Business Analytics Homework 2

Instructor: Zach Zhizhong ZHOU

1. 在 D 盘建立目录：D:/BA/Homework/HW02，将附件中的数据放入该目录中，以该目录做为工作目录。打开 Accidents.xls，可以看到数据和数据的说明，不过下面我们将在 Accidents.csv 上操作。数据集包含 42,183 条 2001 年发生在美国的车祸记录。每条车祸记录的伤亡报告有 3 种可能：NO INJURY，INJURY 或者 FATALITY。车祸记录还记录下其他信息包括：车祸发生在周几，天气状况，路况等。我们的目标是预测车祸是否有伤亡：如果有则原数据集中的 MAX_SEV_IR 等于 1 或者 2，如果无则原数据集中的 MAX_SEV_IR 等于 0。

   打开 Accidents.R，运行以下语句：

```
library(class)
library(plyr) #needed for arrange
library(e1071) ## needed for Naive Bayes
setwd("D:/BA/Homework/HW02")
Accidents <- read.csv("Accidents.csv",header = TRUE)

#将 MAX_SEV_IR 等于 1 或者 2 的值全部设置为 1。
Accidents[Accidents[,'MAX_SEV_IR']>0,'MAX_SEV_IR'] = 1
#将 1-13,15-17,19-20,22-24 列转换成类别型变量。
AA <- c(1:13,15:17,19:20,22:24);
for (i in 1:length(AA)) {   Accidents[,AA[i]] <- factor(Accidents[,AA[i]]) }
```

   这时候我们得到一个数据框 Accidents。

1.1) 取出 Accidents 数据集的子集前 12 条记录，只取 WEATHER_R 和 TRAF_CON_R 为预测变量，然后在数据子集新加一列 Count，取值为 1。该数据子集保存为 AccSub：

```
AccSub <- Accidents[c(1:12),c('WEATHER_R','TRAF_CON_R','MAX_SEV_IR')]
AccSub <- data.frame(AccSub,COUNT=rep(1,nrow(AccSub)))
```

接下来使用 aggregate 函数做分类汇总，统计 WEATHER_R、TRAF_CON_R、MAX_SEV_IR 三个变量取值分布情况，统计值为 MAX_SEV_IR 取值的计数。例如：WEATHER_R=2，TRAF_CON_R=0 情况下，MAX_SEV_IR=0 出现了 5 次，则统计值为 5（如下表）。汇总之后使用 arrange 函数对数据框进行排序，先根据 WEATHER_R 再根据 TRAF_CON_R 从小到大排序。分类汇总表应该如下：

	WEATHER_R	TRAF_CON_R	MAX_SEV_IR	COUNT
1	1	0	0	1
2	1	0	1	2
3	1	1	0	1
4	1	2	0	1
5	2	0	0	5
6	2	0	1	1
7	2	1	0	1

1.2）根据上表计算：如果使用精确贝叶斯分类器，那么当 WEATHER_R=2，TRAF_CON_R=0 时，MAX_SEV_IR=1 的概率是多少?

1.3）运行下面的语句，你将得到给定 MAX_SEV_IR 取值(0 或者 1)，WEATHER_R(或者 TRAF_CON_R)在各个类别的计数值，最后是 MAX_SEV_IR 取值在 12 条记录中的分布。

```
attach(AccSub)
table(MAX_SEV_IR,WEATHER_R)
table(MAX_SEV_IR,TRAF_CON_R)
table(MAX_SEV_IR)
detach(AccSub)
```

使用上面 3 个表计算：如果使用朴素贝叶斯分类器，那么当 WEATHER_R=2，TRAF_CON_R=0 时，MAX_SEV_IR=1 的概率是多少?

1.4）下面使用 naiveBayes 函数建立朴素贝叶斯分类器。可以看到 AccSub 第 2 条记录是 WEATHER_R=2，TRAF_CON_R=0。检查 predict 函数计算结果，确保结果和上题的计算结果相同。

```
AccSub[2,]
Predictors <- AccSub[,c('WEATHER_R','TRAF_CON_R')]
Target <- AccSub[,'MAX_SEV_IR']
classifier<-naiveBayes(Predictors, Target)
Probs <- predict(classifier, Predictors, type='raw',threshold=0.01)
Probs
```

接下来设定截值为 0.4：MAX_SEV_IR 等于 1 的概率如果严格大于 0.4，则将该条记录分类为 MAX_SEV_IR=1，否则分类为 0。

```
AccSub <- data.frame(AccSub,PredictInj = rep(0:1,6)) #在 AccSub 中新增加一列，列名为 PredictInj，初始值为 0,1,0,1…
```

你需要根据 Probs 中计算出 MAX_SEV_IR 等于 0 和 1 的概率，在 AccSub 数据集的 PredictInj 列填写对 MAX_SEV_IR 的预测值（注意：截值为 0.4）。处理结束之后在命令行输入 AccSub[,'PredictInj']并回车应该得到以下结果：

```
> AccSub[,'PredictInj']
 [1] 1 0 0 0 1 0 0 1 0 0 0 0
```

要求：必须使用条件判断语句让 R 自动填写而不能使用人工判断然后对 PredictInj 列逐行赋值。

最后根据 PredictInj 列的预测值和 MAX_SEV_IR 列的实际值给出混淆矩阵。

1.5）运行以下 R 代码：

```
AccidentModel <-
Accidents[,c('HOUR_I_R','ALIGN_I','WRK_ZONE','WKDY_I_R','INT_HWY','RELJCT_I_R',
                                    'REL_RWY_R','TRAF_CON_R','TRAF_WAY','MAX_SEV_IR')]
set.seed(1000)
RowNum <- nrow(AccidentModel)
SampleIndex <- sample(1:RowNum,round(RowNum*0.8),replace = FALSE)
TrainData <- AccidentModel[SampleIndex,]
ValidationData <- AccidentModel[-SampleIndex,]
TargetIndex <- which(colnames(AccidentModel)=='MAX_SEV_IR')
Predictors <- TrainData[,-TargetIndex]
```

解释以上代码是什么意思。

1.6）使用运行上题代码得到的 Predictors 作为预测因子，使用 TrainData[, TargetIndex]作为结果变

量的真实值，建立朴素贝叶斯分类器。将建立好的贝叶斯分类器用于验证数据集的预测：验证数据集的预测因子是 ValidationData[,-TargetIndex]，预测结果存在 MyPredict 中。验证数据集目标变量 MAX_SEV_IR 的真实结果是 ValidationData[,TargetIndex]，使用它和 MyPredict 建立混淆矩阵。

1.7) 如果仅知道训练数据集结果变量 MAX_SEV_IR 的每个取值，即 TrainData[,TargetIndex]，而没有任何预测因子的信息，那么应该对验证数据集的结果变量 MAX_SEV_IR 作何预测？

1.8) 比较前面 2 题预测错误率，一是使用朴素贝叶斯分类器的预测错误率，一是不使用任何预测因子信息进行预测而得的错误率，哪个错误率较小？使用朴素贝叶斯分类器有用吗？

1.9) 注意：本题是附加选做题，不做不会扣分，做对有加分。
进行预测时，错误把本应该是 MAX_SEV_IR=1 的划为 0 的成本为 1，错误把本应该是 MAX_SEV_IR=0 的划为 1 的成本为 1.2。问题：将 MAX_SEV_IR 划为 1 的概率值的截值是多少时，对验证数据集进行预测的成本最小？此时预测错误率为多少？

2. 用课程文件夹/BA/Homework/HW02 中 fgl.csv 的数据，使用玻璃的折射系数（RI - refractive index）以及玻璃中的元素含量推断玻璃的类型。玻璃有 6 种类型，分别是 WinF（float glass window）、WinNF（nonfloat window）、Veh（vehicle window）、Con（container 比如玻璃瓶）、Tabl（tableware）、Head（vehicle headlamp）。玻璃中元素含量考虑 8 种元素，分别是 Na、Mg、Al、Si、K、Ca、Ba、Fe。

2.1) 建立 KNN 模型对玻璃类型进行分类。要求：使用 train 函数对 KNN 模型进行训练，尝试的 k 值是 1 到 8，评价指标是准确率，使用 5-fold 重复交叉检验，该检验重复 20 次。训练结果得到的最佳 k 值是几？画出不同 k 值下的 KNN 模型的预测准确率以及给出最终的模型的混淆矩阵。

2.2) 如果使用 Naive Bayes 方法处理相同的任务，我们不能直接使用 fgl.csv 给出的数据。如果确实想使用 Naive Bayes 方法处理相同的任务，我们应该对数据做什么预处理呢？（你不需要写代码，只需要给出处理数据的想法即可）

3. 小马手头有 20 个上市公司 CEO 年薪的数据。这些 CEO 的年薪可划分为高薪和低薪 2 类，每类 10 人。他们的年薪可能受以下因素影响：股东权益与总资产的比值、股价年收益率、每股收益、资产收益率。小马打算用手头的数据预测某个不在数据集当中的某个上市公司 CEO 拿到的是高薪还是低薪。老马听了小马的打算之后，告诉小马说某种数据挖掘方法适用于他面临的问题，但使用该方法前需要检视一下现有数据看是不是所有的数据都能用。

3.1) 老马说的这种方法是什么方法？

3.2) 为何需要检视数据看是不是所有的数据都能用？

3.3) 小马应该使用什么方法检视数据？

4. 画二维图给出一个二元分类的例子，该例的分类问题非常适合使用决策树算法，可以达到 100%的准确率，但该例却非常不适合使用 kNN 算法。在图中应该使用不同的点形状（point style）或者颜色表示不同分类。你需要解释为什么这个例子适合使用决策树算法进行分类，却不适合使用 kNN 算法。你可以在 Word 上画电子版的图，也可以在纸上画好图然后拍照发电子版做为附件之一通过 Email 发给 TA，也可以在 5 月 25 日课前提交纸质版给我。

提交作业的截止时间：**5 月 25 日周四下午 16:00**。
应该通过电子邮件提交到 zachzhoucourse@hotmail.com。提交作业的邮件标题是："BA HW02 你的名字"。
提交文件列表：

1、对题目的解答，将解答放在一个 Word 或者 PDF 文档里面，将文档命名为 HW02-你的名字。
2、解答题目的完整代码。
以上代码确保放在以下目录当中可以成功执行：D:/BA/Homework/HW02。如果你修改工作目录，导致代码无法在上面的目录中执行，TA 将不会检查你的代码而是直接给你 0 分！