University of Glasgow | School of Computing Science

CSC3005 Laboratory/Tutorial 6 Solution: Clustering Analysis and Anamoly Detection

6. Theory on K-Means Clustering

Iteration 1: Assume A (2,2) and B(1,1) is the cluster mean C1 (2,2) and C2(1,1) respectively

Calculating the distance between A(2,2) and C1(2,2)

$$dist(A, C1) = \sqrt{(2-2)^2 + (2-2)^2} = 0$$

Calculating the distance between A(2,2) and C2(1,1)

$$dist(A, C2) = \sqrt{(2-1)^2 + (2-1)^2} = 1.4142$$

Calculating the distance between B(1,1) and C1(2,2)

$$dist(B, C1) = \sqrt{(1-2)^2 + (1-2)^2} = 1.4142$$

Calculating the distance between B(1,1) and C2(1,1)

$$dist(B, C2) = \sqrt{(1-1)^2 + (1-1)^2} = 0$$

Calculating the distance between C(1.5,0.5) and C1(2,2)

$$dist(C, C1) = \sqrt{(1.5-2)^2 + (0.5-2)^2} = 1.5811$$

Calculating the distance between C(1.5,0.5) and C2(1,1)

$$dist(C, C2) = \sqrt{(1.5-1)^2 + (0.5-1)^2} = 0.7071$$

Calculating the distance between D(3,1) and C1(2,2)

$$dist(D, C1) = \sqrt{(3-2)^2 + (1-2)^2} = 1.4142$$

Calculating the distance between D(3,1) and C2(1,1)

$$dist(D, C2) = \sqrt{(3-1)^2 + (1-1)^2} = 2$$

Calculating the distance between E(3,2) and C1(2,2)

$$dist(E, C1) = \sqrt{(3-2)^2 + (2-2)^2} = 1$$

Calculating the distance between E(3,2) and C2(1,1)

$$dist(E, C2) = \sqrt{(3-1)^2 + (2-1)^2} = 2.2361$$

Point	Distance from C1 mean (2,2)	Distance from C2 mean (1,1)	Point belongs to Cluster
A(2,2)	0	1.4142	C1
B(1,1)	1.4142	0	C2
C(1.5,0.5)	1.5811	0.7071	C2
D(3,1)	1.4142	2	C1
E(3,2)	1	2.2361	C1

Therefore, Cluster C1 contains

A(2,2)
D(3,1)
E(3,2)

Cluster 2 obtains

B(1,1)
C(1.5,0.5)

Recomputing the mean of cluster C1

$$C1\ mean = \left(\frac{2+3+3}{3}, \frac{2+1+2}{3}\right) = (2.667, 1.667)$$

Recomputing the mean of cluster C2

$$C2\ mean = \left(\frac{1+1.5}{2}, \frac{1+0.5}{2}\right) = (1.25, 0.75)$$

Repeating the whole process again at iteration 2

Calculating the distance between A(2,2) and C1(2.667,1.667)

$$dist(A, C1) = \sqrt{(2 - 2.667)^2 + (2 - 1.667)^2} = 0.7455$$

Calculating the distance between A(2,2) and C2(1.25,0.75)

$$dist(A, C2) = \sqrt{(2 - 1.25)^2 + (2 - 0.75)^2} = 1.4577$$

Calculating the distance between B(1,1) and C1(2.667,1.667)

$$dist(B, C1) = \sqrt{(1 - 2.667)^2 + (1 - 1.667)^2} = 1.7955$$

Calculating the distance between B(1,1) and C2(1.25,0.75)

$$dist(B, C2) = \sqrt{(1 - 1.25)^2 + (1 - 0.75)^2} = 0.3536$$

Calculating the distance between C(1.5,0.5) and C1(2.667,1.667)

$$dist(C, C1) = \sqrt{(1.5 - 2.667)^2 + (0.5 - 1.667)^2} = 1.5813$$

Calculating the distance between C(1.5,0.5) and C2(1.25,0.75)

$$dist(C, C2) = \sqrt{(1.5 - 1.25)^2 + (0.5 - 0.75)^2} = 0.3536$$

Calculating the distance between D(3,1) and C1(2.667,1.667)

$$dist(D, C1) = \sqrt{(3 - 2.667)^2 + (1 - 1.667)^2} = 0.7455$$

Calculating the distance between D(3,1) and C2(1.25,0.75)

$$dist(D, C2) = \sqrt{(3 - 1.25)^2 + (1 - 0.75)^2} = 1.7678$$

Calculating the distance between E(3,2) and C1(2.667,1.667)

$$dist(E, C1) = \sqrt{(3 - 2.667)^2 + (2 - 1.667)^2} = 0.4709$$

Calculating the distance between E(3,2) and C2(1.25,0.75)

$$dist(E, C2) = \sqrt{(3 - 1.25)^2 + (2 - 0.75)^2} = 2.1506$$

Point	Distance from C1 mean (2.667,1.667)	Distance from C2 mean (1.25,0.75)	Point belong to Cluster
A(2,2)	0.7455	1.4577	C1
B(1,1)	1.7955	0.3536	C2
C(1.5,0.5)	1.5813	0.3536	C2
D(3,1)	0.7455	1.7678	C1
E(3,2)	0.4709	2.1506	C1

Therefore, Cluster C1 contains

A(2,2)
D(3,1)
E(3,2)

Cluster 2 obtains

B(1,1)
C(1.5,0.5)

Recomputing the mean of cluster C1

$$C1\ mean = \left(\frac{2+3+3}{3}, \frac{2+1+2}{3}\right) = (2.667,1.667)$$

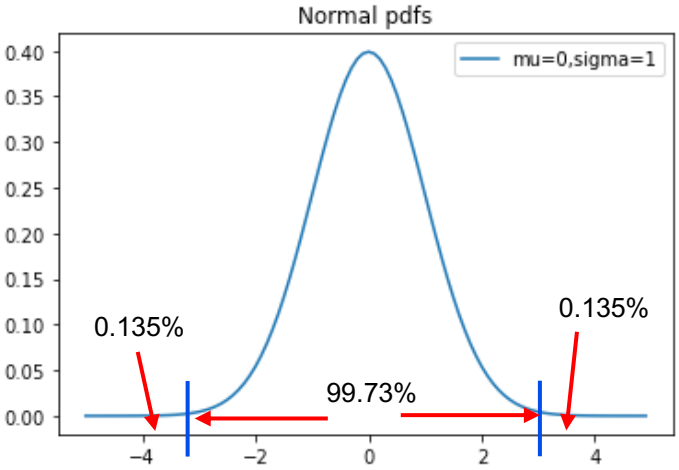
Recomputing the mean of cluster C2

$$C2\ mean = \left(\frac{1+1.5}{2}, \frac{1+0.5}{2}\right) = (1.25,0.75)$$

The cluster mean is the same as iteration 1 and does not change values. It is stable

7. Theory on Anomaly Detection using Statistical Approach

a. This question should have asked how many outliers we would have since the object of this question is to show that even a small probability of an outlier yields a large number of outliers for a large data set. The probability is unaffected by the number of objects. The probability is either 0.00135 for a single sided deviation of 3 standard deviations or 0.0027 for a double-sided deviation. Thus, the number of anomalous objects will be either 1,350 or 2,700.



b. There are thousands of outliers (under the specified definition) in a million objects. We may choose to accept these objects as outliers or prefer to increase the threshold so that fewer outliers result.

8. Theory on Anomaly Detection using K means Clustering Approach

a. The mean of the points is pulled somewhat upward from the center of the compact cluster by point D.

b. No, this point would become a cluster by itself.

c. If absolute distances are important. For example, consider heart rate monitors for patients. If the heart rate goes above or below a specified range of values, then this has an physical meaning. It would be incorrect not to identify any

patient outside that range as abnormal, even though there may be a group of patients that are relatively similar to each other and all have abnormal heart rates.