**注意：**

每个用户R中的设置不同，在运行我提供的程序时，如果读文本文件没有将字符类型（character）自动转换成factor类型，从而造成程序错误，建议在read.table，read.csv函数中加入参数stringsAsFactors =T

例如，df<-read.csv('data/house-votes-84.data',header=F,stringsAsFactors =T)

1. P11

在2.2.4.1节后面补充

R可以使用冒号产生连续的整数序列，也是一个向量。冒号两边的整数是序列起始和结束的整数值。例如

x<-1:3

也可以是

x<-c(1:3)

向量x中的值是

1，2，3

1. P16

P16第3行print(x[x==1]<-25) 改为：

x[x ==1]<-25

print(x)

1. P46

图2.17的绘制程序有错误。见本书github网站上传的程序“图2-17.R”。

1. P57

3.3.1节下面

改为：

如果数据集中存在有缺失值的数据记录（R通常用指示符NA表示，含义是不可用），或者存在Inf（无穷数）、NaN（表示不是一个数）等指示符都应该进行处理。我们简单的用缺失值来表示它们。缺失值的处理通常有下面的几种方法：

1. P60

3.4.3.2 节最后补充

使用scale函数，实现z-score规范化的代码如下：

|  |
| --- |
| m<-matrix(c(1,2,3,4,5,6), nrow=2, byrow=TRUE)  msd<-apply(m,2,sd)  mmean<-apply(m,2,mean)  scale(m,center=mmean,scale=msd) |

1. P60

加3.4.3.3节

（3）最大值规范化

即将每列的数据除以该列的最大值。

1. P63

第二段 y=f(x)，x应该是小写

1. P67

P67 “会达到99.9990％的准确率”

改为：

会达到99.999％的精确度

1. P72

改为：决策树是第7章介绍的集成技术中随机森林的基础模型。

1. P74

KNN算法描述里的 “规范化”前面加一个注释符

1. P77

决策树描述算法中没有缩进

1. P77

决策树算法下面的描述中，增加一句

“创建决策树的算法是一个递归算法，即在算法build\_tree的第13步又调用了算法本身”

1. P78

公式5.6上边

InfoA（D）是属性A对数据集划分后的子集

改为

InfoA（D）是按照属性A的属性值对数据集划分后计算的每个子集信息期望Info(Dj)的加权和，权重是|Dj|/|D|。v是属性值的数量，是划分后的子集。

1. P78

公式5.7 上边

就是属性A划分数据集后的子集

改为

就是属性A的熵。即按照属性A的属性值对数据集的划分后计算的信息期望。v是属性值的数量，是划分后的子集。

1. P85

P85 程序中的第二条删除

deeper.ct.pred.train <- predict(deeper.ct,train.df,type = "class")

P85 倒数第二行

ｃｐ 参数设为 ｃｐ ＝ ０􀆰 ０００ ０１

改为：cp参数设为cp=0

1. P87

P87 在“得到误差均值23270。 房价的均值是 188037。” 这一段后面加上一句

下面的代码对CART的构造过程进行了控制，可以和默认参数时构造的CART比较一下性能。

1. P89

P89 “（１）特征 ｘｉ 为数值类型时” 这一行

在训练数据的特征 ｘｉ 上

改为：在训练数据属于类别的样本上计算特征的均值和方差。

1. P91

P91 5.6.1节上面一段，修改为“详细程序见本书github个人网站上的三个文件：adult\_cart\_C50.R，adult\_KNN.R, adult\_NB.R和adult\_test.R。

P91 表5.1 列ID， 12-15改为 V12-V15

1. P93

P93 当采用KNN建立模型，需要的预处理的可能选项包括：

第3条，后面加上

第4条“形成原始数据集”删除

相应的表5.4中的预处理步骤（4）去掉。

1. P114

P114 第3行，“（6）建议使用neuralnet包所有的…”，改为“（6）建议使用neuralnet前，所有的…”

1. P120

P120 6.4.2 数据集预处理，下面一行 “以及因子类型的列是否有其他符号”，改成“以及字符类型的列是否有其他符号”

下面的程序

df<-read.csv('hmeq.csv')

改成

df<-read.csv('hmeq.csv', stringsAsFactors=T)

1. P125

7.1.1节下第5行：“袋装算法从训练集Ｄ中对数据有放回抽样。”

改为：

“设数据集D中有d个数据样本，袋装算法从训练集Ｄ中有放回抽样d个数据样本，去除重复数据得到数据子集Dk。统计学上证明，此时数据集中不重复的数据样本大小约为0.63\*d”

算法袋装中的第二步，“通过对D有放回抽样，创建大小为d的数据集Di”

改为：

“通过对D有放回抽样d个样本，创建不包含重复数据的数据集Di”

1. P126

倒数第4行，err(Mi)的值在0~1之间

改为

error(Mi)的值在0~1之间

1. P132

P132 在介绍函数randomForest的参数Importance。后面的varImpPlot(modle)改为，varImpPlot（建立的随机森林模型）

1. P133

倒数第9行，即“提升树的预测过程如同7.7所示”下面加上

现在要预测一个家庭中的每个成员是否喜欢玩计算机游戏。每个家庭成员的特征包括：{年龄，性别，是否使用计算机}

倒数第8行，“如果新来的数据是一个男孩” 后面加上，

他的特征是：{年龄<18, 男，使用计算机}，他会得到2.9的评分

P133 最后一行，“完成分类任务时”

后面增加，“的取值是1或-1”

1. P137

P137 倒数第三行里的程序

原： modle<- xgb.train(data=xgb.DMatrix(train, label=y),

改为：

modle<- xgb.train(data=xgb.DMatrix(train, label=y), params=params，

1. P143

P143 最下面的程序的第13行

“idx<-sample(nrow(df),100)test<-df.onehot.data[idx,]”

应该是两条语句

idx<-sample(nrow(df),100)

test<-df.onehot.data[idx,]

1. P144

公式（8.1）下面第3行，

改为

当前数据集中两个类别C和N的数据个数分别是TC=300， TN=700。

1. P145

P145 程序的第19行

adult<-read.table("adult.data",sep=",",header=F, strip.white=TRUE)

应该单独占一行

1. P149

P149 程序下面补充一句

“支持向量机中，所有训练集中所有特征的值被标定为均值为0，方差为1的范围（有的SVM实施中，标定为[-1,+1]或[0,1]的范围）。”

1. P153

P153 算法ADASYN

第二步，应该为“意味着两类完全平衡”

第三步，应该为

为中的每条数据 计算

1. P168

P168 算法第3步，“重复步骤1” 改为

“重复步骤1和2”

1. P170

P170 公式9.14

从R4.0版本开始，factoextra包的get\_clust\_tendency函数使用的是公式9.13而不是9.14计算霍普金斯统计量。因此该值越大，表示聚类趋势越明显。

1. P183

P183 最后一行，“%pin%部分匹配”后面增加

%ain%多个项都匹配

1. P185

P185 10.3.3节下面的代码，第一行插入一条

data(Groceries)

1. P186

P186 图10.1 下面的第三行，前插入一段代码

irules<-apriori(Groceries,parameter=list(confidence=0.4,supp = 0.01, minlen=2))

1. P190

P190 11.1.1节下面第3行

原：“直接处理p值”

改：“直接处理概率p”

1. P191

P191 公式（11.2）和（11.3）改为

或者

1. P195

P195 “以womenroles数据集为例，”

后面加上一句，“以womenroles数据集为例，假设读入后的数据帧名为w”

1. P225

P225 算法Hits的第5步，改为

即满足