CMDA 3654: Assignment #5

Due on Friday, Nov 4, 2016

Jung Choi

"I have neither given nor received unauthorized assistance on this assignment."

November 4, 2016

1 Principal Components Analysis..

(a) Load the **USArrests** dataset in R, and describe the dataset in your own words, in 2-3 lines.

The **USArrests** is a dataset that contains number of arrests for murder, assault and rape for each of the 50 US states in 1973. There is another column that contains percent values of urban population.

(b) Calculate the mean and standard deviation of the four variables.

```
         vars  n   mean    sd median
Murder      1 50   7.79  4.36   7.25
Assault     2 50 170.76 83.34 159.00
UrbanPop    3 50  65.54 14.47  66.00
Rape        4 50  21.23  9.37  20.10
```

(c) Perform principal components analysis of the dataset using R. How many principal components are there? For each principal component, report the standard deviation and the proportion of variance explained.

```
Standard deviations:
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation:
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

(d) How many principal components would you need to explain at least (i) 60% of the total variance? 1 (ii) 75% of the total variance? None (iii) 90% of the total variance? 1

```
                          Comp.1    Comp.2    Comp.3
Proportion of Variance 0.6200604 0.2474413 0.0891408
```

2 (60 points) Hierarchical and K-means clustering.

a) Load the NCI60 dataset from the ISLR package in R, and describe the dataset in your own words, in 2-3 lines.

NCI60 is a dataset that contains microarray data which is expressed in 64 by 6830 matrix. It contains 6830 genes from 64 cancer cell lines.

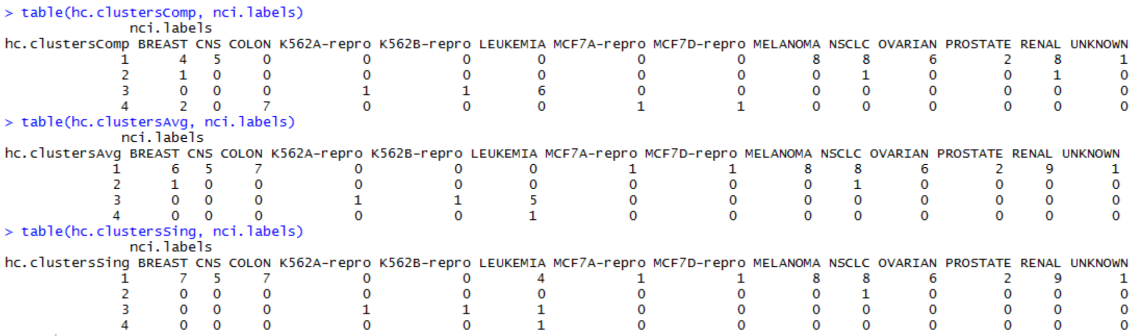
(b) Using Euclidean distance as the dissimilarity measure, perform hierarchical clustering on the data, with (i) Complete Linkage, (ii) Average Linkage, and (iii) Single Linkage.

```

# Cluster method   : complete
# Distance         : euclidean
# Number of objects: 64
# Cluster method   : single
# Distance         : euclidean
# Number of objects: 64
# Cluster method   : average
# Distance         : euclidean
# Number of objects: 64
```

(c) For all three methods in (b), cut the hierarchical clustering tree at 4 clusters, and report the two-way table of actual cancer types (NCI60$labs) and clusters. Are there any differences between the tables from different methods?

It's hard to conclude which linkage type is optimal for us to use, but we can see the obvious difference from the results below. Then depending on the use, we can choose from the following types. The complete takes largest dissimilarities, single uses smallest dissimilarities and average takes mean values of dissimilarities from the observations.



```
> table(hc.clustersComp, nci.labels)
               nci.labels
hc.clustersComp BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
              1      4   5     0           0           0        0           0           0        8     8       6        2     8       1
              2      1   0     0           0           0        0           0           0        0     1       0        0     1       0
              3      0   0     0           1           1        6           0           0        0     0       0        0     0       0
              4      2   0     7           0           0        0           1           1        0     0       0        0     0       0
> table(hc.clustersAvg, nci.labels)
              nci.labels
hc.clustersAvg BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
             1      6   5     7           0           0        0           1           1        8     8       6        2     9       1
             2      1   0     0           0           0        0           0           0        0     1       0        0     0       0
             3      0   0     0           1           1        5           0           0        0     0       0        0     0       0
             4      0   0     0           0           0        1           0           0        0     0       0        0     0       0
> table(hc.clustersSing, nci.labels)
               nci.labels
hc.clustersSing BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
              1      7   5     7           0           0        4           1           1        8     8       6        2     9       1
              2      0   0     0           0           0        0           0           0        0     1       0        0     0       0
              3      0   0     0           1           1        1           0           0        0     0       0        0     0       0
              4      0   0     0           0           0        1           0           0        0     0       0        0     0       0
```

(d) Perform k-means clustering of the data with k=4 clusters. Report the two-way table of actual cancer types (**NCI60$labs**) and clusters.

```
> table(km.clusters,hc.clustersComp) # even with same K, hclus and kmeans can give
     different
           hc.clustersComp
km.clusters  1  2  3  4
          1  9  0  0  0
          2  0  0  8  0
          3 24  3  0  0
          4  9  0  0 11
> table(km.clusters,hc.clustersAvg) # even with same K, hclus and kmeans can give
    different
           hc.clustersAvg
km.clusters  1  2  3  4
          1  9  0  0  0
          2  0  0  7  1
          3 25  2  0  0
          4 20  0  0  0
> table(km.clusters,hc.clustersSing) # even with same K, hclus and kmeans can give
     different
           hc.clustersSing
km.clusters  1  2  3  4
          1  9  0  0  0
          2  4  0  3  1
          3 26  1  0  0
          4 20  0  0  0
```