生统第七次作业

一、“Stocks.txt” is data related to stocks, with the symbols appearing in column 1, and various variables relating to the symbol on the right.

Question:

1. Apply PCA on this data and explain how much variability is explained by the first two principal components.（5’）
2. How many components to keep if we want to have more than 90% variance explained.（5’）
3. Use biplot() to visualize the PCA result, and interpret how many variables comprised by the principal components 1.（10’）
4. 什么是聚类？（5’）

（2）常用的距离有哪些？（5’）

（3）现有A(1,1)，B(1,2)，C(5,6)，D(3,3)，E(6,7)五个点，计算他们之间的欧式距离，并用AGNES算法（采用最小距离）将它们划分到两个簇中。（10’）

三、

1. 什么是PCA，数据降维有何意义？（5’）
2. 哪些数据适合用PCA降维呢（5’）
3. 现有鸢尾花数据集（R语言内置），每组观测值由四个特征值（花萼长度，花萼宽度，花瓣长度，花瓣宽度）和分组变量组成。请对运用PCA对原始的特征数据进行降维，并取得降维后的数据（只取前三个主元）（10’）

四、

1. 聚类的方法大致有哪些？（5’）
2. 聚类和分类的区别是什么？（5’）
3. 列举生物学科研中用到聚类的两个例子？（5’）
4. 请利用广泛使用的iris数据的花瓣属性值进行简单层次聚类。（5’）

五、Consider the two-dimensional data points in “dataPoints.csv”:

(please show your R code. Hint: you can use R function “prcomp”)

1) What is the first principal component (write down the actual vector)?（5’）

2) If we project the original data points into the 1-d subspace by the principal component you choose, what are their coordinates in the 1-d subspace? And what is the sample variance of the projected data?（10’）

3) For the projected data you just obtained above, now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, what is the reconstruction error? (please give your reason)（5’）