

STATS 101C - Statistical Models and Data Mining - Homework 8

Darren Tsang, Discussion 4B

Produced on Thursday, Dec. 10 2020 @ 11:37:02 AM

Question 1 (Exercise 10 from Section 10.7)

Part A

```
set.seed(9999)
one <- data.frame()
two <- data.frame()
three <- data.frame()

for (i in seq(1, 20)){
  t1 <- c()
  t2 <- c()
  t3 <- c()
  for (j in seq(1, 50)){
    t1 <- c(t1, rnorm(1, mean = 0, sd = .5))
    t2 <- c(t2, rnorm(1, mean = -5, sd = .5))
    t3 <- c(t3, rnorm(1, mean = 5, sd = .5))
  }

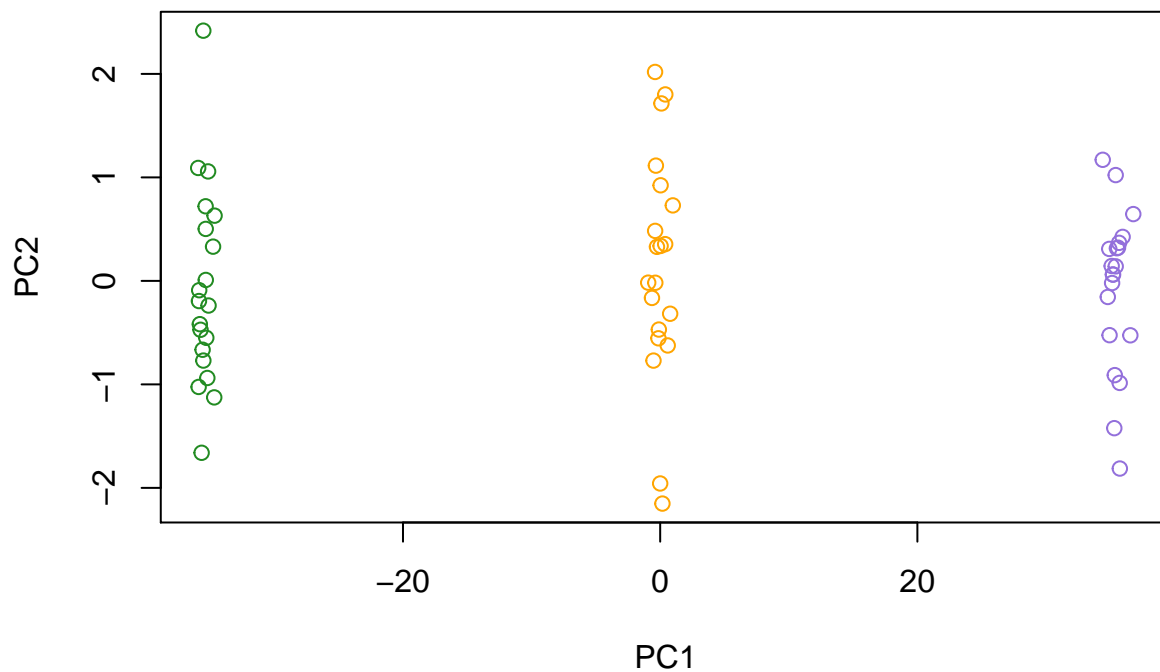
  one <- rbind(one, c(1, t1))
  two <- rbind(two, c(2, t2))
  three <- rbind(three, c(3, t3))
}

names(one) <- c('class', seq(1, 50))
names(two) <- c('class', seq(1, 50))
names(three) <- c('class', seq(1, 50))

data <- rbind(one, two, three)
```

Part B

```
data.pc <- prcomp(data[, -1])
plot(data.pc$x[, 1:2],
     col = c(rep("orange", 20),
              rep("mediumpurple", 20),
              rep("forestgreen", 20)))
```



Part C

```
set.seed(9999)
data.kmeans.three <- kmeans(data[, -1], centers = 3, nstart = 20)
table("true" = data$class,
      "predicted" = data.kmeans.three$cluster)
```

```
##      predicted
## true  1  2  3
##      1  0 20  0
##      2 20  0  0
##      3  0  0 20
```

It got the classes all correct.

Part D

```
set.seed(9999)
data.kmeans.two <- kmeans(data[, -1], centers = 2, nstart = 20)
table("true" = data$class,
      "predicted" = data.kmeans.two$cluster)
```

```
##      predicted
## true  1  2
##      1  0 20
##      2 20  0
##      3  0 20
```

It clumped the true 1's and 3's together into one cluster. The true 2's got clustered together perfectly.

Part E

```
set.seed(9999)
data.kmeans.four <- kmeans(data[, -1], centers = 4, nstart = 20)
table("true" = data$class,
      "predicted" = data.kmeans.four$cluster)
```

```
##      predicted
## true  1  2  3  4
##      1 14  0  0  6
##      2  0  0 20  0
##      3  0 20  0  0
```

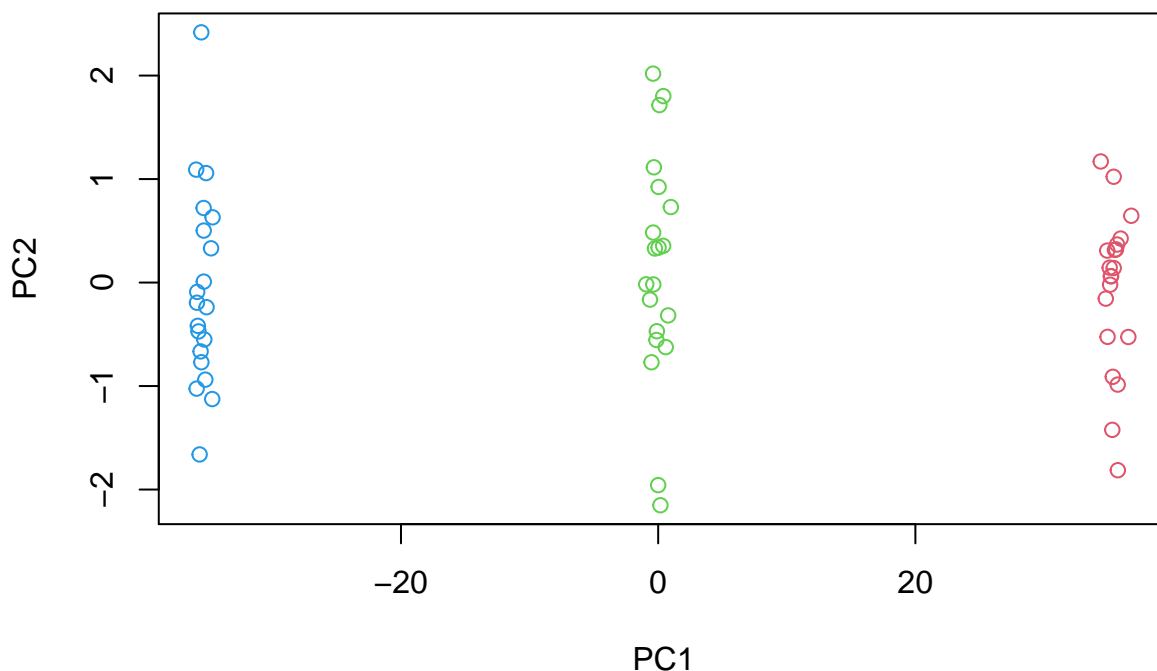
The true 2's and 3's got perfectly clustered into their own cluster. The true 1's got split into two clusters of size 14 and 6.

Part F

```
set.seed(9999)
pc.data.kmeans.three <- kmeans(data.pc$x[, 1:2], centers = 3, nstart = 20)
table("true" = data$class,
      "predicted" = pc.data.kmeans.three$cluster)
```

```
##      predicted
## true  1  2  3
##      1  0 20  0
##      2 20  0  0
##      3  0  0 20
```

```
plot(data.pc$x[, 1:2], col = pc.data.kmeans.three$cluster + 1)
```



It got the classes all correct.

Part G

```
set.seed(9999)
scaled.data.kmeans.four <- kmeans(scale(data[, -1]), centers = 3, nstart = 20)
table("true" = data$class,
      "predicted" = scaled.data.kmeans.four$cluster)
```

```
##      predicted
## true  1  2  3
##      1  0 20  0
##      2 20  0  0
##      3  0  0 20
```

It got the classes all correct. The results are similar to Part B, as they both got all of the classes correct.