K-Means Clustering

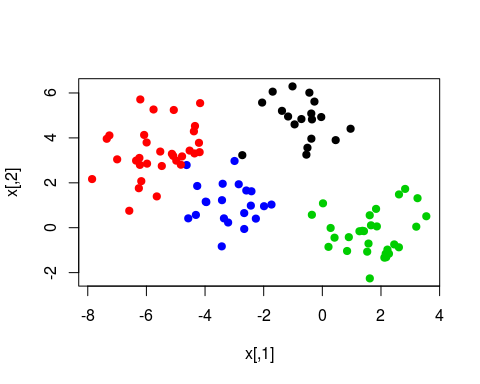
brett

April 4, 2019

# K-Means Clustering

K-means works in any dimension, but is most fun to demonstrate in two, because we can plot pictures. Lets make some data with clusters. We do this by shifting the means of the points around.

set.seed(101)  
x=matrix(rnorm(100\*2),100,2)  
xmean=matrix(rnorm(8, sd=4),4,2)  
which=sample(1:4, 100, replace=TRUE)  
x=x+xmean[which,]  
plot(x, col=which, pch=19)

 We know the “true” cluster IDs, but we wont tell that to the ‘kmeans’ algorithm.

km.out=kmeans(x, 4, nstart=15)  
km.out

## K-means clustering with 4 clusters of sizes 21, 30, 32, 17  
##   
## Cluster means:  
## [,1] [,2]  
## 1 -3.1068542 1.1213302  
## 2 1.7226318 -0.2584919  
## 3 -5.5818142 3.3684991  
## 4 -0.6148368 4.8861032  
##   
## Clustering vector:  
## [1] 2 3 3 4 1 1 4 3 2 3 2 1 1 3 1 1 2 3 3 2 2 3 1 3 1 1 2 2 3 1 1 4 3 1 3  
## [36] 3 1 2 2 3 2 2 3 3 1 3 1 3 4 2 1 2 2 4 3 3 2 2 3 2 1 2 3 4 2 4 3 4 4 2  
## [71] 2 4 3 2 3 4 4 2 2 1 2 4 4 3 3 2 3 3 1 2 3 2 4 4 4 2 3 3 1 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 30.82790 54.48008 71.98228 21.04952  
## (between\_SS / total\_SS = 87.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

plot(x, col=km.out$cluster, cex=2, pch=1, lwd=2)  
points(x, col=which, pch=19)  
points(x, col=c(4,3,2,1)[which], pch=19)

