k-means Clustering of UCI Diabetes Dataset

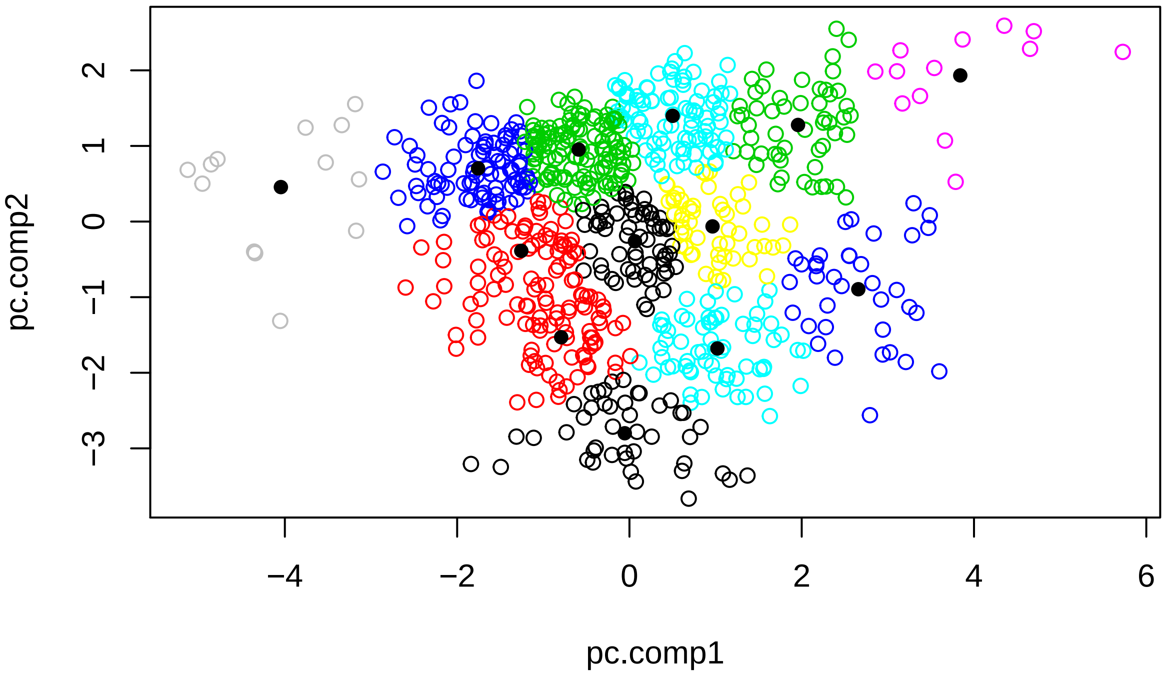
Jon Kinsey

Wed Dec 3 21:06:56 2014

```r
#
# set the working directory
setwd("/Users/Jon/Desktop/R-Projects/diabetes")
# comma delimited data and no header for each variable
RawData <- read.csv("diabetes.csv",sep = ",",header=FALSE)

# In RawData, the response variable is its last column; and the remaining
# columns are the predictor variables.
responseY <- RawData[,dim(RawData)[2]]
predictorX <- RawData[,1:(dim(RawData)[2]-1)]

# For the convenience of visualization, we take the first two principle components
# as the new feature variables and conduct k-means only on these two dimensional data.
pca <- princomp(predictorX, cor=T) # principal components analysis using correlation matrix
pc.comp <- pca$scores
pc.comp1 <- -1*pc.comp[,1] # principal component 1 scores (negated for convenience)
pc.comp2 <- -1*pc.comp[,2] # principal component 2 scores (negated for convenience)


# In R, kmeans performs the K-means clustering analysis, ()$cluster provides the
# clustering results and ()$centers provides the centroid vector (i.e., the mean)
# for each cluster.

X <- cbind(pc.comp1, pc.comp2)
cl <- kmeans(X,13)
cl$cluster
```

```
##   [1]  5 11  2 11  3 10 11 10  4  1  9  5  1  6  5 10  3  2 11 13  3  1  5
##  [24]  9  5  5  5 11  1  2  5  3 12 10  5 13  1  5 13  4  7  5  2  4  2  3
##  [47] 10 11  9  8 12 12 10  4  4 12  4  3  7 13  8  2 10 13  2 10  7  5 12
##  [70]  9 11  7  1  7 11  8  2  9 12 12 12  8  9 12  5 13  5 11  5 12 12  9
##  [93]  7  2 11  7 11 12 11  3  7 12 12 12 11 13 10  9 11 11  7  4 11 12  7
## [116]  1  2 12 11 12  6  9 13  1 12 13 13 13  7  2  7  2  3  9 12 11 11 11
## [139] 11 13  2  9 11  2 13 12  2 13  1 12  3 10  4  6  5  4 12 11 11  4  7
## [162]  7  3 11  9  9 11 10 10  9  2  7 12 13 11  4  2  6  5  9 10 13 12 10
## [185]  2  5  4  3  7  7 12  7  2  2  2  3 12 12 13  7 11  9 11 12  5 10  4
## [208]  5 11  5 12  3  5 13  7  4 13  9  9  2  3  5 10  5 11 11 12 13  6 13
## [231]  9  4 11 10 11  9  4  3  5 12 12 11 10  7  3  5  2  6  7 11  2 10 12
## [254] 11  5 11 11 11 13  4  7 12 11  5 10  9 12 13 12 12  4 11  2 13  1 13
## [277] 10 11  2 11 11  5  5  1  2  5  6  3 12  7 11 13  3 13  2  7  3 13  5
## [300]  1 12 13  9  9  2 13  5 11 13 13 10 13  9 11  7 11 10  9 13  1  9 11
## [323] 10  5 11 11 13  1 13  9  5 12 10  1 12  3 12  2  4  5 11 11 12  2  1
## [346]  5 11  8 12 11 10 10  9 12 11  1 13  2  5  3  3  1  5  1  7 11 10 12
## [369] 11  7  6 11 11 13 13  4 11 13  9  3 13 12 11 11 11 12  9  5  4 11 13
## [392]  9 13 10  9 13 11 13 12  7 10  2  7  2  2 13  2 12  5  6  7 13  3 11
## [415] 13  3 11  7 12 11  3 11 13 11  4  3  8  7  3  7  8  9 11 10 10 12  5
## [438] 10 12  9  7 11 11  2 10  6 11 13 13 11 12 10 13  2 13  5  2 11  4  1
## [461]  2 12  9 10  2 12 12 13 10  3  3 11 11  1 10  9  3  9  9  5  3 13 12
## [484] 13 12  3  3  4 10  1 11  9  9  7  8  1 10 11  5  7 11 11 11  9  9  2
## [507]  7 13 11  1  2 11  1 12 12  9  5  1  2  5 11 13  8  1 10 12 12 11 13
## [530] 12 11 11 13 12 13 12 10  2  3  3  7 13  5  9 11  4  4  9  7  5 11 11
## [553]  1 11 11  9 11  1  5  2  1  3 13 10 12 11 11  9  7 13 10 10 11 11  3
## [576] 13  9 11  2  4  3 10  1  2  4 12  2 10  4  8  4 13  2 13  3 13 10 12
## [599]  9 12 11  8 11  5 10 11  6 12  3 11 11  7  4  9  5 10  2 12  5 12 13
## [622] 11  5 13 12 13 12 11  2 12  2 13 12 11  2  1  2 11  7 12 11 10  1 12
## [645] 13  3  9 13  5 12 12 11  7 12 11  3 11  3  1 13  1  3  4  4  9 13  1
## [668]  2  7  5  5 12  5  6  1  2  1 12 10 11 12  3 13 10  1 13 10 12 13  3
## [691]  2  1 13  7 12  4  9  8  7  9 13  5  7 10  9  9  8 13  1 13 13  9  5
## [714] 13 10  3  3  1 13  9 10 13  7  7  2  7 13 11 10 12  9  2  3 11  2 11
## [737] 13  2 11 10  4 11 11  1  4  5  3  3  7  2  9 13 12  6  5  3  5  2 10
## [760]  1 11  5  2  5 11  9 10 11
```

```
plot(pc.comp1, pc.comp2,col=cl$cluster)
points(cl$centers, pch=16)
```



```
# Take k = 13 as the number of clusters in K-means analysis.
# The figure shows the resulting scatter plot with different clusters in different
# colors. The solid black circles are the centers of the clusters.
```