数据分析与可视化 第三-四次作业

说明：

1. 作业上交日期为 2022 年 12 月 15 日中午 12 点，将作业发给助教邮箱，逾期将无法提交，视为放弃此次作业。

2. 请将作业保存成 pdf（不要用 word）上传；文件名为 XXXX_YY_Ass3.pdf，其中 XXXX 为你的学号名，YY 为你的名字。

3. 作业中的每个问题，涉及到代码问题的，都需要在该题目位置附上相应的 R（Rstudio）源代码，不要用截屏记录代码，否则视为没有作答（助教将复制并运行你的源代码）。

4. 作业中涉及到简答问题的，请给出你的答案和理由。

5. 涉及到生成随机数的问题，请设定“种子”；其中 seed 的值请设定为你的 9 位数学号，请在合适的位置用如下代码： set.seed(XXXXXXXX)

辛苦大家在圣诞和新年期间完成此次作业，预祝大家新年快乐！期末取得好成绩！

1. 利用你的学号，生成一个 1000*p 的矩阵 X，如下所示。

```
studno <- 1234567890 # 改成你的学号！！！！！
set.seed(studno)
n <- 1000
p <- 10
beta0 <- 1
beta <- c(c(1,2,3,4,5), rep(0, p-5))
X <- matrix(rnorm(n*p, 0, 1), nrow=n, ncol=p)
e <- rnorm(n, 0, 0.2)
Y <- beta0 + X %*% beta + e
dat <- data.frame(Y,X)
colnames(dat) <- c("Y", paste("X", 1:p, sep=""))
```

（a）请描述目前生成的响应变量中，有用的自变量是哪些。

（b）请用 AIC 估计 Y ~ X 的线性回归中，依次估计出来的系数非零的变量分别是哪些。

（c）请用 lasso 和 ridge，依次估计出来的系数非零的变量分别是哪些，绘制 solution path

（d）设定 p 为 100 重复（b）-（c），结果有什么变化？

2. 请基于 wineTrain 数据集，进行主成分分析。

(a) 请计算 wineTrain 的主成分，并输出计算结果，你应该得到一个 13*13 的得分矩阵。

(b)通过合适的图表，将 fviz_pca_ind(), fviz_pca_var() 和 fviz_pca_var()进行展示。这三个图展示的分别是什么？

(c) 计算每个主成分对原始变量的解释程度，按照降序排列，并计算累计贡献率。你觉得选多少个主成分比较合适？

(d) 展示原始变量到前 m 个主成分的变换矩阵，其中 m 为你在(c)中的主成分的个数。在前 m 个主成分上，原始的 13 个变量对每个主成分的影响有多少是正的影响，有多少是负影响？请通过 apply 函数进行展示。

(e) 将你得到的 m 个主成分作为新的变量，基于这 m 个主成分进行 k-means 聚类，nstart=25.你认为 k 取多少比较合适？与基于原始数据的 k-means 相比，k 的值是否相同？

(f) 基于 m 个主成分进行 3-menas 聚类，该模型得到的分组结果与原始变量 Cultivar 的标签吻合度如何？

(g) 请通过任何一种你学过的分类方法，将 wine 进行分类，其中 Cultivar 作为响应变量，前 m 个主成分作为解释变量，得到每个样本点的分类的预测值。此时的预测效果怎么样？与基于原始变量进行预测相比，此时的精准度如何？基于此，你觉得基于主成分分析后再进行模型预测，是否可取？