Assignment 6 - Unsupervised Learning (Principal Component Analysis and Clustering)

Saha Debanshee Gopal

November 29, 2016

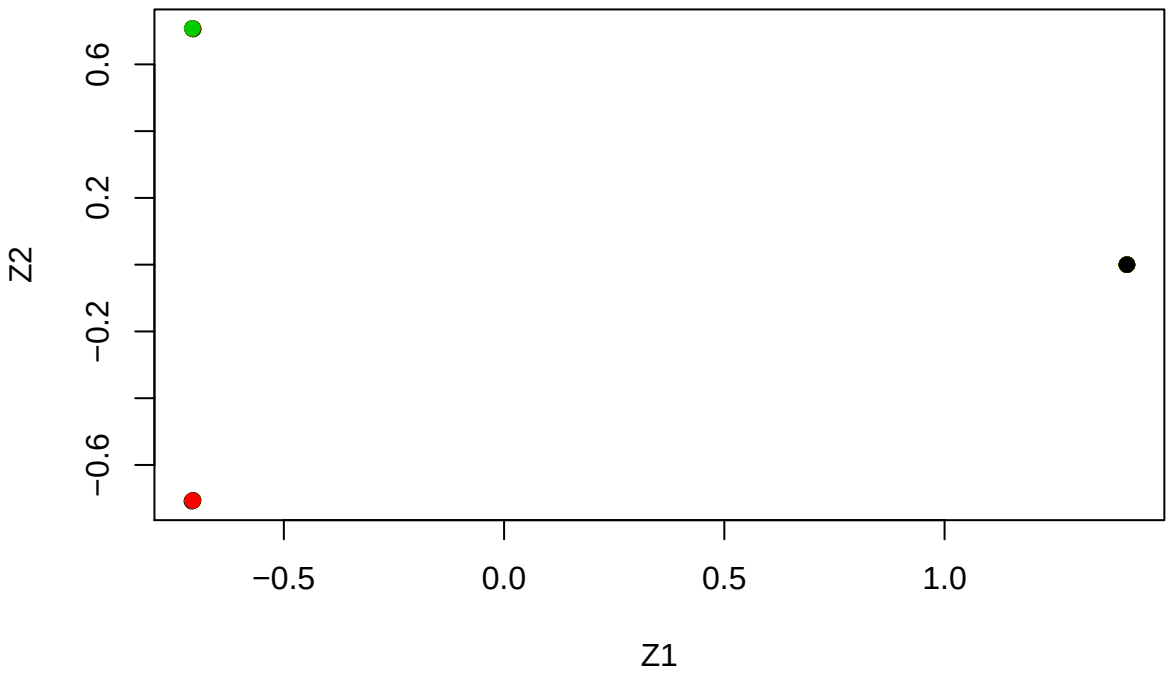
Q1. In this problem, you will generate simulated data, and then perform PCA and KK-means clustering on the data.

a)Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.

```
set.seed(2)
x <- matrix(rnorm(20 * 3 * 50, mean = 0, sd = 0.001), ncol = 50)
x[1:20, 2] <- 1
x[21:40, 1] <- 2
x[21:40, 2] <- 2
x[41:60, 1] <- 1
true.labels <- c(rep(1, 20), rep(2, 20), rep(3, 20))
```

b)Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, the return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.

```
pr.out <- prcomp(x)
plot(pr.out$x[, 1:2], col = 1:3, xlab = "Z1", ylab = "Z2", pch = 19)
```



c)Perform KK-means clustering of the observations with K=3K=3. How well do the clusters that you obtained in KK-means clustering compare to the true class labels ?

```
km.out <- kmeans(x, 3, nstart = 20)
table(true.labels, km.out$cluster)
```

```
##
## true.labels  1  2  3
##           1 20  0  0
##           2  0 20  0
##           3  0  0 20
```

The observations are perfectly clustered.

d)Perform KK-means clustering with K=2K=2. Describe your results.

```
km.out <- kmeans(x, 2, nstart = 20)
table(true.labels, km.out$cluster)
```

```
##
## true.labels  1  2
##           1 20  0
##           2  0 20
##           3 20  0
```

All observations of one of the three clusters is now absorbed in one of the two clusters.

e)Now perform KK-means clustering with K=4K=4, and describe your results.

```
km.out <- kmeans(x, 4, nstart = 20)
table(true.labels, km.out$cluster)
```

```
##
## true.labels  1  2  3  4
##           1  9  0  0 11
##           2  0 20  0  0
##           3  0  0 20  0
```

The first cluster is splitted into two clusters.

f)Perform KK-means clustering with K=3K=3 on the first two principal component score vectors, rather than on the raw data. That is, perform KK-means clustering on the 60x2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

```
km.out <- kmeans(pr.out$x[, 1:2], 3, nstart = 20)
table(true.labels, km.out$cluster)
```

```
##
## true.labels  1  2  3
##           1  0  0 20
##           2  0 20  0
##           3 20  0  0
```

All observations are perfectly clustered once again.

g)Using the "scale()" function, perform KK-means clustering with K=3K=3 on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (b) ? Explain.

```
km.out <- kmeans(scale(x), 3, nstart = 20)
table(true.labels, km.out$cluster)
```

```
##
## true.labels  1  2  3
##           1  8  2 10
##           2  0 19  1
##           3 11  1  8
```

We may see that we have worse results than with unscaled data, as scaling affects the distance between the observations.