Business Analytics Homework 2

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\HW02 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\HW02 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\HW02 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. 使用课程文件夹/BA/Homework/HW02 中 Wine.csv 的数据。对第 2 列到第 14 列数据（从“Alcohol”列至“Proline”列）进行主成分分析。要求：使用主成分变量捕捉到至少 90%的总方差。

1.1）如果不对数据进行标准化处理，需要使用几个主成分变量可捕捉到 90%以上的总方差？
1.2）如果对数据进行标准化处理，需要使用几个主成分变量可捕捉到 90%以上的总方差？
1.3）比较以上两种方法，哪种处理方法较好？为什么？

2、Tayko Software 是一家出售教育软件和游戏软件的软件销售公司。最近它更新了软件目录并把不同的软件目录寄给它的用户。这次邮件营销产生了 1000 个购买行为。根据数据，公司打算用一个模型预测一个购买软件的用户会花多少钱。课程文件夹/BA/Homework/HW02 中的 Tayko.csv 包含了 1000 次购买行为的信息。表 1 描述了将要用在模型当中的变量。

2.1）使用 R 语言算出在软件目录 source_a 购买了软件的消费者平均消费金额是多少，在 source_b 购买了软件的消费者平均消费金额又是多少。

表 1：线性回归方程使用的变量：

Var. #	变量名	描述	变量类型	代码描述
1.	US	是美国地址吗？	binary	1: yes 0: no
17.	Freq.	去年在软件编目的交易次数	numeric	
18.	last_update_days_ago	几天前消费者记录被更新？	numeric	
20.	Web.order	消费者至少下了一次网上订单	binary	1: yes 0: no
21.	Gender.male	消费者是男性	binary	1: yes 0: no
22.	Address_is_res	地址是家庭住址	binary	1: yes 0: no
24.	Spending	该消费者在测试性邮件营销后的购买金额	numeric	

2.2）使用 R 语言画图：横轴为 Freq 纵轴为 Spending 的散点图、横轴为 Freq 纵轴为 Spending 的抖动（jittering）散点图、横轴为 last_update_days_ago 纵轴为 Spending 的散点图。
2.3）使用 R 语言做线性回归：以 Spending 作为被解释变量，以 US、Freq、last_update_days_ago、

Web.order、Gender.male、Address_is_res 作为解释变量。使用所有数据进行回归分析，给出线性回归的结果并指出哪些解释变量在 5%的显著性水平上显著。
2.4）使用 R 语言进行线性模型的变量选择：使用 backward elimination 方法选择线性回归模型的变量并给出最后被选择的模型形式。

3、我们打算使用课程文件夹/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。
3.1）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。
3.2）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。那么要想使用 Naive Bayes 方法处理相同的任务，我们应该对数据如何进行处理呢？

4、我们手头有一个海南省的联通用户订购联通数据包套餐的数据，我们打算根据现有数据使用 Naive Bayes 方法预测新的消费者最有可能订购哪个联通数据包套餐。现有数据集当中消费者的信息有一列是消费者目前的详细居住地址。联通数据挖掘项目组讨论数据挖掘时有 2 种意见：A）使用 Naive Bayes 方法训练数据时，地址精确到海南的县市一级即可。B）地址应精确到小区，这样才能提高精确度。假设你是负责联通数据挖掘项目的主管，请你评价以上 2 种意见。你认为地址列数据较好的处理方法应该是怎样的呢？

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：12 月 11 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 HW02 你的名字”。