HW8_kmeans.R

lukemcevoy

2021-11-22

```r
# clear the environment
rm(list=ls())

# select the data
filename<-’/Users/lukemcevoy/Develop/stevens/f21/dataMining/week10/hw7/wisc_bc_ContinuousVar.csv’
cancer<-read.csv(filename)
# cancer_df<-data.frame(lapply(na.omit(cancer),as.numeric))
cancer_df<-data.frame(cancer)
cancer_df<-cancer_df[-1]
cancer_df$diagnosis <- ifelse(cancer_df$diagnosis == ’M’, 1, 0)
View(cancer_df)

normalized_cancer_df<-as.data.frame(apply(cancer_df[,1:ncol(cancer_df)], 2, function(x) (x-min(x))/(max

# We want to cluster with all features BUT diagnosis, we remove here
normalized_cancer_df<-normalized_cancer_df[-1]

# split data
index<-sort(sample(nrow(normalized_cancer_df), round(.3*nrow(normalized_cancer_df))))
training<-normalized_cancer_df[-index,]
test<-normalized_cancer_df[index,]

kmeans_2<- kmeans(normalized_cancer_df[,-ncol(normalized_cancer_df)],2,nstart = 10)
kmeans_2$cluster
```

```
##   [1] 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 1 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 1 1 2 1 1 2 2 1 2 1 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 2 1 2 2 1 1 2 1 2 1 2 1
##  [75] 1 2 1 2 2 1 1 2 2 2 1 2 1 2 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1
## [112] 1 2 1 1 1 1 2 2 1 1 2 2 1 1 1 1 2 2 2 1 2 2 1 2 1 1 1 2 1 1 2 1 1 1 1 2 1
## [149] 1 1 1 1 2 1 1 1 2 1 1 1 1 2 2 1 2 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 2 2 2 1 1
## [186] 1 2 1 1 1 2 1 1 2 2 1 2 1 2 2 1 2 2 2 1 1 1 1 1 1 2 1 2 2 2 2 1 1 2 2 1 1
## [223] 1 2 1 1 1 1 1 2 2 1 1 2 1 1 2 2 1 2 1 1 1 1 2 1 1 1 1 1 2 1 2 2 2 1 2 2 2
## [260] 2 2 1 2 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1
## [297] 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 1 2 2 2 1 1
## [334] 1 1 2 1 2 1 2 1 1 1 2 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2
## [371] 2 1 2 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1
## [408] 1 2 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 2 1 2 2 1 2 1 1 1 1 1 2 1 1
## [445] 2 1 2 1 1 2 1 2 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1
## [482] 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 2 2 1 2 1 2 1 1 1 1 1 2 1 1 2 1 1 1 2 2
## [519] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [556] 1 1 1 1 1 1 1 2 2 2 2 2 2 1
```

```r
kmeans_2$centers
```

```
##   radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## 1   0.2551389    0.2883485      0.2468546 0.1437167       0.3574644
## 2   0.5079424    0.3967220      0.5087786 0.3664586       0.4710213
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1        0.1814862      0.1052516           0.1313797     0.3411801
## 2        0.4222137      0.4180701           0.4714324     0.4580997
##   fractal_dimension_mean  radius_se texture_se perimeter_se    area_se
## 1              0.2577646 0.06420433  0.1884376   0.06016572 0.02866725
## 2              0.2961483 0.19242952  0.1911349   0.17947434 0.13202609
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1     0.1817308      0.1346190   0.05966163         0.1825441   0.1725144
## 2     0.1798694      0.2557809   0.12318898         0.3070239   0.1896424
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1           0.08541214    0.2049828     0.3203932       0.1923723 0.09927551
## 2           0.13038679    0.4839449     0.4530746       0.4685514 0.31723198
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1        0.3567883         0.1497726       0.1330193            0.2634654
## 2        0.5008625         0.3641053       0.3897802            0.6601540
##   symmetry_worst
## 1      0.2266010
## 2      0.3382888
```

```r
table(kmeans_2$cluster,cancer_df$diagnosis)
```

```
##
##       0   1
##   1 350  32
##   2   7 180
```