Business Analytics Homework 3

Instructor: Zach Zhizhong ZHOU

回答以下的题目，你需要提交：1、一个 Word 文件回答题目中问到的问题或把图形输出结果贴贴到 Word 文件中，2、针对每个题目编写的 R Script 文件。你的 R Script 文件和数据文件应放在 D:\BA\Homework\HW03 文件夹内测试和运行，确保我把你写的 R Script 和你使用的数据文件放在我自己 PC 的 D:\BA\Homework\HW03 文件夹中也可以运行。如果你的电脑没有 D 盘，那么代码和数据应该可以放在 C:\BA\Homework\HW03 当中运行，不过你需要在 Word 文件第一行就特别说明这一点。

1、数据文件 RealEstate.csv 中列出了国内几个城市在 2012 年的房价均值（AvgPrice）、人均 GDP（GDP_PC）、人均月工资（AvgMSalary）、人均月可支配收入（AvgMNetIncome）和 CPI 指数（CPI）。请根据以上指标使用聚类分析将数据文件中的 26 个城市分成 4 类。要求：画出树图、并指出各个城市分属哪一类。 如果将城市分成 2 类，那么以上 4 类城市将如何进行类别合并？

2、数据文件 Segmentation.csv 被用来进行市场细分（Market Segmentation），其中 Class 列表示消费者类型，其余 58 列被用来预测消费者类型。（提示：如果你把 58 个预测因子名称都写出来将会是非常大的工作量，但有非常简单的方法可以让你避免把 58 个预测因子名称全部写出来）。请使用现有数据建立一个决策树模型。要求：每个树叶节点至少包含 100 个数据才能继续长出枝叶，把决策树画出来，写出决策树给出的至少 3 条规则。

3、数据文件 NN.csv 中有 3 列，分别是 x.1，x.2 和 y。建立一个用 x.1 和 x.2 来预测 y 的人工神经网络（Neural Network）模型。将 500 个数据使用随机抽样的方法分成 350 个数据用于训练模型和 150 个数据用于验证模型（提示：使用 sample 函数）。在随机分组之前执行 set.seed(1)命令。要求：

a. 人工神经网络的输出节点对加权和做 Logistic 函数处理，隐藏层有 3 个节点，输出错误对权重的偏导数小于 0.01。请把最终得到的人工神经网络画出来。将得到的人工神经网络模型用在验证数据集上并给出混淆矩阵。
b. 做与（a）相同的事情，不同之处是将隐藏层设置为 2 个节点。比较两个混淆矩阵的结果说明隐藏层有几个节点较好？

4、数据文件 Cosmetics.csv 记录了 1000 条女士购买化妆品的记录。一行表示一条消费记录。列名表示化妆品，取值 yes 代表购买了该化妆品，取值 no 则代表未购买。我们希望了解以下问题：

a. 如果某个女士购买了 Nail.Polish，那么在支持度 support 至少 0.1，信心水平 confidence 至少 0.5 的情况下该女士还会买什么化妆品？
b. 如果某个女士购买了 Mascara，那么在支持度至少 0.2，信心水平至少 0.5 的情况下该女士还会买什么化妆品？
c. 利用以上得到的任意一条规则解释支持度和信心水平是什么意思，应该如何计算？
d. 问题（a）和（b）如果只让你分别给出最有用的一条规则，那么它们是什么？

5、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。

a. 老马说的这种方法是什么方法？
b. 为何需要检视数据看是不是所有的数据都能用？
c. 小马应该使用什么方法检视数据？

6、以下是某投资银行人事部门 2 人的对话：

人事部门干事李先生：郑总，今年有 2000 人通过在线求职平台申请我们的 5 个岗位。我刚来公司，不太清楚我们这边如何筛选面试对象。

人事部门主管郑女士：你手头不是有去年的数据吗？去年也有 2000 人申请 5 个岗位，你分析一下去年的数据应该就知道我们怎么选人了。

人事部门干事李先生：但问题是去年的 5 个岗位和今年不同，而且今年我们提高了薪水，申请者的平均水平和去年相比有很大提升，比如有不少上海交大毕业的学生。

人事部门主管郑女士：那你能不能帮我先初步处理一下数据，我下午再过来看一下申请我们公司都有些什么样的人，不过我可不想去把那 2000 份申请材料全都读一遍。

a. 如果按照郑女士的建议去分析去年的数据，应该使用什么方法？预期得到什么结果？
b. 李先生要准备下午给郑女士阅读的报告，应该使用什么方法？预期得到什么结果？

7、在 R Script 中输入 library(fpp); data(euretail) 调入名为 euretail 的数据集。

a. 时间序列画图：

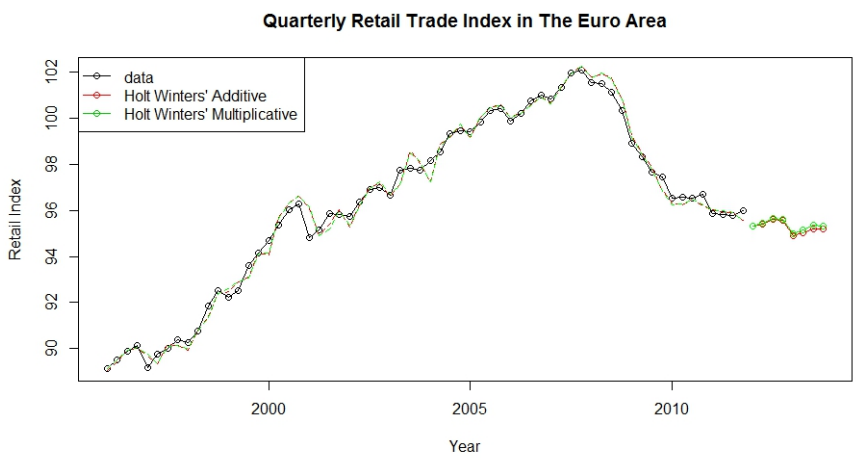
画出时间序列 euretail，x 轴名称为 Year，y 轴名称为 Retail Index，整个图的名称为 Quarterly Retail Trade Index in The Euro Area。

b. 分解时间序列：

分别使用 multiplicative form 和 additive form 将时间序列进行分解并画出分解之后各模块的图。

c. Holt-Winters 指数平滑：

对时间序列 euretail 做 Holt-Winters 指数平滑并画图，应得到以下的图：



注意：如果画图的时候 legend 函数生成的图例说明太大，可以在 lengend 函数当中加上 cex=0.8 或者 cex=0.5 调整图例大小。

8、数据文件 hmohiv.csv 记录了对艾滋病患者的观察记录，其中 time 代表观察到的时间长度（月），status 代表状态（1=死亡，0=数据删失），age 代表年龄，drug 代表有无吸毒史（1=有，0=无）。使用这个数据文件分析：

a. 使用 Kaplan-Meier 方法画出生存函数。
b. 使用 Kaplan-Meier 方法分别画出有吸毒史和无吸毒史两组人群的生存函数，并比较有吸毒史和无吸毒史两组人群的生存函数在 5%显著性水平上有无显著区别。
c. 使用 Cox 回归模型估计危险比（Hazard Ratio），模型需要使用年龄和有无吸毒史作为协变量（Covariate）。指出协变量在 5%的显著性水平下是否显著，最后解释回归结果。

你可以独立完成作业，也可以组队完成，每个队伍人数不得超过 4 人。人数越少的队伍，在期末总评时会有适当加分奖励（奖励 0.5－3.5 分，最多可以将你的成绩提升一个绩点比如从 B+提高到 A-，但不会提升 2 个或者以上的绩点）。

提交作业的截止时间：1 月 15 日周日晚上 20:55 分。
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 HW03 你的名字”。