

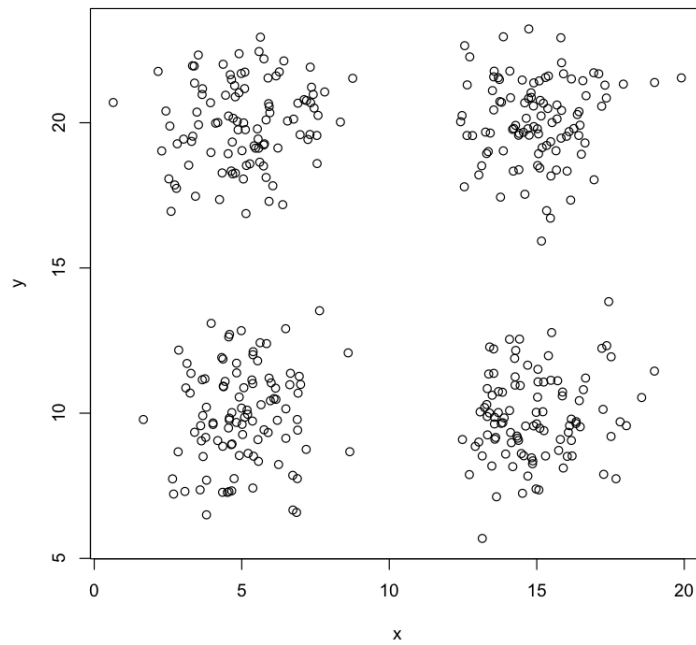
Math 5365

Data Mining 1

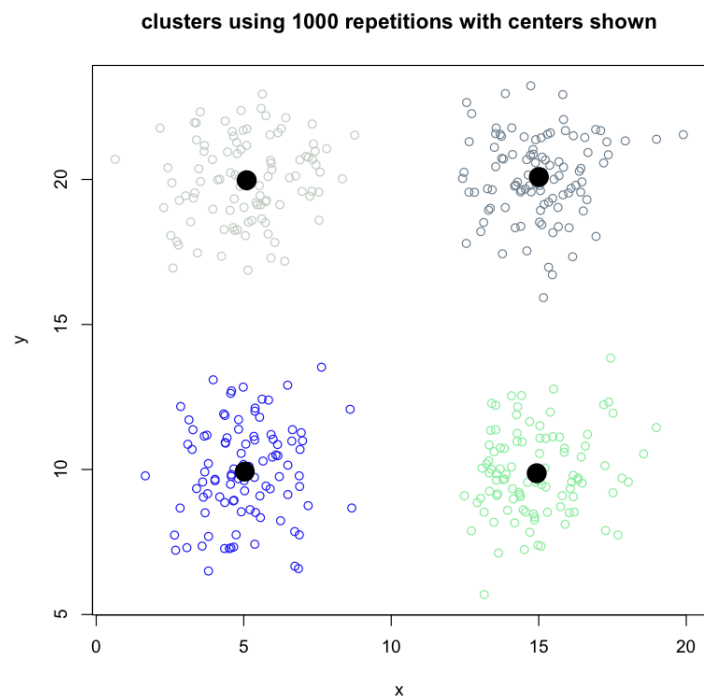
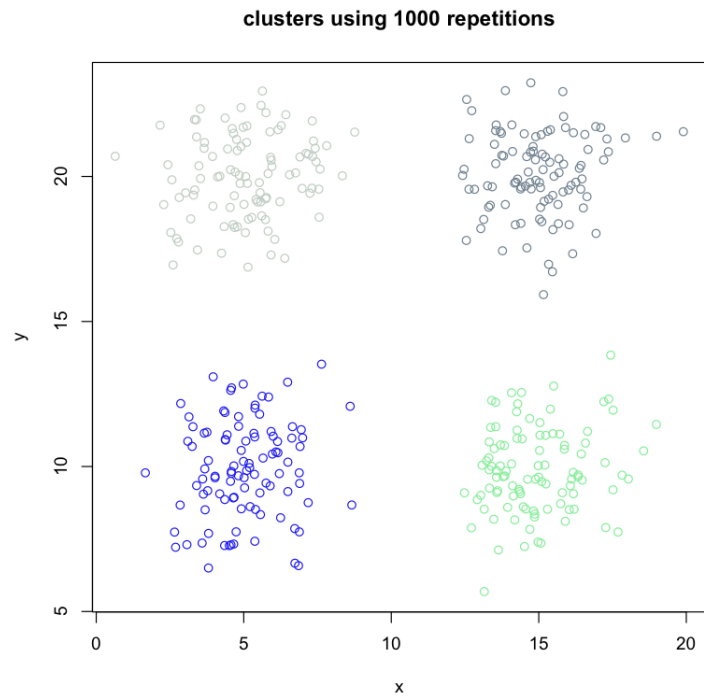
Homework 17

Mary Barker

1. Generate a data set similar to the one displayed, where each of the four clusters has 100 points.



- (a) Perform a K-means clustering with $K = 4$ and 1000 repetitions.
- (b) Plot the points and color them based on which clusters they are in.



(c) Find the total, total within, and between sums of squares

The total sum of squares is 21777.85.

The total within sum of squares is 1711.74.

The total between sum of squares is 20066.11.

- (d) Find the centers of the clusters.

The centers are in the table below

	x	y
1	5.034003	9.927154
2	15.004838	20.085618
3	14.931673	9.870831
4	5.100832	19.972832

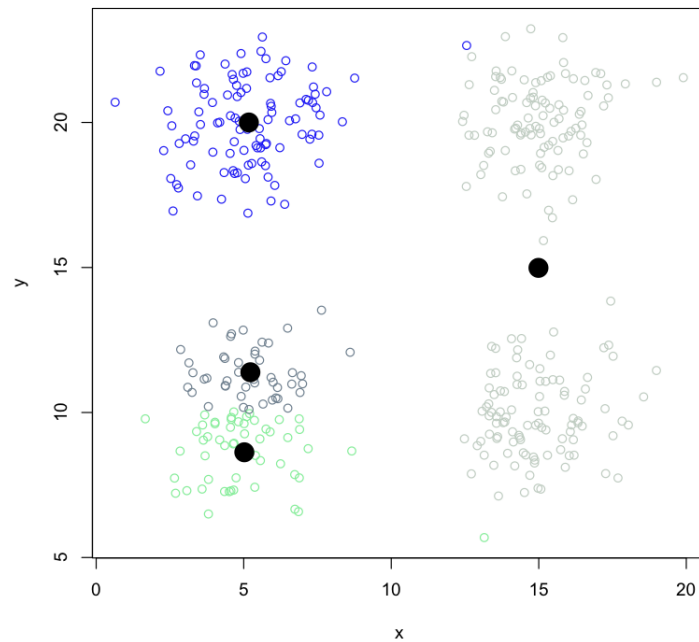
- (e) Suppose we want at least one of our repetitions of K-means to have the property that every cluster contains exactly one initial centroid. How many repetitions would be necessary to ensure that this happens with at least 99% probability?

The minimum number of necessary repetitions is 47.

- (f) Can you find a choice of initial centers that does not result in the optimal clusters?

What is the total within sum of squares for that clustering?

The center of each cluster is plotted together with the colored points



The total within sum of squares for this case is 6728.091.

2. Perform a K-means clustering with $K = 2$ and 1000 repetitions for the wdbc data set. What classification accuracy would be achieved if the clusters were used to predict the diagnosis in this data set?

The classification accuracy using $K = 2$ and 1000 repetitions is roughly 85.41301%

```

1 #Data Mining hw 17
2 library(stats)
3 wdbc <- read.csv('~/.Dropbox/Tarleton/data_mining/dfiles/wdbc.data',
4                 header=F,sep=',')
5 wdbc <- wdbc[,-1]
6 source('~/.Dropbox/Tarleton/data_mining/generic_functions/dataset_ops.R')
7
8
9 # 1. Generate a data set similar to the one displayed, where each of the

```

```

10 #    four clusters has 100 points.
11 set.seed(0)
12 x <- c(rnorm(100, 5, 1.5), rnorm(100, 15, 1.5),
13        rnorm(100, 5, 1.5), rnorm(100, 15, 1.5))
14 y <- c(rnorm(100, 10, 1.5), rnorm(100, 10, 1.5),
15        rnorm(100, 20, 1.5), rnorm(100, 20, 1.5))
16 plot(x, y)
17 points <- data.frame(x = x, y = y)
18
19 #    (a) Perform a K-means clustering with K = 4 and 1000 repetitions.
20
21 kmeans_reps <- function(dset, centers, reps){
22   w_ss = Inf
23   for(i in 1:reps){
24     k_cluster <- kmeans(x = dset, centers = centers)
25     if((k_cluster$tot.withinss) < w_ss){
26       ssw = k_cluster$tot.withinss
27       my_k_cluster <- k_cluster
28     }
29   }
30   return(my_k_cluster)
31 }
32 reps = 1000
33 mycluster <- kmeans_reps(points, 4, reps)
34
35 #    (b) Plot the points and color them based on which clusters they are in.
36 plot(x, y, col=c('blue',

```

```

37         'slategray',
38         'lightgreen',
39         'honeydew3',
40         'orange',
41         'brown')[mycluster$cluster],
42     main='clusters using 1000 repetitions')
43 plot(x, y, col=c('blue',
44                 'slategray',
45                 'lightgreen',
46                 'honeydew3',
47                 'orange',
48                 'brown')[mycluster$cluster],
49     main = 'clusters using 1000 repetitions with centers shown')
50 lines(mycluster$centers,type='p',pch=16,col='black', cex = 2.5)
51 # (c) Find the total, total within, and between sums of squares
52 mycluster$totss
53 mycluster$tot.withinss
54 mycluster$betweenss
55
56 # (d) Find the centers of the clusters.
57 mycluster$centers
58
59 # (e) Suppose we want at least one of our repetitions of K-means to have
60 #     the property that every cluster contains exactly one initial centroid.
61 #     How many repetitions would be necessary to ensure that this happens
62 #     with at least 99% probability?
63

```

```

64  minimum_reps <- function(k, err){
65    ceiling(log(err) / log(1 - factorial(k) / (k^k)))
66  }
67  minimum_num = minimum_reps(4, 0.01)
68
69  # (f) Can you find a choice of initial centers that does not result in the
70  #       optimal clusters? What is the total within sum of squares for that
71  #       clustering?
72
73  # try with alternate centers:
74  centers04 <- cbind(c(10, 15, 10, 5), c(10, 15, 20, 15))
75  newcluster04 <- kmeans(x = points, centers = centers04)
76  plot(x, y, col=c('blue',
77                  'slategray',
78                  'lightgreen',
79                  'honeydew3',
80                  'orange',
81                  'brown')[newcluster04$cluster],
82        main = 'clusters using 4 centers')
83  lines(newcluster04$centers,type='p',pch=16,col='black', cex = 2.5)
84
85  centers03 <- cbind(c(4, 15, 5, 16), c(10, 20, 10, 20))
86  newcluster03 <- kmeans(x = points, centers = centers03)
87  plot(x, y, col=c('blue',
88                  'slategray',
89                  'lightgreen',
90                  'honeydew3',

```

```

91         'orange',
92         'brown')[newcluster03$cluster],
93     main = 'clusters using 4 centers')
94 lines(newcluster03$centers,type='p',pch=16,col='black', cex = 2.5)
95
96 centers02 <- cbind(c(5, 15, 15, 15), c(10, 15, 20, 12))
97 newcluster02 <- kmeans(x = points, centers = centers02)
98 plot(x, y, col=c('blue',
99                 'slategray',
100                 'lightgreen',
101                 'honeydew3',
102                 'orange',
103                 'brown')[newcluster02$cluster],
104     main = 'clusters using 4 centers')
105 lines(newcluster02$centers,type='p',pch=16,col='black', cex = 2.5)
106
107 centers01 <- cbind(c(2, 10, 10, 20), c(2, 10, 20, 25))
108 newcluster01 <- kmeans(x = points, centers = centers01)
109 colors = 5 * c(1:50)
110 plot(x, y, col=c('blue',
111                 'slategray',
112                 'lightgreen',
113                 'honeydew3',
114                 'orange',
115                 'brown')[newcluster01$cluster],
116     main = 'clusters using 4 centers')
117 lines(newcluster01$centers,type='p',pch=16,col='black', cex = 2.5)

```



```

118
119 # try with k = 1
120 newcluster1 <- kmeans(x = points, centers = 1)
121 plot(x, y, col=c('blue',
122                 'slategray',
123                 'lightgreen',
124                 'honeydew3',
125                 'orange',
126                 'brown')[newcluster1$cluster],
127       main = 'clusters using 1 center')
128 lines(newcluster1$centers,type='p',pch=16,col='black', cex = 2.5)
129
130 # try with k = 2
131 newcluster2 <- kmeans(x = points, centers = 2)
132 plot(x, y, col=c('blue',
133                 'slategray',
134                 'lightgreen',
135                 'honeydew3',
136                 'orange',
137                 'brown')[newcluster2$cluster],
138       main = 'clusters using 2 centers')
139 lines(newcluster2$centers,type='p',pch=16,col='black', cex = 2.5)
140
141 # try with k = 2
142 newcluster21 <- kmeans(x = points, centers = cbind(c(5, 16), c(10, 21)))
143 plot(x, y, col=c('blue',
144                 'slategray',

```

```

145         'lightgreen',
146         'honeydew3',
147         'orange',
148         'brown')[newcluster21$cluster],
149     main = 'clusters using 2 centers')
150 lines(newcluster21$centers,type='p',pch=16,col='black', cex = 2.5)
151
152 # try with k = 6
153 newcluster6 <- kmeans(x = points, centers = 6)
154 plot(x, y, col=c('blue',
155                 'slategray',
156                 'lightgreen',
157                 'honeydew3',
158                 'orange',
159                 'brown',
160                 'yellow',
161                 'plum')[newcluster6$cluster],
162     main = 'clusters using 6 centers')
163 lines(newcluster6$centers,type='p',pch=16,col='black', cex = 2.5)
164
165 # try with k = 8
166 newcluster8 <- kmeans(x = points, centers = 8)
167 plot(x, y, col=c('blue',
168                 'slategray',
169                 'lightgreen',
170                 'honeydew3',
171                 'orange',

```

```

172         'brown',
173         'yellow',
174         'plum')[newcluster8$cluster],
175         main = 'clusters using 8 centers')
176 lines(newcluster8$centers,type='p',pch=16,col='black', cex = 2.5)
177
178 # try with k = 10
179 newcluster10 <- kmeans(x = points, centers = 10)
180 plot(x, y, col=colors[newcluster10$cluster],
181       main = 'clusters using 10 centers')
182 lines(newcluster10$centers,type='p',pch=16,col='black', cex = 2.5)
183
184 #brute force:
185 d <- function(x1, x2){
186     return(sqrt(sum( (x1 - x2)^2 )))
187 }
188 count = 1
189 list_of_clusters <- list()
190 keep_going = TRUE
191 for(i in 1:100000){
192     if(keep_going){
193         newcluster05 <- kmeans(x = points, centers = 4)
194         plot(x, y, col=c('blue',
195                         'slategray',
196                         'lightgreen',
197                         'honeydew3',
198                         'orange',

```

```

199         'brown')[newcluster05$cluster])
200     lines(newcluster05$centers,type='p',pch=16,col='black', cex = 2.5)
201     x1 <- as.numeric(newcluster05$centers[1,])
202     x2 <- as.numeric(newcluster05$centers[2,])
203     x3 <- as.numeric(newcluster05$centers[3,])
204     x4 <- as.numeric(newcluster05$centers[4,])
205     mindist <- min(c(
206         d(x1, x2), d(x1, x3), d(x1, x4),
207         d(x2, x3), d(x2, x4),
208         d(x3, x4)
209     ))
210     if(mindist < 5){
211         list_of_clusters[[count]] <- newcluster05
212         count = count + 1
213         keep_going = FALSE
214     }
215 }
216 }
217
218 # 2. Perform a K-means clustering with K = 2 and 1000 repetitions for the
219 #     wdbc data set. What classification accuracy would be achieved if the
220 #     clusters were used to predict the diagnosis in this data set?
221
222 wdbc_cluster <- kmeans_reps(wdbc[,2:ncol(wdbc)], 2, 1000)
223 acc <- confmatrix(wdbc$V2, wdbc_cluster$cluster)

```