IS 622 Week 8 Homework

Ben Arancibia

October 18, 2015

7.1.3

Suppose we have a d-dimensional Euclidean space. Consider vectors whose components are only $+1$ or -1 in each dimension. Note that each vector has length d^(1/2), so the product of their lengths (denominator in the formula for the cosine of the angle between them) is d. If we chose each component independently, and a component is as likely to be $+1$ as -1, what is the distribution of the value of the numerator of the formula (i.e., the sum of the products of the corresponding components from each vector)? What can you say about the expected value of the cosine of the angle between the vectors, as d grows large?

If we have a d-dimensional Euclidean space, a problem exists of "curse of dimensionality". In high dimensions almost all pairs are equally far away from one other. As defined in the problem, you have two equally likely possibilitys +1 and -1. Since they are equally likely the value of the summing would be 0 (1 + -1 = 0). For large d, the cosine of an angle is close to 0 because the summation is 0. If the cosine of an angle is 0 then the angle has to be close to 90 degrees.

7.2.1

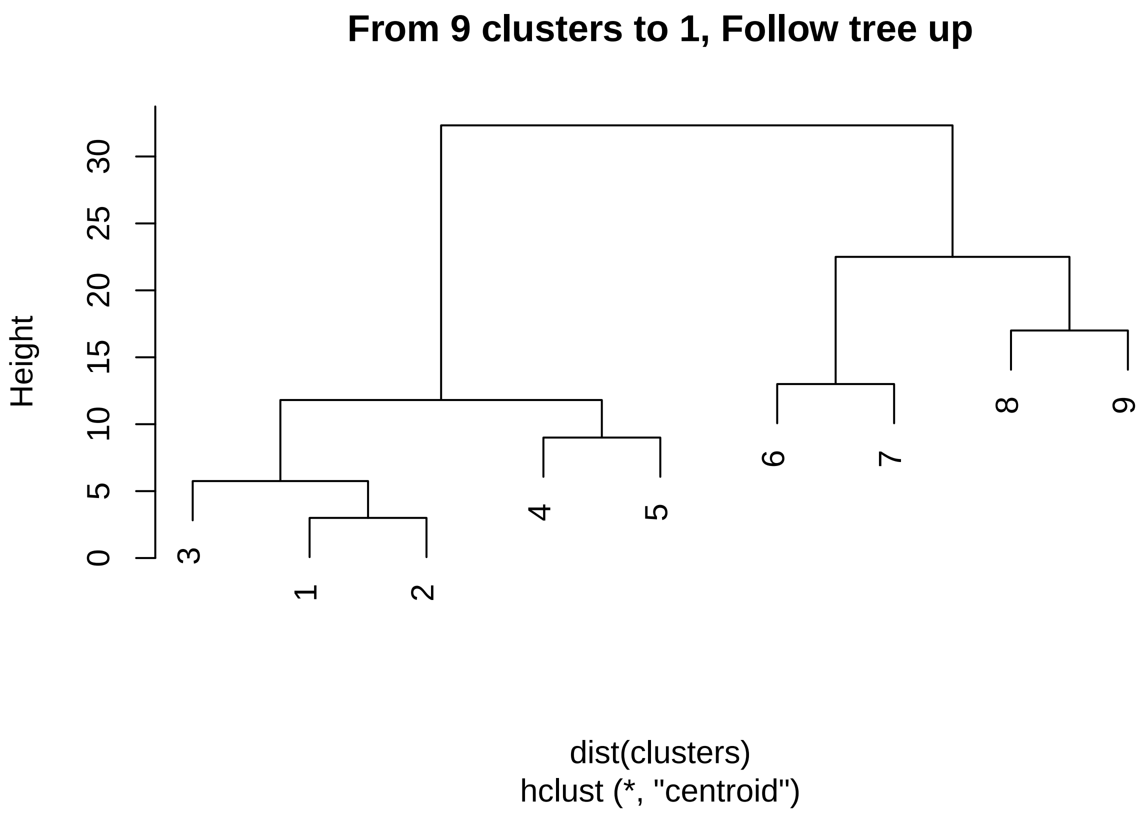
Perform a hierarchical clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81, assuming clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged.

```
clusters <- list( c(1), c(4), c(9), c(16), c(25), c(36), c(49), c(64), c(81))

hc <- hclust(dist(clusters), "cen")
hc$merge
```

```
##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3    1
## [3,]   -4   -5
## [4,]    2    3
## [5,]   -6   -7
## [6,]   -8   -9
## [7,]    5    6
## [8,]    4    7
```

```
plot(hc, main = "From 9 clusters to 1, Follow tree up")
```



7.2.2

How would the clustering of Example 7.2 change if we used for the distance between two clusters:

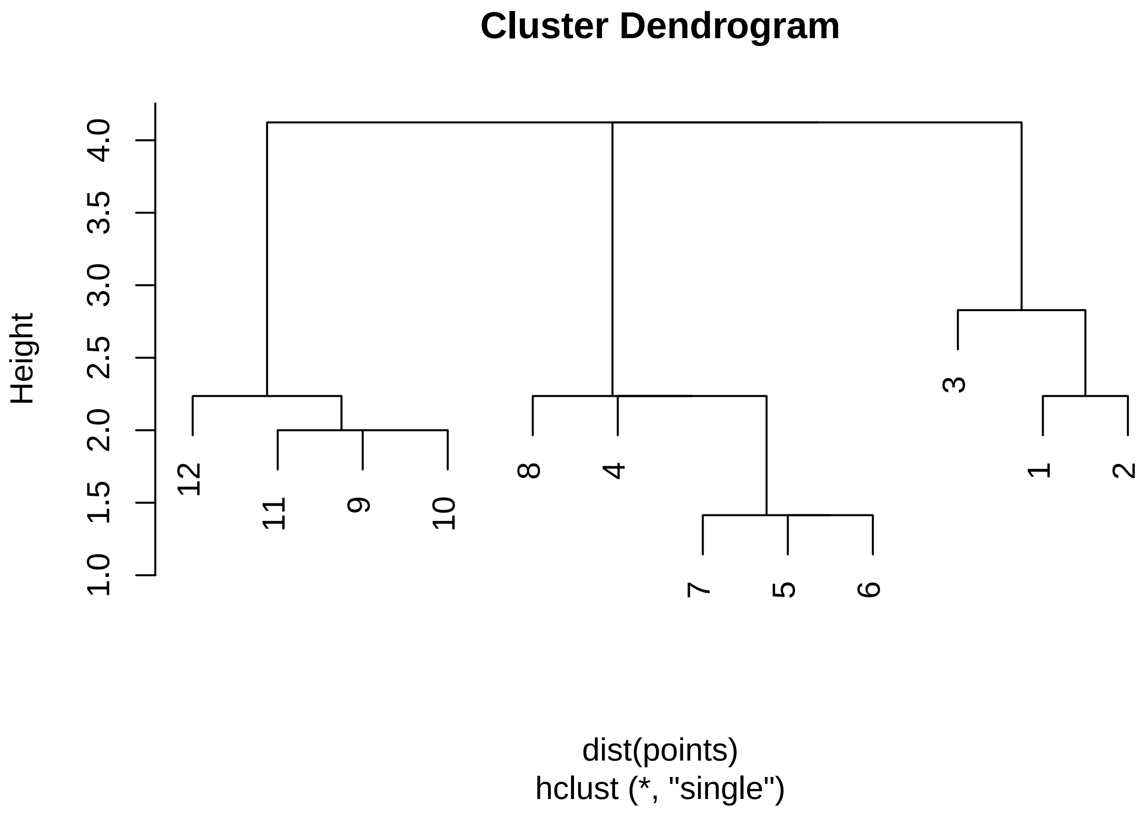
```r
points <- c(2,2,3,4,5,2,9,3,12,3,11,4,10,5,12,6,6,8,4,8,4,10,7,10)
points <- matrix(points, nrow=12, ncol=2, byrow=TRUE)
points <- as.data.frame(points)

points
```

```
##    V1 V2
## 1   2  2
## 2   3  4
## 3   5  2
## 4   9  3
## 5  12  3
## 6  11  4
## 7  10  5
## 8  12  6
## 9   6  8
## 10  4  8
## 11  4 10
## 12  7 10
```

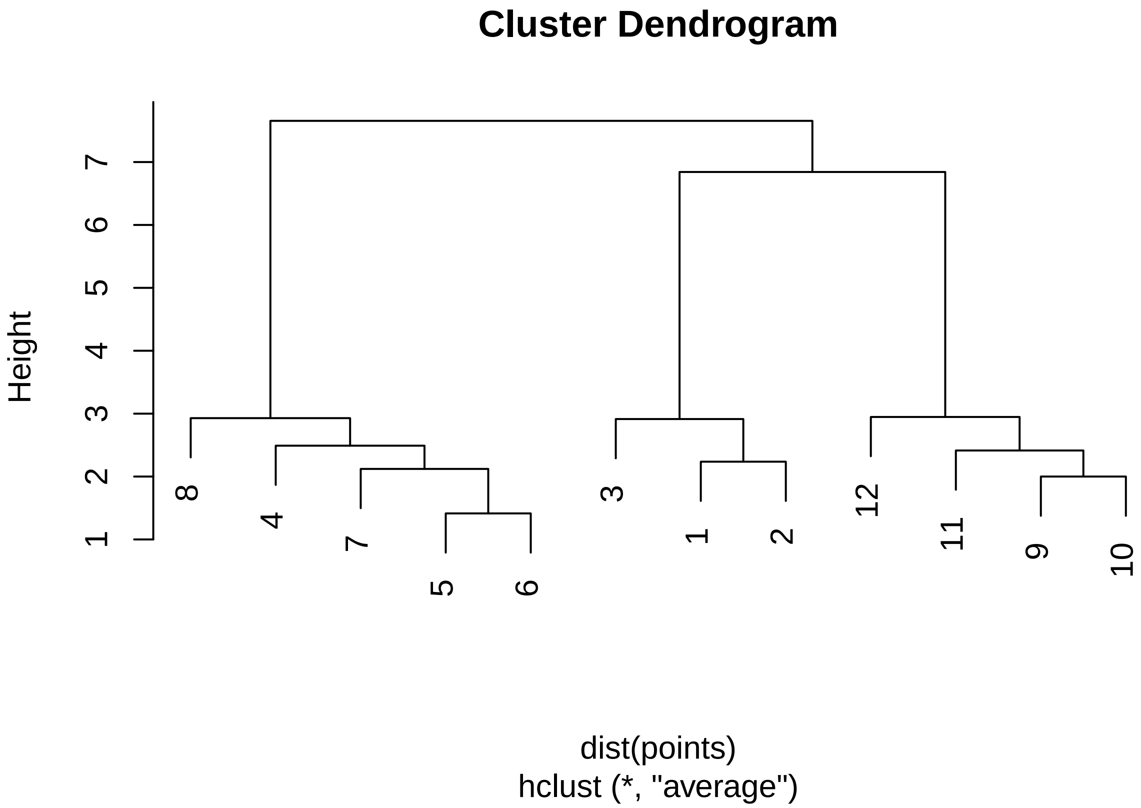
(a) The minimum of the distances between any two points, one from each cluster.

```
hc <- hclust(dist(points), "single")
plot(hc)
```



(b) The average of the distances between pairs of points, one from each of the two clusters.

```
hc <- hclust(dist(points), "average")
plot(hc)
```


Cluster Dendrogram
Height
1
2
3
4
5
6
7
8
4
7
5
6
3
1
2
12
11
9
10
dist(points)
hclust (*, "average")