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.
2. How many components to keep if we want to have more than 90% variance explained.
3. Use biplot() to visualize the PCA result, and interpret how many variables comprised by the principal components 1.

Answer:

> data <- read.table("Stocks.txt", row.names = 1, header = T)

> pca <- princomp(data, cor = T)

> summary(pca)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8

Standard deviation 2.070766 1.2971806 1.0228343 0.70727217 0.62193416 0.24914879 0.179082732 0.0433452635

Proportion of Variance 0.536009 0.2103347 0.1307738 0.06252924 0.04835026 0.00775939 0.004008828 0.0002348515

Cumulative Proportion 0.536009 0.7463437 0.8771174 0.93964667 0.98799693 0.99575632 0.999765149 1.0000000000

可以看出PC1解释了53.6%的variability ， PC2解释了21%的variability 。

根据Cumulative Proportion，可以知道，需要保留PC1-PC4。

> biplot(pca)

图片包含 文字, 地图

描述已自动生成

> loadings(pca)

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8

quantity 0.606 0.443 0.511 0.329 0.254

entry\_price 0.148 -0.639 -0.167 0.418 0.504 0.325

profit\_dollar 0.336 -0.817 0.438 -0.141

market\_cap 0.470 0.301 0.217 -0.619 0.506

cash\_and\_marketable 0.375 0.308 -0.233 -0.450 0.398 -0.553 0.119 -0.171

tev 0.479 0.456 -0.100 -0.740

revenues 0.463 -0.164 -0.305 0.302 0.632 0.406

ebitda 0.413 -0.117 0.214 0.375 -0.523 -0.581 -0.133

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8

SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000

Proportion Var 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125

Cumulative Var 0.125 0.250 0.375 0.500 0.625 0.750 0.875 1.000

这⾥看出，PC1主要由entry\_price, market\_cap, cash\_and\_marketable, tev, revenues, ebitda组成。

二、（1）什么是聚类？（20分）（2）常用的距离有哪些？（20分）（3）列举两个个常用的聚类方法。（20分）（4）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。（40分）

1. 聚类是指将样本分到不同的组中，使得同一组中的样本差异尽可能的小，不同组中的样本差异尽可能的大。
2. 欧式距离，曼哈顿距离，民科夫斯基距离。
3. K-means聚类，层次聚类。
4. 各点之间的欧式距离如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E |
| A | 0 |  |  |  |  |
| B | 1 | 0 |  |  |  |
| C | 6.40 | 5.66 | 0 |  |  |
| D | 2.83 | 2.24 | 3.61 | 0 |  |
| E | 7.81 | 7.07 | 1.41 | 5 | 0 |

上表所有距离中，A、B之间距离最小，首先将A、B聚类到一个簇中；新的距离矩阵如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AB | C | D | E |
| AB | 0 |  |  |  |
| C | 5.66 | 0 |  |  |
| D | 2.24 | 3.61 | 0 |  |
| E | 7.07 | 1.41 | 5 | 0 |

此时，距离最小的是C、E，将其聚类到一个簇中；新的距离矩阵如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | AB | CE | D |
| AB | 0 |  |  |
| CE | 5.66 | 0 |  |
| D | 2.24 | 3.61 | 0 |

此时，AB和D的距离最小，将AB、D聚类到一个簇中，此时仅有两个簇，聚类完成。A、B、D聚类到一个簇，而C、E聚类到另一个簇。

三、

1. 什么是PCA，数据降维有何意义？
2. 哪些数据适合用PCA降维呢
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）

答案：

1：PCA即主成分分析，是一种广泛使用的数据降为算法，主要思想是将n维特征的数据映射到k维（k<n）。从而使下游的数据处理变得快速简单。新构造的主元相互正交（线性无关），且按方差降序排列。

2：适用于维度较多，特征冗余（特征间具有相关性），噪音较多的数据。

3：

library(dplyr)

iris<-select(iris,-Species)

iris.pca<-princomp(iris,cor = T)

summary(iris.pca,loadings = T)

new.iris<-as.data.frame(iris.pca$scores[,1:3])

四、

1. 聚类的方法大致有哪些？
2. 聚类和分类的区别是什么？
3. 列举生物学科研中用到聚类的两个例子？
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. 第一题

– 划分聚类方法

– 层次聚类方法

– 密度聚类方法

– 网格聚类方法

– 基于模型的方法

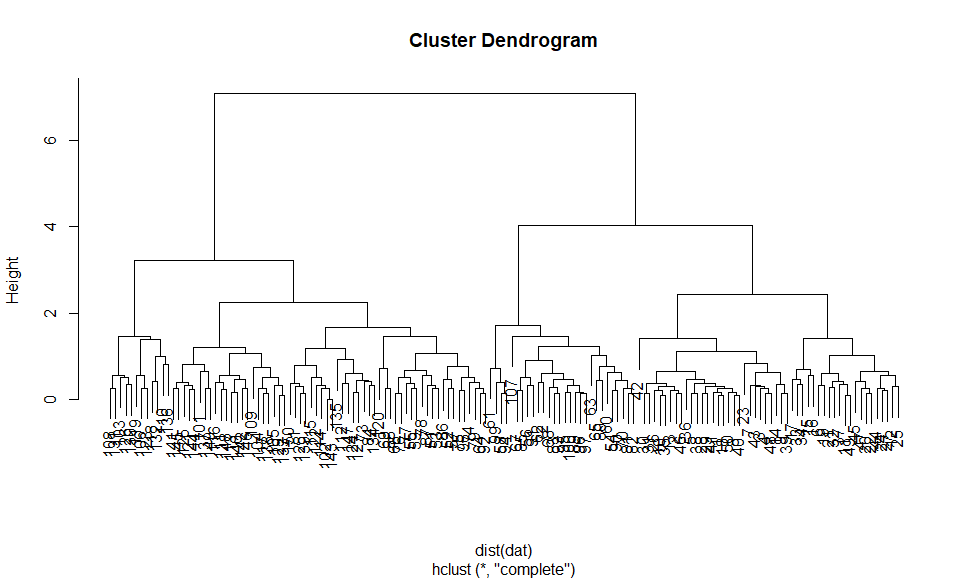
– 其它聚类方法

1. 分类是监督学习，聚类是非监督学习
2. 如统计肿瘤细胞参数对肿瘤细胞进行聚类，学生自圆其说即可。

dat <- iris[,1:4]

hc = hclust(dist(dat))

plot(hc)



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)?(10)

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?(15)

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)(10)

Answers:

#1)

> rm(list = ls())

> data <- read.csv('/dataPoints.csv',header = T,sep = ',')

> data.pca <- prcomp(data,center = TRUE,scale = TRUE)

> data.pca

Standard deviations (1, .., p=2):

[1] 1.414214e+00 2.372134e-17

Rotation (n x k) = (2 x 2):

PC1 PC2

x 0.7071068 0.7071068

y 0.7071068 -0.7071068

That is:

PC1: (0.707106, 0.7071068) or (-0.707106, -0.7071068) or () or ()

#2)

#a

> data.pca$x[,1]

[1] -1.414214 0.000000 1.414214

Or

#b

> var(data.pca$x[,1])

[1] 2

#3)

The reconstruction error is 0, since all three points are perfectly located on the direction of

the first principal component. Or, you can actually calculate the reconstruction:

> PC1 <- c(sqrt(2)/2,sqrt(2)/2)

> transformData <- data.pca$x[,1]

> transformData[1]\*PC1

[1] -1 -1

> transformData[2]\*PC1

[1] 0 0

> transformData[3]\*PC1

[1] 1 1

which are exactly the row data points.