IS622 Week9 - Clustering pt 2

James Quacinella

10/24/2015

Exercise 7.4.1 (section 7.4.3)

Consider two clusters that are a circle and a surrounding ring, as in the running example of this section. Suppose:

i. The radius of the circle is c.

ii. The inner and outer circles forming the ring have radii i and o, respectively.

iii. All representative points for the two clusters are on the boundaries of the clusters.

iv. Representative points are moved 20% of the distance from their initial position toward the centroid of their cluster.

v. Clusters are merged if, after repositioning, there are representative points from the two clusters at distance d or less.

In terms of d, c, i, and o, under what circumstances will the ring and circle be merged into a single cluster?

Answer

For the purpose of merging, only the inner radius would matter since we know that all representitive points are on the boundries. If the points on the inner boundry are > d, so are the points on the outer boundry.

The inner points will move from having radius i to $0.8i$ since it is reduced in size by 20%. The points on the circle, by the same reasonsing, move from c to $0.8c$.

We merge clusters when the representitive points are at a distance d or less. So the condition would be that $0.8i - 0.8c <= d$ or $0.8(i - c) <= d$.

Exercise 7.5.1 (section 7.5.5)

Using the cluster representation of Section 7.5.1, represent the twelve points of Fig. 7.8 as a single cluster. Use parameter k = 2 as the number of close and distant points to be included in the representation. Hint: Since the distance is Euclidean, we can get the square of the distance between two points by taking the sum of the squares of the differences along the x- and y-axes.

Answer

From section 7.5.1, the following features form the representation of a cluster:

1. N, the number of points in the cluster.

In this case, $N = 12$.

2. The clustroid of the cluster, which is defined specifically to be the point in the cluster that minimizes the sum of the squares of the distances to the other points; that is, the clustroid is the point in the cluster with the smallest ROWSUM.

```
# Copy points from figure
points <- matrix(c(2, 2,
                   3, 4,
                   5, 2,
                  4, 10,
                  7, 10,
                  4, 8,
                  6, 8,
                  10, 5,
                  12, 6,
                  11, 4,
                  9, 3,
                  12, 3), byrow=TRUE, nrow=12)
points
```

```
##       [,1] [,2]
##  [1,]    2    2
##  [2,]    3    4
##  [3,]    5    2
##  [4,]    4   10
##  [5,]    7   10
##  [6,]    4    8
##  [7,]    6    8
##  [8,]   10    5
##  [9,]   12    6
## [10,]   11    4
## [11,]    9    3
## [12,]   12    3
```

```
# Centroid is the point minimizing distance^2 to all other points in cluster
distances <- as.matrix(dist(points, upper=TRUE, diag=TRUE))^2
clustroid_idx <- which.min(apply( distances, 1, sum))
clustroid <- points[clustroid_idx, ]
clustroid
```

```
## [1] 6 8
```

Therefore, the clustroid is $(10, 5)$. This intuitivley / visually makes sense.

3. The rowsum of the clustroid of the cluster.

This would be calculated as follows:

```
rowsum <- sum(distances[clustroid_idx, ])
```

4. For some chosen constant k, the k points of the cluster that are closest to the clustroid, and their rowsums. These points are part of the representation in case the addition of points to the cluster causes the clustroid to change. The assumption is made that the new clustroid would be one of these k points near the old clustroid.

For k = 2, we need to find the two closest points to the clustroid. By looking at the distances matrix, we see the points with idx 5 and 6, which correspond to:

```
points[5, ] # 7,10
```

```
## [1]  7 10
```

```
points[6, ] # 4,8
```

```
## [1] 4 8
```

5. The k points of the cluster that are furthest from the clustroid and their rowsums. These points are part of the representation so that we can consider whether two clusters are close enough to merge. The assumption is made that if two clusters are close, then a pair of points distant from their respective clustroids would be close.

Same reasoning as above, we can manually inspect the distance matrix and find the points that are furthest away, which are points with idx 12 and 1:

```
points[12, ] # 12,3
```

```
## [1] 12  3
```

```
points[1, ] # 2,2
```

```
## [1] 2 2
```