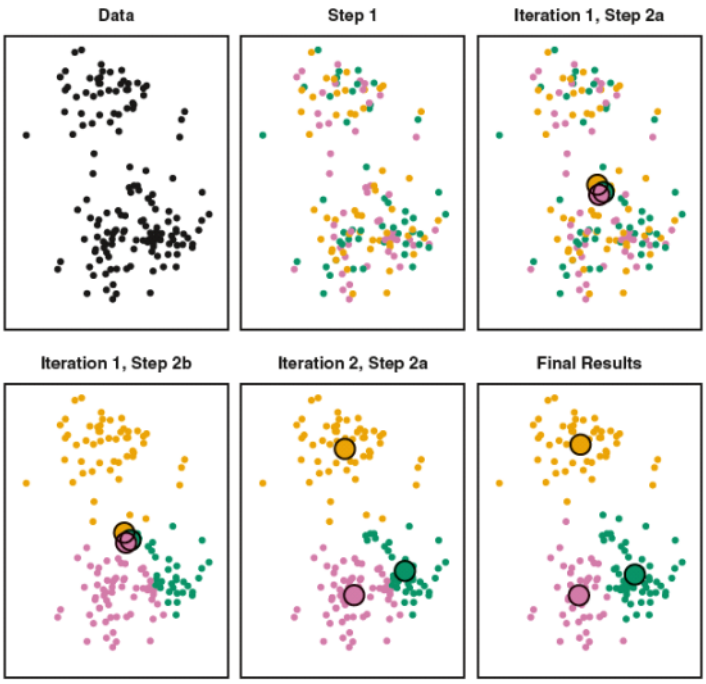
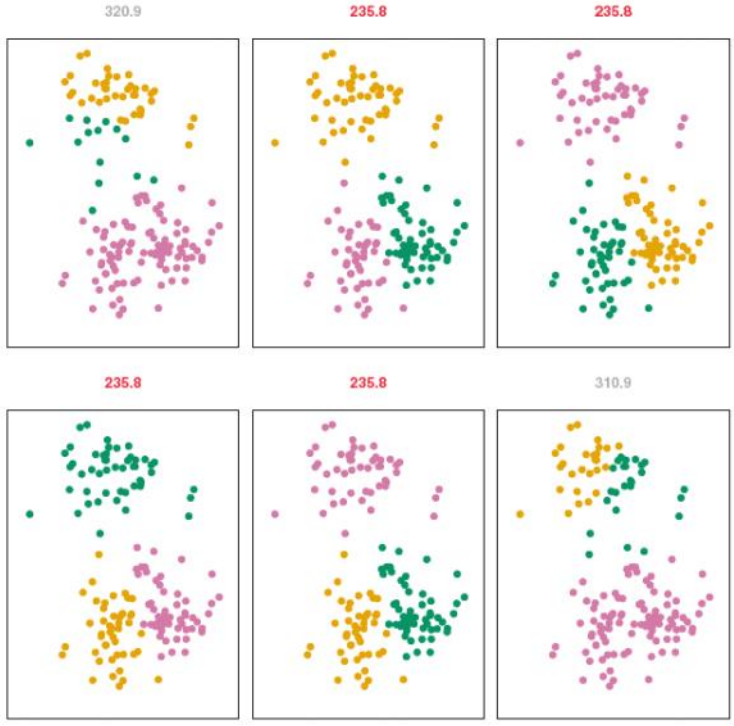


图 10-5 中 $K=3$ 时的 K 均值聚类过程。左上：原始观测点。上排中间：算法的第 1 步，每个观测被随机分配到一个类中。右上：第 2（a）步中类中心的计算，这些类中心在图中用彩色大圆片表示。可以看到，因为 K 均值聚类的初始类分配是随机的，所以初始的类中心几乎完全重叠。左下：在第 2（b）步中，每个观测被分配到了与之最近的类中。下排中间：第 2（a）步再次执行后得到了一个新的类中心。右下：10 次迭代后得到的结果。

对不同的随机初始类别状态运行这个算法，选择使式10.11值最小的方案。比如下图：



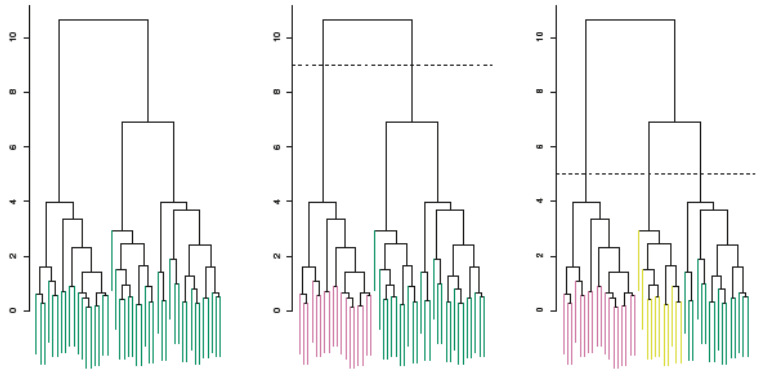
用图 10-5 中的数据运行 6 次 $K=3$ 的 K 均值聚类法后得到的结果。在每次运行时，K 均值聚类算法的第 1 步都会给每个观测随机分配一个不同的初始类。以上每幅图上方的数字表示聚类后目标函数（10.11）的一个值。我们得到了 3 种不同的局部最优解，其中一种局部最优解的目标值相对较小且提供了各类之间相对较好的划分。图上方数字为红色的聚类都实现了一致的最优分配，它们共同的目标值都是 235.8。

2.局限

必须先确定类数K，不好确定。

10.2.2系统聚类法（hierarchical clustering）——自下而上的方法，这里介绍最为常见的凝聚法（agglomerative）

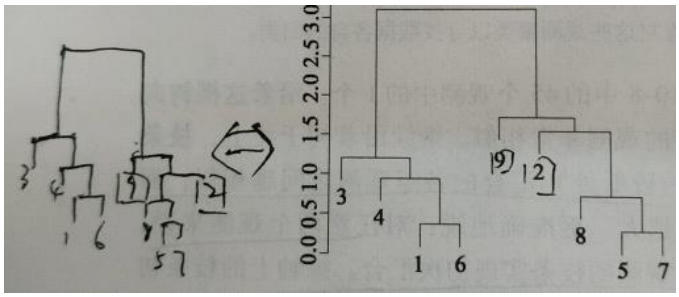
1.预备知识：谱系图（dendrogram）



最底下每片叶子代表每个样本。沿着这棵树向上，一些树叶开始汇入某些枝条中，这表示相应的样本非常相似。继续沿着树干往上，枝条本身也开始向叶子或其他枝条汇合。越早(在树的较低处)汇合的各组样本越相似，越晚(接近树顶)汇合的各组样本之间的差异越大。对于任意两个样本，都可以在树上找到一个点，在这个点上，包含这两个样本的枝条实现初次汇合。纵轴上枝条初次汇合高度表示两个样本的差异程度。

在谱系图上切割的高度控制着类的数量。比如上图分别得到一类、两类、三类。

注意：一个谱系图有另外2^{n-1}种可能排序，n是叶子（样本）数量。因为在n-1个汇合点上，2支已汇合枝条的位置可以交换，这并不影响谱系图的意义。因此，不能根据2个观测在横轴上的接近程度判断它们之间的相似程度，而需根据纵轴上包含两个样本的枝条第1次汇合的高度来判断这2个样本的相似程度。比如下面2和9：



2.思路

反复执行如下步骤：从谱系图的底部开始，n个样本各自被看作一类，然后将两个最为相似的类汇合到一起，得到n-1个类，然后再把两个最近的类汇合到一起，就得到了n-2个类。以此类推，直到所有样本都属于某一个类。（完成谱系图）

归类时需要用到差异度量指标。每两个数据之间的差异度量指标有欧式距离、基于相关性距离等；每两组数据的差异度量指标有最长距离法（complete linkage）、类平均法（average linkage）、最短距离法（single linkage）、重心法（centroid）等，用类平均法和最长距离法得到的类分布比较均衡。

1. 首先，计算 n 个观测中所有 $\binom{n}{2}=n(n-1)/2$ 对每两个数据之间的相异度（比如欧式距离），将每个观测看做一类。

2. 令 $i=n, n-1, \cdots, 2$：

（a）在 i 个类中，比较任意两类间的相异度，找到相异度最小的（即最相似的）那一对，将它们结合起来。用两个类之间的相异度表示这两个类在谱系图中交汇的高度。

（b）计算刚下的 $i-1$ 个新类中，每两个类间的相异度。

最后处理一张谱系图就可以得到任意数量的聚类。

例子：

例 1　设抽取五个样品，每个样品只测一个指标，它们是 1，2，3.5，7，9，试用最短距离法对五个样品进行分类。

（1）定义样品间距离采用绝对距离，计算样品两两距离，得距离阵 $D_{(0)}$ 如下：

	$G_1=\{X_1\}$	$G_2=\{X_2\}$	$G_3=\{X_3\}$	$G_4=\{X_4\}$	$G_5=\{X_5\}$
$G_1=\{X_1\}$	0				
$G_2=\{X_2\}$	1	0			
$G_3=\{X_3\}$	2.5	1.5	0		
$G_4=\{X_4\}$	6	5	3.5	0	
$G_5=\{X_5\}$	8	7	5.5	2	0

（2）找出 $D_{(0)}$ 中非对角线最小元素是 1，即 $D_{12}=d_{12}=1$，则将 G_1 与 G_2 并成一个新类，记为 $G_6=\{X_1,X_2\}$。

（3）计算新类 G_6 与其它类的距离，按公式：

$$G_{i6}=\min(D_{i1},D_{i2}) \qquad i=3,4,5$$

即将表 $D_{(0)}$ 的前两例取较小的一列得表 $D_{(1)}$ 如下：

	G_6	G_3	G_4	G_5
$G_6=\{X_1,X_2\}$	0			
$G_3=\{X_3\}$	1.5	0		
$G_4=\{X_4\}$	5	3.5	0	
$G_4=\{X_4\}$	7	5.5	2	0

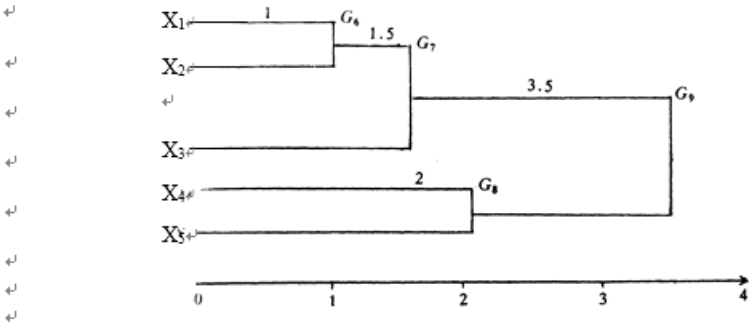
（4）找出 $D_{(1)}$ 中非对角线最小元素是 1.5，则将相应的两类 G_3 和 G_6 合并为 $G_7=\{X_1,X_2,X_3\}$，然后再按公式计算各类与 G_7 的距离，即将 G_3, G_6 相应的两行两列归并一行一列，新的行列由原来的两行（列）中较小的一个组成，计算结果得表 $D_{(2)}$ 如下：

	G_7	G_4	G_5
$G_7=\{X_1,X_2,X_3\}$	0		
$G_4=\{X_4\}$	3.5	0	
$G_4=\{X_4\}$	5.5	2	0

(5)找出 $D_{(2)}$ 中非对角线最小元素是 2，则将 G_4 与 G_5 合并成 $G_8=\{X_4,X_5\}$，最后再按公式计算 G_7 与 G_8 的距离，即将 G_4，G_5 相应的两行两列归并成一行一列，新的行列由原来的两行（列）中较小的一个组成，得表 $D_{(3)}$ 如下：

	G_7	G_8
$G_7=\{X_1,X_2,X_3\}$	0	
$G_8=\{X_4,X_5\}$	3.5	0

最后将 G_7 和 G_8 合并成 G_9，上述并类过程可用下图表达。横坐标的刻度是并类的距离。



由上图看到分布两类 $\{X_1,X_2,X_3\}$ 及 $\{X_4,X_5\}$ 比较合适，在实际问题中有时给出一个阈值 T，要求类与类之间的距离小于 T，因此有些样品可能归不了类。

最短距离法也可用于指标（变量）分类，分类时可以用距离，也可以用相似系数。但用相似系数时应找最大的元素并类，也就是把公式 $D_{ik}=\min(D_{ip},D_{iq})$ 中的 min 换成 max。

3.局限

系统聚类法通过切割谱系图得到的类必然嵌套与在更高高度切割谱系图得到的类中，但在实际中这个假设可能并不成立。比如有一组样本：男女各占一半，美国人、日本人、法国人各占三分之一。分成两类的最好办法是通过性别分类，分成三类的最好办法是通过国籍分类。但它们并不是嵌套关系。因此在这种情况下，系统聚类法得到的类代表性不强。因为当类数给定时，系统聚类法表现不佳，K均值聚类法更合适。

10.2.3聚类分析在实践中的注意点

1.聚类分析之前先回答下面几个问题：

（1）变量需要先经过某种标准化处理吗？比如变量中心化为均值为0，或标准化为标准差为1。

（2）应用K均值聚类法：数据分成多少类（K取多少）合适。

（3）应用系统聚类法：

a. 用什么指标衡量每两个样本间的相异度？

这个取决于聚类数据的类型和要解决什么问题。综合两者选择相异度指标。详见书P275

b. 用什么距离计算两组样本间的聚类？

c. 在谱系图的哪个位置切割？

2.验证聚得的类是否正确：现在有很多方法可以通过给每个类一个p值的方式评估聚类是否正确。但哪种方法最好还没有达成共识。

3.聚类分析的缺陷：

无论是 K 均值还是系统聚类都对每个观测实施了类别的分配，但有时这么分类可能并不合适。比如，假设大部分观测真实的类是小众（未知）子类，但这些观测中的小众群不仅彼此之间差异很大，而且与其他观测之间的差异也相当大。由于 K 均值聚类法和系统聚类法会强行地把每个观测分配到不同的类中，这使得找到的类的意义可能会因为不属于任何类的离群点的存在而受到严重扭曲。混合模型因能够允许离群点的存在使类免受干扰而获得人们的关注。混合模型相当于 K 均值聚类法的温柔版，Hastie et al. (2009) 中有这类方法的介绍。

另外，聚类分析方法一般不能很有效地处理受干扰的数据。比如，假设首先对 n 个观测进行聚类，然后随机删除由这 n 个观测的部分组成的子集，对余下的观测进行聚类。希望前后两次聚类得到的两组类非常接近，但常常事与愿违！

4.几点实用建议：

（1）数据标准化和距离选择会对结果产生巨大影响，因此建议用这些参数的不同取值进行多次聚类分析，统一进行比较，仔细观察有没有一些固定的模式始终存在。

（2）由于聚类分析有时候不那么有效，建议对数据的子集进行聚类，这样可以对所得到类的稳定性有一个整体的认识。

（3）聚类分析的结果不应被视为数据集的绝对真理，牢记无指导学习方法是对数据的探索。