= 习题

R 语言介绍

1. 请分别利用命令在 Rconsole 和 Rstudio 的菜单安装扩展包 ISLR 和 ElemStatLearn, 并
2. 分别查看这两个包里都有哪些函数以及这些函数的使用方法。2.请在 D 盘新建一个文件夹 RDataAnaly, 并把你的当前工作目录更改到这个目录。

R 语言基础

1. 使用 R 计算：pi， exp(1)
   a) 计算 31079 除 170166719 后的余数
   b) 计算π^e，e^π，$[\exp(\pi)]^e$，$[\exp(\pi^e)]$和$[\pi^e - e^\pi]$
   c) 计算$(2.3)^8 + \log(7.5) - \cos(\pi/\sqrt{2})$
2. 必要时使用 rep()和 seq()函数，生成向量
   a) （1 2 3 4 5 2 3 4 5 6 3 4 5 6 7 4 5 6 7 8 5 6 7 8 9）。
3. 计算$\sum_{i=1}^{N}\frac{1}{i}$，并和$\log(N)+0.6$比较，其中$N=500$。
4. x=rep(c(NA,seq(1,by=0.5, length=10)), times=5),将 x 中非缺失的元素提取出来赋值给 z, 并将 z 大于 3 的元素提取出来，并计算其均值
5. 使用 R 对以下矩阵
   a) A=matrix(c(35,4,12,2,14,11,9,5,11,3,38,12,1,0,4,2),4)
      - A
      - [,1] [,2] [,3] [,4]

   b) [1,] 35 14 11 1
   c) [2,] 4 11 3 0
   d) [3,] 12 9 38 4
   e) [4,] 2 5 12 2

   f) 求其行列式，A^{-1}，AA^T，$A^TA, A^{-1}A, A^{-1}A - AA^{-1}$。
6. 请将 MASS 包中的 Boston 数据集的 crim, lstat,rm, dis, age, medv 等变量提出出来并组成一个新的数据框 House
   (1) 把该新的数据集里按 medv 变量大于中位数的所有房子筛选出来，
   (2) 把 rm 大于 5 的房子也筛选出来。
   (3) 同时符合第（1）和第（2）问的房子筛选出来

编写函数

3. 编写一个函数可以计算向量里奇数的个数以及具体的奇数是哪些？
4. 编写一个函数模拟博饼，并试博 5 次，看看博饼的结果如何？是否博得状元？（博饼起源于泉州，后由郑成功带入厦门，是闽南地区特有的由饼文化外延的一种民俗活动，同时扔 6 颗骰子）。
5. 用 while 循环语句编写函数生成 Fibonacci 序列，要求告诉序列的长度而生成对应的序列。
6. 编写一个函数生成 Fibonacci 序列，要求可以输入生成序列的长度或者最大值都可以生成对应的序列。

数据读写与预处理

1. 新建一个 txt 文档，输入一些数据，尝试读入到 R 里，并把 txt 文档转换对应的 csv 文档，并读入到 R 里，然后将读入的数据再写出保存在本地的目录里。
2. 请分析 ISLR 包里的 Wage 数据：
   （1） 请按 race 把不同的种族的分别筛选出来，然后比较一下不同种族的平均工资是否一样？
   （2） 请把 wage 按 100 划分为高工资和低工资两类，并新建一个变量，分别用 high 和 low 代替高工资和低工资
3. 请分析 ISLR 包里的 Hitters 数据：
   （1） 请找出 Salary 变量里有缺失的所有球员，并统计一下总共有多少缺失
   （2） 请用均值和中位数分别插值缺失的数据

探索性分析与作图

1. 请分析 ISLR 包里的 Wage 数据：

(1)统计该数据的 race 的人数，做条线图和饼图分别分析

(2)计算 wage 的均值、方差、标准差、极差, 中位数，下四分位数，上四分位数

(2)绘制 wage 的直方图、密度估计曲线、QQ 图，并将密度估计曲线与正态密度曲线想比较

(3)绘制 wage 的箱线图、并计算五数概括。

2. 请分析 ISLR 包里的 Hitters 数据
(1)把有缺失数据的行全部删掉
（2）绘制 Years 对于 Salary 的散点图，并计算两者之间的相关系数；
(3)请分析 League 和 Division 之间的关系
（4）请用箱线图分析 Hits 与 League 的关系
(5)请画 AtBat， Hits， HmRun， Runs， RBI， Walks， Years 的多重散点图，并分析它们之间的关系

KNN

请分析 ISLR 包中的 Default 数据集。请将 Default 数据集按照 70：30 的比例随机划分为训练集和测试集，基于训练集建立 KNN 模型，自变量 X 是 balance，income,因变量 y 是 default, 分别计算 K=1,3,5,10 等几种的训练模型在在测试集的预测准确率。

线性回归

1. 假设一元线性回归模型 Y=2+3*X+e，其中 X 是服从 N（2，2）分布，扰动项 e 服从 N（0，1）分布。

   （1）. 请模拟样本容量为 100 的随机数，做出 Y 和 X 的散点图，并用最小二乘法估计出系数，在散点图上添加估计出来的回归线，并添加上真实的回归线，比较两者的差异。

   (2). 重复模拟 1000 次，请 1000 次模拟的系数估计结果用箱线图进行分析，并计算它们的均值和中位数，看看与真实参数是否存在差异。

2. 请分析MASS包里的Boston数据集，想要研究该地区的房屋价格中位数（medv）与该地区的底层人口比例（lstat）的关系。请

（1）请先分析 medv 的分布情况，做直方图，箱线图以及密度函数图分析一下分布情况。请计算 medv 的最小值，最大值，中位数，下四分位数和上四分位数以及标准差。

（2）请做散点图分析 medv 与 lstat 的关系，并计算它们之间的相关系数

（3）请做回归分析 medv 和 lstat 的关系，mdev 是因变量，这些系数是否都显著？并解释这些回归系数的含义，lstat 每增加一个单位，medv 会怎么变化？

（4）请预测当 lstat 分别为 5，10，15 的 medv 的值

3. 假设多元线性回归模型 Y=2+3*X1+1.5X2+e，其中 X1 和 X2 是服从均值为（1，1），边际方差为（2，2），协方差为 1 的二元正态分布，扰动项 e 服从 N（0，2）分布。

（1）. 请模拟样本容量为 100 的随机数，分别画出 Y 和 X1 和 X2 的散点图，并用最小二乘法估计出系数。

(2). 重复模拟 1000 次，请 1000 次模拟的系数估计结果用箱线图进行分析，并计算它们的均值和中位数，看看与真实参数是否存在差异。

4. 请分析MASS包里的Boston数据集，想要研究该地区的房屋价格中位数（medv）与其他影响因素的关系。

（1）请用矩阵式散点图分析各个因素之间的关系，以及分析哪些因素可能会影响房屋价格？

（2）请用逐步回归建立最优的模型，并估计出最优模型的系数以及解释这些系数的含义

logistic 回归

1. 请分析 ISLR 包中的 Default 数据集。

   （1） 请分析一下 default 的分布情况，并计算违约率，分析 default 与 student, balance 和 income 的关系

   （2） 请用一元 logistic 建模分析 default 分别与 student, balance，income 的关系

   （3） 请用多元 logistic 建模分析 default 与 student, balance，income 的关系，并比较分析一元 logistic 回归的分析结果与多元 logistic 的分析结果。

   （4） 请基于多元 logistic 回归的分析预测当一个申请者是 student, balance=2500, income=50000 的违约概率。

   （5） 请将 Default 数据集按照 70：30 的比例划分为训练集和测试集，基于训练集建立最优的 logistic 回归模型，计算训练集的预测准确率，并用测试集验证模型的可靠性，计算测试集的预测准确率。

2. 请结合自己的专业，思考什么场景是可以利用 logistic 回归进行建模分析？

变量选择

1. 请分析 MASS 包中的 Boston 数据集：

   （1） 利用 LASSO, MCP,SCAD 三种惩罚方法分析找出影响房屋价格 medv 的因素，比较一下这些方法找出的影响因素。

   （2） 请比较一下 LASSO 方法与逐步回归方法筛选出来的结果

2. 请分析 ISLR 包中的 Smarket 数据集

   （1） 以 Direction 为因变量，请用 LASSO, MCP,SCAD 三种惩罚方法分析找出影响股票价格涨跌方向的因素，并比较一下三种方法找出的影响因素是否一样？

3. 请模拟生成 X 从多元正态分布产生，， X_i, X_j 对应的相关系数是 $\rho^{|i-j|}$，$\rho = 0.1, 0.5, 0.9$，回归系数 $\beta = (1,1,1,1,1,0.5,0.5,0.5,0.5,0.5,0,\ldots,0)$,随机扰动项是标准正态分布，请模拟 100 次，分别用 lasso,SCAD,MCP 去筛选变量，比较变量筛选的 FNR,FDR

决策树、随机森林与 Boosting

1. 请分析 ElemStatLearn 包中 SAheart 数据集，

2. 请分析 MASS 包中 Boston 数据集，请将数据集按 70：30 拆分为训练集和测试集，利用决策树对训练集进行建模预测房价 medv，并对决策树修剪枝，然后分析测试集的预测 MSE。

3. 请分析 ISLR 包中的 Default 数据集：

（1）请将数据集按 60：40 的比例划分训练集和测试集， 请分别利用决策树、随机森林和 logistic 回归对训练集构建模型分析 default，利用测试集对所构建的模型进行测试。比较这三个模型的训练集预测准确率以及测试集的预测准确率。

（2）请思考如何在这三个模型中选择最优模型来预测？

支持向量机

1. 请分析 ISLR 包中的股票数据 Smarket，以股票的涨跌方向 Direction 为因变量，以 Lag1-Lag5 以及 Volume 为自变量，进行如下分析：
   （1） 分析一下该数据集的股票涨跌天数分别是多少以及它们的比例是多少？
   （2） 请以 2005 年之前数据为训练集，2005 年的数据为测试集。用训练集数据进行建模，先分析当 cost=1 时的建模结果，然后利用交叉验证方法选取最优的 cost 参数，并分析最优的模型结果
   （3） 利用得到的最优模型对测试集进行预测，分析预测准确率

2. 请分析 ISLR 包中的 Auto 数据集：

（1） 将 Auto 数据集中的 mpg 按照中位数划分为两类，新增一个变量 grade，并用 0 和 1 分别表示。

（2） 从该数据集随机抽取 292 个样本作为训练集，剩下的作为测试集

（3） 利用 maximal margin classifier 进行建模，利用交叉验证选取最优的模型，分析该最优模型的结果，并利用该最优模型对测试集进行预测分析

（4） 请利用 radial kernel 的 SVM 对训练集进行建模，利用交叉验证选择最优的模型，分析该最优模型的结果，并利用最优模型对测试集进行预测分析

聚类分析

1. 请聚类分析 datasets 包中的 USArrests 数据集。

（1） 请描述统计分析一下该数据集里各变量的关系

（2） 基于提供的变量，分别用系统聚类和 K-means 聚类方法将美国的 50 个州划分几类，并分析每类的特点。

思考如何确定最优的聚类数目？比如在这里是划分为 3 类好还是划分为 4 类好？

2. 请分析 class 包中 flower 数据集，基于提供的变量，分别用系统聚类和 K-means 聚类方法将花分成几类。并分析每类的特点。思考如何确定最优的聚类数目？

关联规则

1. 请分析 arules 包中的 Groceries 数据集：

   （1）请先统计该数据集中各个商品的频率

   （2）找出该数据集中有用的关联规则