第十五章 数据预测建模实验

上一章讨论了数据的探查建模，也就是给一个数据集合，没有任何其它可以了解的知识，仅仅是希望探查这些数据有什么规律性的东西，也就是这些数据有什么样的结构组织形式，比如使用聚类分析就发现 wine 数据集中的所有样本根据其酒中的成分，可以聚为三类，每一类酒具有相似的成分。这就说明我们成功地从这批数据中探查到了一个模型，那就是这些不同类型的酒其实可以分为三类，如果要评估酒的营养成分，只要从三类中各抽取一个样本就可以了。

现在给出一个刚刚研发出来的新酒，判断它到底属于上面三类中的哪一类酒，这时候怎么办呢？我们立刻想到可以计算这个新酒和上面三类酒中心点的距离，距离那个中心点近就属于那类酒，这就是预测了，也就是给定一个样本，能够立刻推断出其所属类别。

信息处理的末端就是决策，决策的结果就是给出一个是和非的判断。我们在具体生产活动中得出大量样本数据，也知道它的性质，希望从这些数据中建立一个模型，以此可以预测给定新样本的性质。比如一个医院，获得了很多关于肿瘤的数据，判断肿瘤的指标有很多个，比如肿块厚度，细胞大小，细胞形状的均匀性，边际附着力，单个上皮细胞大小等等，根据这些指标获得一批样本数据，一些样本最后发现是恶性肿瘤，一些样本最后发现是良性肿瘤。这就意味着已经获得了关于这批数据的分类特征，现在来了一个病人，化验得出关于肿瘤的多个指标数据，那么如何判断这个病人的肿瘤是良性肿瘤还是恶性肿瘤呢？这就是预测了，也就是根据已有的样本所属类别，预测未知样本的所属类别。这种判断决策分类预测有点特殊，就是所有的样本只分为两类，对应实际含义就是“是”和“非”。现实生活中这类例子太多了：暑假来了，根据过去销售经验，哪类学生才会买笔记本电脑；银行放贷，根据用户过去的信用和资产等情况，哪些用户是潜在客户；邮件过滤，根据关键词，图像和超文本等数据，哪些邮件是垃圾邮件等等，这些都要做出判断，都是根据过去的知识，对给定的观测值做出“是”或者“非”的判断，进而进行决策。

对已知类别的样本建模，已经提出了很多算法，比如线性判别分类，KNN、SVM、朴素贝叶斯分类、人工神经网络、决策树、C5.0、随机森林、adaboost 和 bagging 等等许多种算法。这些算法构建分类模型的步骤一般都是先将已知样本数据分为两批，比如随机挑选三分二的样本作为训练样本进行建模，然后使用剩下的三分一的样本代入模型进行预测，看预测的分类结果和已知的分类是否一致，这就是分类验证，如果大多数都正确的分类了，那么这个模型就证明是有效的，可以使用了，如果大多数分错了，那么这个模型是无效的，可能要使用其它模型来建模了。本章实验的编写主要参考了文献[20]和文献[23]等。

第一节 逻辑回归

我们在统计分析实验中对回归分析做了实验。回归分析建立模型以后就可以根据自变量取值计算出因变量值，因此回归分析也应该归于预测建模类任务之列，只是为了讨论统计检验和分析，将其放到那个专题讲述而已。对于逻辑回归，因变量是类型变量，很明显就涉及到分类了，因此本节做逻辑回归的预测建模实验。

逻辑回归主要用于使用连续型或者类别型预测变量来预测二值型结果变量。它实质是一种广义线性模型，也就是根据数值变量预测二元分类，显然这时候的因变量就是二值型的类别变量。

拟合逻辑回归模型在 R 语言中使用函数 glm()。如果预测变量中有类别型变量，这时候 glm()函数就会自动将其编码为虚拟变量。下面以某国一个地区的乳腺癌数据为例来建立逻辑回归模型。

这个乳腺癌数据集共有 699 个样本，可以选其中 70%为训练样本来建立预测模型，30%的样本来验证模型是否有效，当然这种选择应该是随机的。具体数据集可以使用下面的 R 语言到网上下载：

```
loc = "http://archive.ics.uci.edu/ml/machine-learning-databases/"
ds = "breast-cancer-wisconsin/breast-cancer-wisconsin.data"
```

```
url = paste(loc, ds, sep="")
breast = read.table(url, sep=",", header=FALSE, na.strings="?") #以上四行就是到 uci 网站读取数据
```

下面的语句是对读取的数据进行预处理

```
names(breast) = c("ID", "clumpThickness", "sizeUniformity", "shapeUniformity", "maginalAdhesion",
                  "singleEpithelialCellSize", "bareNuclei", "blandChromatin", "normalNucleoli", "mitosis", "class")
df = breast[-1]  # 将第一列去除，因为第一列是样本的编号，对我们数据处理没有意义
df$class = factor(df$class, levels=c(2,4),labels=c("benign", "malignant")) #将 class 因变量变成 factor 数据类型
```

下面语句对预处理后的数据进行随机抽样

```
set.seed(1234)
train = sample(nrow(df), 0.7*nrow(df)) # 抽取 70%的样本作为训练样本，抽取的是其序号
trainSet = df[train,]                  # 根据序号抽取具体的训练样本
validateSet = df[-train,]              # 去掉训练样本，剩下的就是验证样本。
table(trainSet$class)
table(validateSet$class)
```

上面语句中使用 sample()函数进行随机抽样，70%抽查的样本放到 trainSet 中，剩下的放入 validateSet 中，通过 table()可以查看有多少样本是良性的，有多少样本是恶性的。运行上面代码，具体可以得到训练样本中良性 319，恶性为 170，而验证样本中良性为 139，而恶性为 71。下面通过函数 glm()建立逻辑回归模型，具体 R 语言为

```
fitLogit = glm(class~., data = trainSet, family=binomial())
```

解释一下 glm 中的参数含义：class~.意思是变量 class 是因变量此处指的是恶性肿瘤还是良性肿瘤，~后面的点表示数据集中剩下的其它变量都是自变量，具体自变量是哪些，可以查看 name 向量定义的字符串。“data=”是数据集的变量名。“family=”是因变量属于什么分布，在本实验中，对于肿瘤的发生，假设服从二项式分布。下面使用 summary()查看拟合逻辑回归模型的结果如下：

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24605  -0.08012  -0.03110   0.00266   2.11576
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -12.4412     2.0547  -6.055  1.4e-09 ***
clumpThickness             0.7407     0.2262   3.275  0.00106 **
sizeUniformity            -0.0320     0.3399  -0.094  0.92500
shapeUniformity            0.2073     0.3715   0.558  0.57680
maginalAdhesion            0.5194     0.1708   3.041  0.00236 **
singleEpithelialCellSize  -0.3217     0.2613  -1.231  0.21831
bareNuclei                 0.5851     0.1881   3.111  0.00187 **
blandChromatin             0.8599     0.2923   2.942  0.00326 **
normalNucleoli             0.4036     0.1828   2.208  0.02725 *
mitosis                    0.8923     0.3552   2.512  0.01200 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 621.04  on 477  degrees of freedom
Residual deviance:  52.39  on 468  degrees of freedom
  (因为不存在，11 个观察量被删除了)
AIC: 72.39
Number of Fisher Scoring iterations: 9
```

观察上面回归系统的 Pr 值，很容易可以发现其中的指标变量 sizeUniformity，shapeUniformity 和 singleEpithelialCellSize 对回归方程的贡献都不显著，因为假设是不相关的，它们的检验值都大于 0.05，因此对回归方程都不显著，因此可以去除这三个因变量进行重新逻辑回归。如果仍然以 9 个变量参与建立了逻辑回归模型，那么现在就可以对 validateSet 的数据集进行验证了，也就是这些数据样本根据已经建立的逻辑回归模型，预测属于恶性肿瘤还是良性肿瘤。另外注意训练样本集合中有 11 个样本由于缺项被删除了，没有参加逻辑回归。

R 语言中的 preditct()函数用来预测 validateSet 数据集中肿瘤数据是良性还是恶性的，具体实现形式：prob = predict(fitLogit, validateSet, type="response")，参数 type=指定值为"response"，这表示该函数返回的结果默认为恶性的对数概率，我们当然可以取定阈值 0.5，将大于这个阈值的样本认为是恶性肿瘤，小于 0.5 的样本认为是良性肿瘤。这样就将得到的预测

分类和实际分类进行对比，生成列联表就可以观察分类准确情况，这个列联表也称为混淆矩阵。具体 R 语言实现代码如下：

```
prob = predict(fitLogit, validateSet, type="response")
logitPred = factor(prob > .5, labels=c("benign", "malignant"))
   #添加 labels，良性肿瘤名称在前的原因是它的值 prob <0.5,而 factor 是按照从小到大排列的
logitPerf = table(validateSet$class, logitPred,dnn=c("Actual", "Predicted"))
logitPerf
```

验证得到的混淆矩阵如下

```
           Predicted
Actual      benign   malignant
benign         129           6
malignant        1          69
```

可以观察到，有一个恶性肿瘤被误分类到良性，有 6 个良性被误分类到恶性中。可以计算分类准确率，就是(129+69)/205=97%。需要说明的是验证样本实际上为 210，而现在使用的验证样本为 205，这是因为其中的 5 个样本由于缺项没有参加验证。这个模型 97%的有效性还是可以接受的。

逻辑回归的时候，发现其中三个指标是不显著的，应该给以去除，进而精简模型，可以使用 step()函数逐步回归，去除不显著的指标，然后重新进行逻辑线性回归。具体 R 语言代码：logitReduced = step(fitLogit)，获得重建的模型，然后重新验证，得到结果如下

```
           Predicted
Actual       benign     malignant
  benign         130             5
  malignant        1            69
```

可以观察到，模型的确得到改进，其中良性误判恶性比没改进前少了一个。此处需要说明的是上面的实验由于样本挑选是随机的，尽管使用了 set.seed()函数，但是每次系统关闭重启或者不同的同学做实验，得到的结果可能不完全一样，因为随机抽取的训练样本不一样。

第二节 决策树

本节介绍经典决策树，主要用于根据预测变量预测二元因变量，这是经典决策树特点，其具体算法步骤就是(1)可以根据已知训练样本的分类情况，计算所有预测变量的信息增益，将具有信息增益最大的预测变量作为测试属性，创建一个节点，将所有样本按照该预测变量取值进行分组；(2)每一组样本数据继续按照(1)的步骤进行；(3)重复上面两个步骤；(4)结束的条件就是给定节点的所有样本都属于同一个类，或者没有预测变量了，或者分支没有样本。

这种决策树建模算法可能造成过拟合现象，过拟合现象出现的原因就是建模时将噪声和误差都当作真实信息参与分类，因此建立的模型用来预测样本数据，其效果会很差。为了解决这个问题，通常需要对生成的决策树进行剪枝。

我们还是以上一节的肿瘤为例，使用决策树对该肿瘤数据建立预测模型，即使用上一节训练样本构造决策树模型。R 语言中 rpart 包中的 rpart()函数可以实现决策树构建，prune()函数可以实现对构建的决策树进行剪枝，从而得到更优的决策树，下面的 R 语言实现具体决策树构建

```
library(rpart)
set.seed(1234)
dtree = rpart(class ~ ., data = trainSet, method = "class",parms=list(split="information"))
```

上面语句运行时，如果没有安装 rpart 包，则先安装该包。rpart()函数的第一个参数同样是自变量和因变量公式，class 是数据集 trainSet 中的 class 变量，它只取两种类型的值，也就是恶性肿瘤和良性肿瘤，“~”后面的点代表其它预测变量，data=后面是数据集，method=后面取值可以为 class 或者为 anova，这取决于因变量是否为 factor 类型。paras=后面是一个 list 列表，其中 split=后面是决策树对每一个预测变量对应的样本进行分割的方法，如果是 information，表示使用的是信息增益法，至于这些算法的思想，可以参考商务智能或者数据挖掘方面的文献。返回的结果 dtree 里面有一个数组， dtree$cp 即可查看该数组，如表 15.1 所示。可以使用 dtree$查看有哪些数据集合。

表 15.1：决策树返回的 cptable 内容

	CP	nsplit	rel error	xerror	xstd
1	0.81764706	0	1.00000000	1.0000000	0.06194645
2	0.04117647	1	0.18235294	0.1823529	0.03169642
3	0.01764706	3	0.10000000	0.1588235	0.02970979
4	0.01000000	4	0.08235294	0.1235294	0.02637116

这个表中 CP 表示复杂度参数的意思，复杂度指的是决策树分枝的多少，当然决策树分枝越多越复杂。此处的 CP 是决策树每一次分裂时最小的提升量。在决策树构建时，引进耗费复杂度(cost complexity)。具体公式为

$$CC = \sum_{\text{终端节点数目}} \text{错分类数目} + \lambda \times \text{划分次数}$$

其中λ是惩罚项。所谓划分次数如表 15.1 中 nsplit 所示，就是决策树大小，即分支数目。一个新的划分要满足 CC 取最小值，这个最小值就是 CP；rel error 对应的是训练样本对每一次划分构建的树所计算的分类误差；xerror 是对训练样本进行 10 折交叉验证误差，也就是将训练样本集分成 10 份，轮流将其中 9 分作为训练样本，一份作为测试样本进行实验，每次实验都得到误差率，10 次实验得到的误差平均值就是 10 折交叉验证误差；xstd 是 10 折交叉误差的标准差。

R 语言中有一个函数 plotcp()专门用来绘制复杂参数与交叉验证误差关系图，其参数就是 rpart()函数返回的决策树模型，即 plotcp(dtree)，绘制的图形如图 15.1 所示。

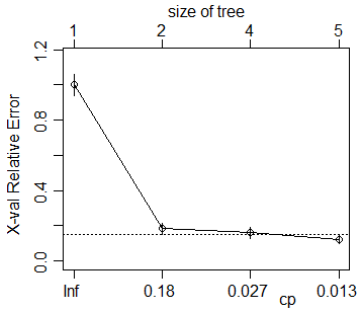

图 15.1： CP 和 xerror 关系图

从图 15.1 可以看到，CP 取值 0.027，xerror 取值基本稳定，也就是决策树模型误差率基本不变，查阅表 15.1，可以发现此时 nsplit=3，构建的决策树模型最优，因此决策树可以选择 nsplit=3 时 CP=0.0176 进行剪枝。R 语言有一个函数 prune()实现剪枝功能，具体语句为：

```
dtreePruned = prune(dtree,cp=0.0177)
```

返回的 dtreePruned 就是决策树模型，可以使用 rpart.plot 包中的函数 prp()来绘制最终的决策树，如果没有安装包，则需要先安装。函数 prp()的参数特别多，可以通过 help()来查看其参数含义。运行语句 prp(dtreePruned,type=5,varlen=0)得到图 15.2。

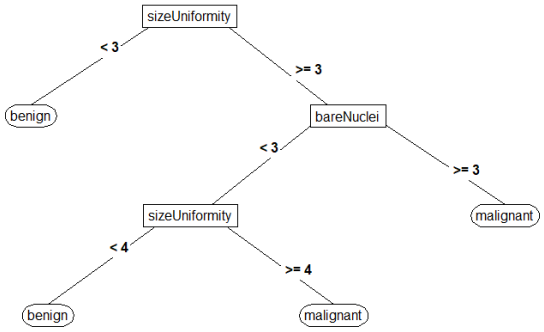

图 15.2： 修剪后肿瘤数据的决策树模型图

注意修剪语句中，如果 cp 的值选取小于 nsplit=3 时对应的 cp 值，那么就无法实现修剪，也就是得到的决策树模型保持不变，此时的决策树模型如图 15.3 所示。

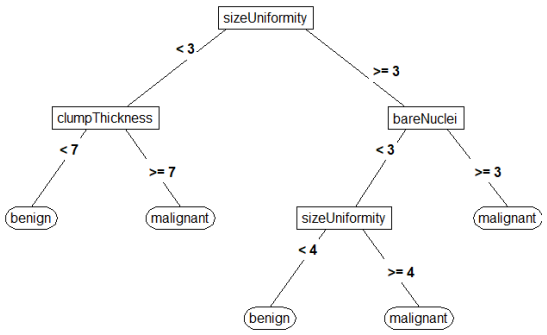

图 15.3：没有修剪的肿瘤数据决策树模型图

比较图 15.2 和图 15.3，可以发现，图 15.2 少了一个 clumpThickness 变量的分割。建立修剪后的模型，可以对测试样本进行分类预测了。具体 R 语言代码如下

```
dtreePred = predict(dtreePruned, validateSet, type="class")
dtreePerf = table(validateSet$class, dtreePred,dnn=c("Actual", "Predicted"))
dtreePerf
```

运行的结果如下

```
Actual      benign malignant
  benign       129        10
  malignant      4        67
```

可以发现该模型的预测准确率不如上一节的逻辑回归准确率高，这就说明不同的算法对不同的样本，有效性是不一样的，这正是研究的价值，也是提出很多各种分类算法的原因。另外需要注意的是在逻辑回归中，由于缺项，只有 205 个预测样本参与验证，而决策树却是 210 个样本全部参与验证，原因就是只要在建立的决策树模型中，分类属性值不缺项的样本就可以参与分类验证。

第三节 综合实验十五

一、实验目的

1. 掌握逻辑回归模型实现给定样本的预测
2. 掌握基于决策树模型的数据分类和预测

二、实验平台和软件

1. 互联网平台
2. 计算机系统
3. Rstuio 环境和 R 语言编译系统

三、实验内容和步骤

1. 逻辑回归建模

(1) 根据实验指导，读取 uci 网站中的乳腺癌数据集

(2) 生成乳腺癌 dataframe 数据结构

(3) 取生成的乳腺癌数据集中 70%样本为训练集，30%样本为验证集合

(4) 使用 glm()函数进行逻辑回归

(5) 使用构建的逻辑回归模型对验证样本集进行预测

(6) 计算该模型在样本验证集上的分类准确率

(7) 去掉不显著的预测变量，重新逻辑回归和验证预测

2. 决策树建模

(1) 使用上面的乳腺癌数据进行决策树建模

(2) 绘制 cp_xerror 表，寻找最佳 cp 取值

(3) 实现对生成的决策树剪枝

(4) 绘制剪枝后的决策树图形

(5) 使用优化的决策树模型对验证变量进行预测计算

(6) 计算该决策树模型在验证集上的分类准确率

四、实验报告（总分 8 分）

1.使用 read.table 直接读取给定链接的数据集 http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/australian/australian.dat，抽取 70%和 30%分别作为训练集和验证集，以数据集中 v15 为影响变量，对其它变量进行逻辑回归，优化回归模型，预测验证集，给出验证结果和精确度估计(注意 read.table 以 sep=空格字符串读取，read.table 返回的数据类型为 dataFrame，每列名称系统自动命名为 V 加上列号，注意这个列号的 V 是大写的) (3 分)。说明：这是一个信用卡数据集，将上面的 dat 修改为 doc 可以下载文档说明。

2.对于 iris 数据集，使用决策树建模，70%为训练集，30%为验证集。要求 rpart()函数使用 control=rpart.control(minsplit=2, cp=0)作为该函数其中的参数生成决策树，然后分析剪枝，生成最优决策树用来测试验证集，给出模型准确率 (3 分)

3.记录实验过程，分析实验问题和解决办法 (2 分)