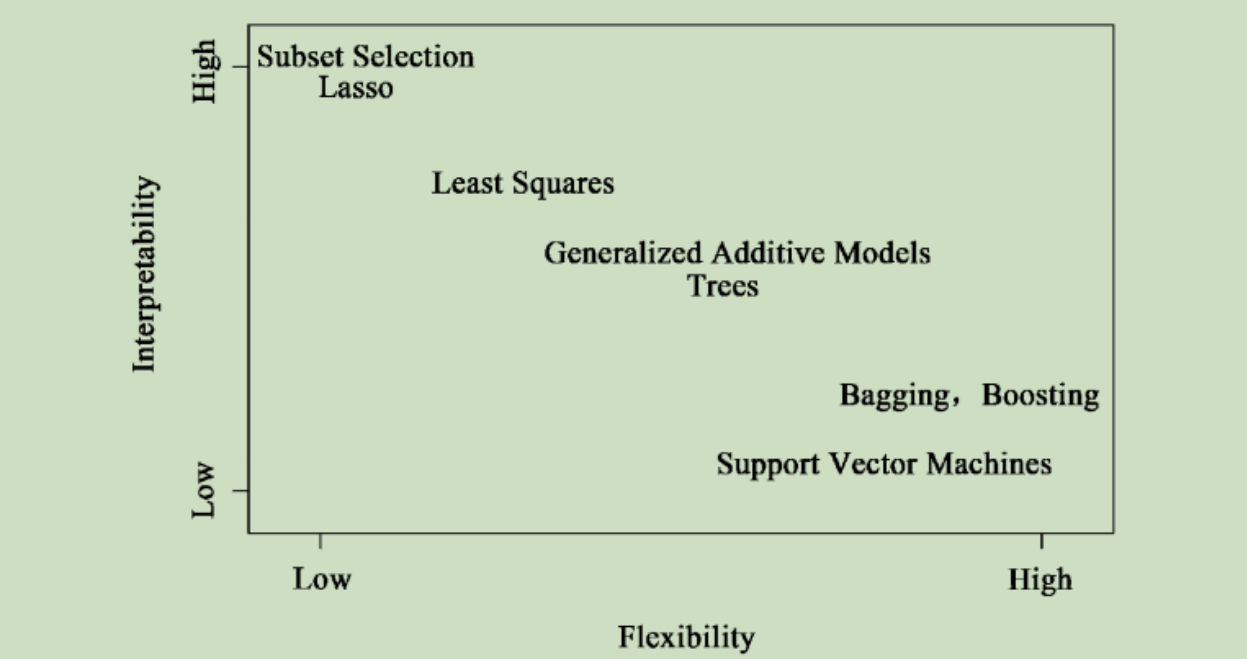
高级统计简答题汇总

2022.11.18 xhsioi

第二章

重要的图表以及公式：



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

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

$$1 - E\left(\max_j \Pr(Y = j \mid X)\right)$$

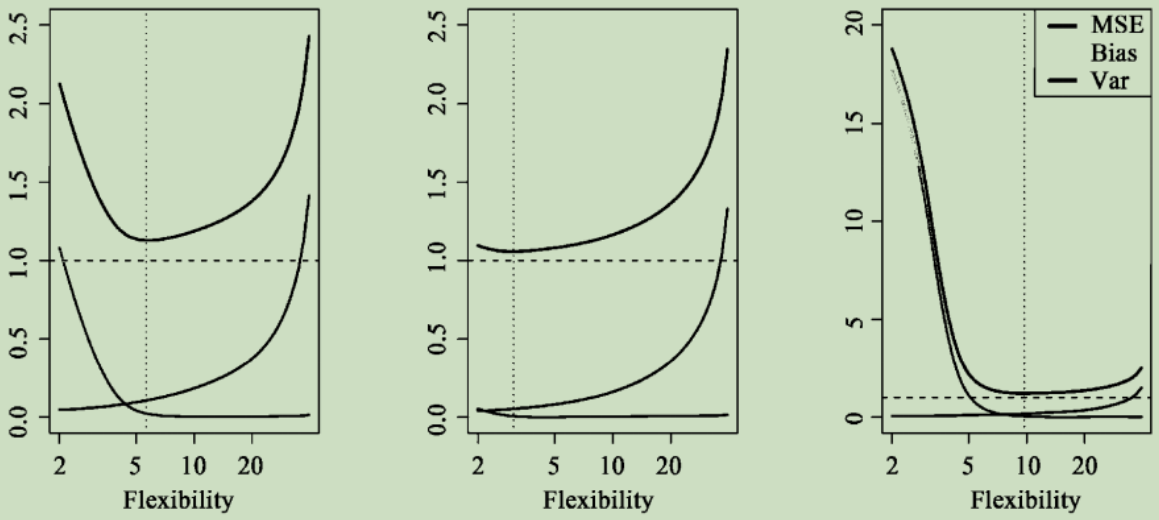


图 2-12　分别表示图 2-9 ~ 图 2-11 的三个数据集的平方偏差（蓝色曲线）、方差（橙色曲线）、不可约误差（虚线）、测试均方误差（红色曲线）。垂直的点线表示最小测试均方误差所对应的光滑度。

基本问题：

- 什么情况下需要进行估计f？46
- Y作为预测，其精确度依赖于哪些量？46
- 如何区分推断和预测？47
- 推断中常用的基本问题有哪些？48
- 如何利用均方误差计算可约误差和不可约误差？47
- 估计f的方法有哪些？
- 半指导学习的适用的数据模型为哪些？52
- 模型的拟合效果如何评价（针对回归类模型）？54
- 描述曲线光滑度的量是什么？56
- 光滑度和偏差、方差的关系？58
- 分类模型最常用的估计精度的方法是什么？59
- 贝叶斯错误率如何计算？
- KNN算法的实现步骤？61

- 光滑度高的模型优缺点是什么？适用的情况是什么？5
- KNN算法中当K逐渐增大，边界将如何变化？7

实验基本命令：62

- ls()
- rm()
- x=matrix(data = c(1,2,3,4), nrow=2, ncol=2, byrow=TRUE)（默认为false）
- x=rnorm(100, mean = . sd =)（默认呈现01正态分布）
- pdf(””) dev.off()
- x = seq(0, 1, length = 10)（创建0和1之间的等距的10个数序列）
- contour(x, y, f)（x值的向量、y值的向量、每对xy坐标上标记某个矩阵的元素）
- image(x, y, fa)
- read.table()
- fix()
- dim()
- names()
- hist()
- identify(x , y . name)

第三章 线性回归

图表以及公式：

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

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

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

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

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\mathrm{Cor}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

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

细节知识：

- 一个无偏估计不会系统地高估或者低估真实参数；
- 系数的解释：所有其他预测变量保持不变的情况下增加一个单位对Y产生的平均效果；
- 每个变量的t检验都等价于不含该变量，但包含所有其他变量的模型的F检验。
- 共线性降低回归系数估计的准确性，导致相关系数的标准误差变大；
- 预测区间的长度永远比置信区间宽；
- KNN的维数灾难问题；
- 对于非线性关系，测试集使用线性模型还是非线性模型得到的误差是无法比较的；
- 不含截距的线性回归构造；

相关问题：

- 线性回归经常讨论的一些问题？74（是否有关系、关系有多强、那些变量有关系、关系是否线性、预测精度如何、是否有协同作用、那种变量促进响应变化）
- 估计参数测量接近程度的常用方法是？76
- 总体均值的标准误差计算？残差的标准误差呢？
- 置信区间如何计算？（95 99）

- 如何判断预测变量和响应变量之间存在相关性？79
- t检验中p值是如何确定的？80
- 典型的临界p值有哪些？80
- 判断线性回归的拟合质量的标准？80
- 如何计算随机变量之间的相关性？82
- 为什么有些变量在简单线性回归中表现出很强的相关性，而在多元线性回归中显现出较低的相关性？84
- 如何判断多个响应变量和预测变量之间有关系？（多元分析）
- 如何选择重要的变量？87
- 拒绝向后选择的条件有哪些？87
- 可约误差和不可约误差的判断方式是什么？89
- 哑变量和水平数的关系？什么是基准水平？92
- 线性模型的假设有哪些？93
- 如何去除可加性假设？什么是实验分层原则？94
- 如何去除线性假设？95
- 线性模型存在的问题？96
- 如何精确地估计某一个预测变量对响应变量的影响？如何判断未来的预测精度呢？104
- KNN回归和分类的具体区别？105

实验基本命令：

- confint(model)
- predict(model , data)
- par(mfrow = c(x,y))
- hatvalues(model)　计算杠杆统计量；
- which.max(data)　识别数据中最大元素的索引；

- vif()　　计算方差的膨胀因子；
- anova(model1 , model2)　　比较两个模型的优劣程度；

第四章 分类

图表及公式：

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

$$\Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

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

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

$$\hat{\pi}_k = n_k / n \qquad (4.16)$$

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k \qquad (4.17)$$

细节知识：

- 在使用一个预测变量进行逻辑斯蒂回归时，如果其他预测变量与之有关系，那么预测模型会存在一定的风险（简单和多元出现不同的结果）
- 贝叶斯分界也是ovo的；
- QDA每一类观测都有自己的协方差矩阵，lda是假定每一类的协方差矩阵是相同的；

相关问题：

- 为什么定性变量的回归问题不能使用线性回归？124
- 逻辑斯蒂函数表达形式以及对数发生比的构造？其含义？125
- 估计系数的基本方法？126
- 不使用逻辑斯蒂回归的原因？129
- 线性判别分析实现的基本步骤？130
- 贝叶斯分类器的分类结果如何表达？（表达式）130
- 灵敏度、特异度、召回率和精确度的计算？136
- 为什么LDA的灵敏度这么低？134
- ROC图像的基本使用？136
- 第一类错误和第二类错误的区分？136
- QDA在实现过程中和lda存在哪些不同？137
- K近邻算法需要输入的多组参数有哪些？K值对分类的影响？147

第五章 重抽样算法

图表以及公式：

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \qquad (5.1)$$

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \qquad (5.2)$$

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \qquad (5.3)$$

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \qquad (5.4)$$

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r}\right)^2} \qquad (5.8)$$

细节知识：

- 在拟合过程中，保留训练观测的一个子集，然后对保留的观测运用统计学习方法估计测试错误率；
- 对于留一交叉验证法，对应的验证集MSE提供了对于测试误差的一个渐进无偏估计；
- 最小二乘法来拟合线性或者多项式回归模型时，LOOCV方法花费的时间将会缩减为之拟合一个模型的时间；

相关问题：

- 交叉验证常用的误差估计方法？155
- 验证集方法存在的问题有哪些？156
- 留一交叉验证的测试误差估计的均值如何计算？156
- 留一交叉验证和验证集方法的比较？157
- 使用交叉验证的目的是什么？是对MSE的真值感兴趣吗？159
- LOOCV方法和k折方法的异同点？如何进行偏差-方差的均衡？159
- 对于分类问题，交叉验证的评判标准是什么？160
- 使用KNN分类器时，使用训练错误率还是交叉验证错误率对K值进行选择？为什么？161
- 自助法适用的条件

实验相关操作：

第六章 线性模型选择与正则化

图表以及公式：

$$C_p = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2) \tag{6.2}$$

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

$$\mathrm{BIC} = \frac{1}{n}(\mathrm{RSS} + \log(n)d\hat{\sigma}^2) \tag{6.3}$$

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{6.5}$$

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j| \tag{6.7}$$

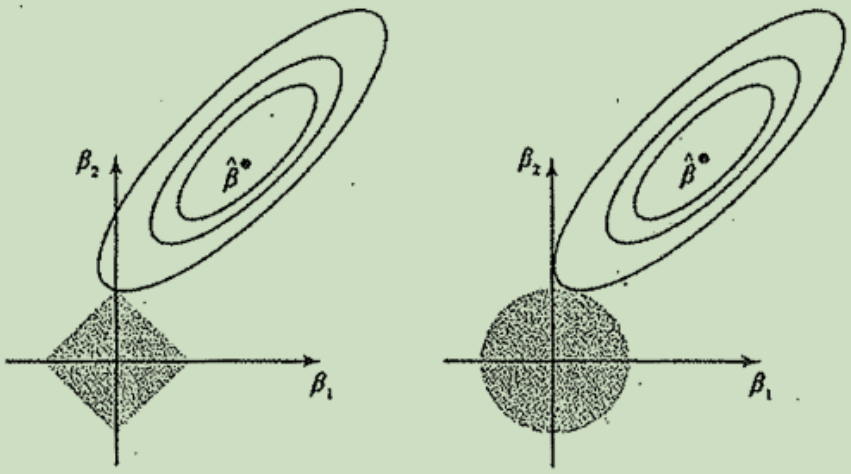


图 6-7　误差等高线和限制条件区域（左：lasso；右：岭回归）。实心区域是限制条件 $|\beta_1| + |\beta_2| \leqslant s$ 和 $\beta_1^2 + \beta_2^2 \leqslant s$，椭圆是 RSS 等高线。

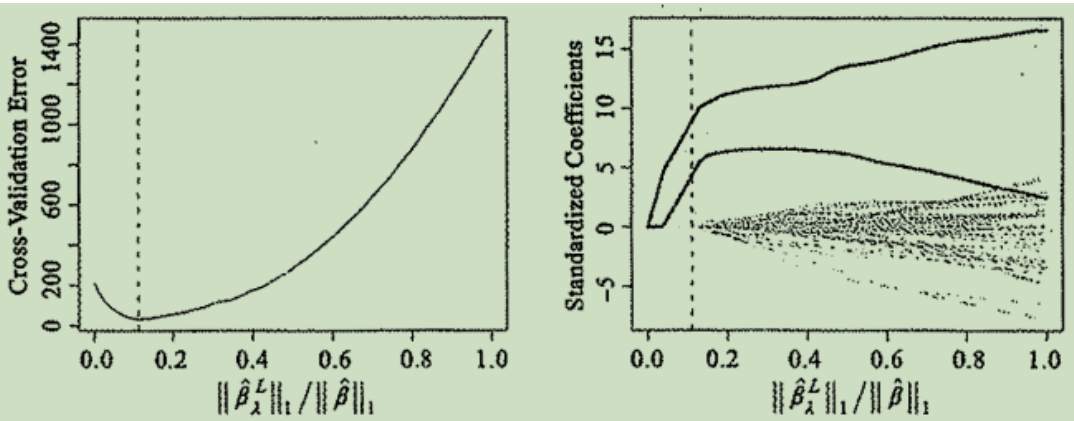


图 6-13　左：用 lasso 对图 6-9 稀疏模拟数据进行 10 折交叉验证的均方误差。右：相应的lasso系数。垂直虚线表示使得交叉验证误差最小的系数。

$$Z_m = \sum_{j=1}^{P} \phi_{jm} X_j \tag{6.16}$$

其中，ϕ_{1m}，ϕ_{2m}，…，ϕ_{pm}是常数，$m=1$，…，M。可以用最小二乘拟合线性回归模型

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \ i = 1, \cdots, n \tag{6.17}$$

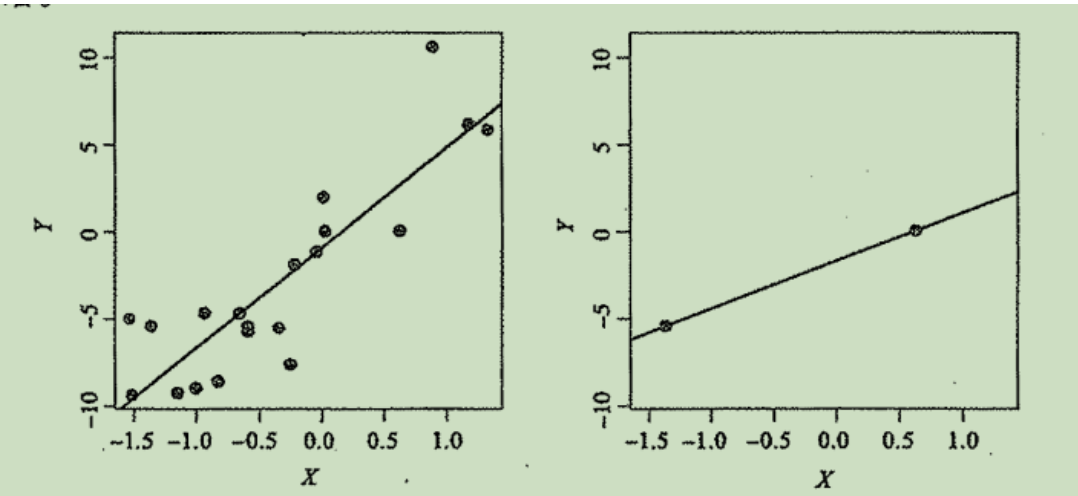


图 6-22　左：低维最小二乘回归；右：$n=2$ 个观测和两个系数估计（1 个截距和 1 个系数）的最小二乘回归。

细节知识：

- 不同的n和p的大小关系会影响向前逐步选择的最优子集方法的结果；

- 随着岭回归中的λ增大，岭回归估计值朝着0的方向缩减，当λ极大的时候，所有的岭回归系数估计值几乎为0，此时得到的模型相当于不包含任何变量的零模型；
- 对预测变量的参数乘以一个常数，可能会导致岭回归系数估计值产生显著的改变；
- 主成分分析和主成分回归在分析过程中没有考虑响应变量对主成分方向的影响；
- 偏最小二乘在计算第一主成分时，会将最大的权重赋给与响应变量相关性最强的变量；

相关问题：

- p和n的大小会对最小二乘的结果产生怎样的影响？最小二乘是如何减小计量方差的？173
- 对线性回归模型的拓展方法有哪些？基本原理是什么？173
- 最优子集方法的构建过程是怎样的？174
- 子集选择方法结果的评判标准是什么？174
- 逐步选择相较于最优子集的优点？176
- 向前选择和向后选择实现过程？176
- 为什么向后选择在n<p的情况下不成立？177
- 最优模型选择的方法有哪些？178
- 岭回归中为什么不对β0进行惩罚？181
- 为什么在使用岭回归之前需要对数据进行标准化处理？标准化的公式是什么？182
- 为什么岭回归优化课最小二乘的结果？182
- 为什么提出lasso回归？183
- 为什么lasso可以将系数估计完全压缩为0？185
- 岭回归和lasso回归的不同适用条件？187
- 岭回归的贝叶斯解释中对应的密度函数是什么？lasso呢？188
- 降维方法是如何实现的呢？190
- 降维常用的方法有哪些？191

- 主成分分析法需要进行标准化处理吗？偏最小二乘呢？195

第七章 非线性模型

图表以及公式：

$$\Pr(y_i > 250 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)} \tag{7.3}$$

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4 \tag{7.2}$$

$$\begin{aligned} C_0(X) &= I(X < c_1) \\ C_1(X) &= I(c_1 \leqslant X < c_2) \\ C_2(X) &= I(c_2 \leqslant X < c_3) \\ &\vdots \\ C_{K-1}(X) &= I(c_{K-1} \leqslant X < c_K) \\ C_K(X) &= I(c_K \leqslant X) \end{aligned} \tag{7.4}$$

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + \beta_K b_K(x_i) + \varepsilon_i \tag{7.7}$$

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & x > \xi \\ 0 & \text{否则} \end{cases} \tag{7.10}$$

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \tag{7.11}$$

$$\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y} \tag{7.12}$$

细节知识：

- 如果预测变量本身不具有明显的分割点，那么分段固定拟合就不十分恰当；
- 多次样条：在每个节点的后面添加截断幂积；
- 三次样条不能够解决边界区域方差较大的问题；

- 对于光滑样条的惩罚项，λ越大函数的光滑程度越高；
- 有效自由度就是矩阵Sλ的对角元素之和；
- 对于光滑样条求解λ，一般使用留一交叉验证进行计算，原因是只需要一个模型的计算量；

相关问题：

- 非线性回归有哪些类型？218
- 多项式回归最高阶数？为什么这样设置？218
- 三次样条的自由度是多少？222
- 三次样条函数的基本形式是什么？222
- 为什么使用自然样条？223
- 拟合样条的节点位置如何确定？节点的数量如何确定呢？223
- 相较于多项式回归，样条拟合的优点在哪？225
- 如何理解光滑样条中的惩罚项？226
- 为什么光滑样条更加注重有效自由度？226
- 局部回归的实现步骤？228
- GAM模型的优缺点？231

第八章 基于树的方法

图表以及公式：

$$\sum_{m=1}^{|T|}\sum_{i:x_i\in R_m}(y_i-\hat{y}_{R_m})^2+\alpha|T| \tag{8.4}$$

$$E=1-\max_k(\hat{p}_{mk}) \tag{8.5}$$

$$G=\sum_{k=1}^{K}\hat{p}_{mk}(1-\hat{p}_{mk}) \tag{8.6}$$

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{8.7}$$

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

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

(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：

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

细节知识：

- 树内部的各个节点的连接部分称为分支；
- 分类树的关注点：给定的观测值被预测为它所属区域内的训练集中最常出现的类；
- 决策树具有高方差的特点；
- 对于装袋法，每棵树能利用约三分之二的观测值。
- 装袋法得到的模型一般解释性较低；

相关问题：

- 回归树的建立过程？246
- 为什么要引入树的剪枝？本书采用的哪一种剪枝方式？246
- 回归树终端节点的判断条件是什么？247
- 代价复杂度剪枝的是如何实现的？相关的系数如何求解？248
- 分类树和回归树的评价标准是什么？249
- 基尼系数的大小对分类树的影响？互熵呢？249
- 线性回归和决策树的适用情况？树方法相较于传统方法的优点？251

- 树方法相较于其他回归和分类方法的缺点是什么？252
- 装袋法实现的基本原理？定性变量如何处理呢？253
- 随机森林方法的实现过程？255
- 应用提升法的回归树的构建过程？256
- 提升法的三个重要的调整参数？257
- 提升法应用的函数是什么？

第九章 支持向量机

图表以及公式：

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{9.1}$$

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

满足

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{9.10}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geqslant M, \quad i = 1,\cdots,n \tag{9.11}$$

$$\underset{\beta_0,\beta_{11},\beta_{12},\cdots,\beta_{p1},\beta_{p2},\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n}{\text{maximize}} M \tag{9.16}$$

$$\text{满足 } y_i\left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ji}^2\right) \geqslant M(1 - \varepsilon_i)$$

$$\sum_{i=1}^{n} \varepsilon_i \leqslant C,\ \varepsilon_i \geqslant 0,\ \sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1$$

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{9.17}$$

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{9.18}$$

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \tag{9.22}$$

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right) \tag{9.24}$$

$$\underset{\beta_0, \beta_1, \cdots, \beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i f(x_i)] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{9.25}$$

细节知识：

- 最大间隔分类器在线性不可分的情况下的推广叫做支持向量分类器，对支持向量分类器进行拓展得到支持向量机；
- 只有落在间隔上的观测以及穿过间隔的观测会影响超平面；
- 对于线性支持向量分类器而言，如果一个训练观测并不是支持向量，那么它对应的参数a为0；
- 一对一模型最终预测类别是预测次数最多的一类，一对多的模型最终预测的是边界函数相对最大的一类；
- cost越大，支持向量之间的间隔越小，容易产生过拟合

相关问题：

- 超平面是如何定义的？267
- 对于分隔超平面，其具有哪些性质？268
- 如何构造最大分类器，即它的约束条件有哪些？270
- 什么是软间隔？
- 最大分类器和支持向量分类器的支持向量相同吗？为什么？273
- 对于非线性分类情况，如何利用支持向量机进行分类？275
- 如何区分最大间隔分类器、支持向量分类器以及支持向量机？273
- 支持向量机如何处理多分类问题？279
- 一对一、一对多分别需要多少模型参与？具体的实现过程是怎样的呢？279

实验基本命令：

- svm()
- cost 惩罚因子，对最大间隔进行惩罚；
- tune（） 支持向量机中内置的交叉验证函数；

第十章 无指导学习

图表以及公式：

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 \qquad (10.12)$$

其中 $\bar{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$ 是第 C_k 类中第 j 个分量的均值。在第 2（a）步中，每个变量的类中心是

表 10-2　系统聚类法中的 4 种最为常用的距离形式汇总表。

距离方式	描　述
最长距离法	最大类间相异度。计算 A 类和 B 类之间的所有观测的相异度，并记录最大的相异度。
最短距离法	最小类间相异度。计算 A 类和 B 类之间的所有观测之间的相异度，并记录最小的相异度。最短距离法会导致观测一个接一个地汇合延伸拖尾的类。
类平均法	平均类间相异度。计算 A 类和 B 类之间的所有观测的相异度，并记录这些相异度的平均值。
重心法	A 类中心（长度为 p 的均值向量）和 B 类中心的相异度。重心法会导致一种不良的倒置现象的发生。

算法 10.1　K 均值聚类法

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”）。

算法 10.2　系统聚类法

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) 计算剩下的 $i-1$ 个新类中，每两个类间的相异度。

细节知识：

- 主成分的方向被截石位特征空间中原始数据高度变异的方向；
- 每个主成分的载荷向量都是惟一的；
- 主成分的得分向量可以用来预测变量的回归分析；
- PCA从观测中的低微表示解释更多的方差，而聚类分析想要从观测中寻找不同的类；
- 系统聚类：自上而下的凝聚法；

相关问题：

- 主成分的两种理解？296
- 主成分分析是否需要对变量进行标准化处理？297
- 主成分的数量如何确定？299
- 如何区分聚类分析和主成分分析？300
- 聚类分析有哪几种基本形式？
- K均值聚类的实现原理？系统聚类呢？302 307
- 倒置现象产生的原因？307
- 相异度的指标如何确定？308
- 系谱图的高度是如何确定的？

实验基本命令：

- promp()

- biplot()
- nstart 随机产生nstart个类的质心；
- hclust(,…, method = “complete/average/single”) 最长距离、平均、最短距离；
-