Intelligent Data Mining - Exercise 6

Michael Debono Mrđen

7 December 2017

1 Assignment 1: Hierarchical clustering

a. 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 format: $\{(centroid_n : point_n, point_{n+1}), point_{n+2}\}$

   1. **Clusters**: $\{1, 4, 9, 16, 25, 36, 64, 81\}$
   2. Smallest distance is $|1 - 4| = 3$
   3. Combine 1 and 4 into a cluster with centroid $\frac{1+4}{2} = 2.5$
   4. **Clusters** $\{(2.5 : 1, 4), 9, 16, 25, 36, 49, 64, 81\}$
   5. Next smallest distance is $|2.5 - 9| = 6.5$
   6. Combine 1, 4, and 9 into a cluster with centroid $\frac{1+4+9}{3} = 4.7$
   7. **Clusters**: $\{(4.7 : 1, 4, 9), 16, 25, 36, 49, 64, 81\}$
   8. Next smallest distance is $|16 - 25| = 9$
   9. Combine 16 and 25 into a cluster with centroid $\frac{16+25}{2} = 20.5$
   10. **Clusters**: $\{(4.7 : 1, 4, 9), (20.5 : 16, 25), 36, 49, 64, 81\}$
   11. Next smallest distance is $|36 - 49| = 13$
   12. Combine 36 and 49 into a cluster with centroid $\frac{36+49}{2} = 42.5$
   13. **Clusters**: $\{(4.7 : 1, 4, 9), (20.5 : 16, 25), (42.5 : 36, 49), 64, 81\}$
   14. Next smallest distance is $|4.7 - 20.5| = 15.8$
   15. Combine 1, 4, 9, 16, and 25 into a cluster with centroid $\frac{1+4+9+16+25}{5} = 11$
   16. **Clusters**: $\{(11 : 1, 4, 9, 16, 25), (42.5 : 36, 49), 64, 81\}$
   17. Next smallest distance is $|64 - 81| = 17$
   18. Combine 64 and 81 into a cluster with centroid $\frac{64+81}{2} = 72.5$
   19. **Clusters**: $\{(11 : 1, 4, 9, 16, 25), (42.5 : 36, 49), (72.5 : 64, 81)\}$

20. Next smallest distance is $|42.5 - 72.5| = 30$
21. Combine 36, 49, 64, and 81 into a cluster with centroid $\frac{36+49+64+81}{4} = 57.5$
22. **Clusters**: $\{(11 : 1, 4, 9, 16, 25), (57.5 : 36, 49, 64, 81)\}$
23. Next (smallest) distance is $|57.5 - 11| = 46.5$
24. Combine all points into a cluster with centroid $\frac{1+4+9+16+25+36+49+64+81}{9} = 31.7$
25. **Clusters**: $\{(31.7 : 1, 4, 9, 16, 25, 36, 49, 64, 81)\}$

b. We can select clustroids for clusters, even if the space is Euclidean. Consider the three natural clusters in Fig. 7.2, and compute the clustroids of each, assuming the criterion for selecting the clustroid is the point with the minimum sum of distances to the other points in the cluster.

Cluster 1	(4,10)	(7,10)	(4,8)	(6,8)	Sum
(4,10)	0	3	2	$\sqrt{8}$	7.83
(7,10)	3	0	$\sqrt{13}$	$\sqrt{5}$	8.84
(4,8)	2	$\sqrt{13}$	0	2	7.60
(6,8)	$\sqrt{8}$	$\sqrt{5}$	2	0	7.06

Cluster 2	(3,4)	(2,2)	(5,2)	Sum
(3,4)	0	$\sqrt{5}$	$\sqrt{8}$	5.06
(2,2)	$\sqrt{5}$	0	9	11.24
(5,2)	$\sqrt{8}$	9	0	11.83

Cluster 3	(10,5)	(12,6)	(11,4)	(9,3)	(12,3)	Sum
(10,5)	0	$\sqrt{5}$	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{8}$	8.71
(12,6)	$\sqrt{5}$	0	$\sqrt{5}$	$\sqrt{18}$	3	11.71
(11,4)	$\sqrt{2}$	$\sqrt{5}$	0	$\sqrt{5}$	$\sqrt{2}$	7.30
(9,3)	$\sqrt{5}$	$\sqrt{18}$	$\sqrt{5}$	0	3	11.71
(12,3)	$\sqrt{8}$	3	$\sqrt{2}$	3	0	10.2

- Clustroid 1: (6,8)
- Clustroid 2: (3,4)
- Clustroid 3: (11,4)

2 Assignment 2: K-means clustering

a. For the three clusters of Fig. 7.8, compute the representation of the cluster as in the BFR Algorithm. That is, compute N, SUM, and SUMSQ.

	Cluster 1	Cluster 2	Cluster 3
N	4	3	5
SUM	(21,36)	(10,8)	(54,21)
SUMSQ	(117,328)	(38,24)	(590,95)

b. For the three clusters of Fig. 7.8, compute the variance and standard deviation of each cluster in each of the two dimensions.

Variance	Cluster 1	Cluster 2	Cluster 3
$i = 1$	$\frac{117}{4} - \frac{21}{4}^2 = 1.69$	$\frac{38}{3} - \frac{10}{3}^2 = 1.56$	$\frac{590}{5} - \frac{54}{5}^2 = 1.36$
$i = 2$	$\frac{328}{4} - \frac{36}{4}^2 = 1$	$\frac{24}{3} - \frac{8}{3}^2 = 0.89$	$\frac{95}{5} - \frac{21}{5}^2 = 1.36$

Standard deviation	Cluster 1	Cluster 2	Cluster 3
$i = 1$	$\sqrt{1.69} = 1.3$	$\sqrt{1.56} = 1.25$	$\sqrt{1.36} = 1.17$
$i = 2$	1	$\sqrt{0.89} = 0.94$	$\sqrt{1.36} = 1.17$

c. Suppose a cluster of three-dimensional points has standard deviations of 2, 3, and 5, in the three dimensions, in that order. Compute the Mahalanobis distance between the origin (0,0,0) and the point (1,-3,4).

$$\sqrt{\sum_{i=1}^{d} \left(\frac{p_i - c_i}{\sigma_i}\right)^2} = \left(\frac{1-0}{2}\right)^2 + \left(\frac{-3-0}{3}\right)^2 + \left(\frac{4-0}{5}\right)^2 = 1.89$$

3 Assignment 3: K-means clustering with Mahout

This assignment is done separately.