

Data Mining Assignment 3

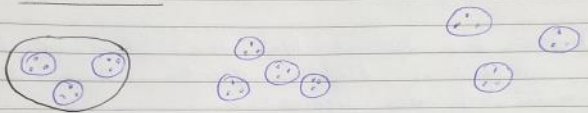
06 April 2018 14:07

Tan Chapter 8

DATE:


1.1

~~Exercise~~ Exercise 2



6.

(a) There are an infinite number of ways to split the circle into two clusters - just take any line that bisects the circle.



$K=2$

This line can make an angle $0^\circ \leq \theta \leq 180^\circ$ with the x-axis. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All these solutions will have the same, globally minimal, error.

DATE:

(b) $k=3$



~~$k=3$~~ The distance b/w the edges of the circles is slightly greater than the radii of the circles.

If you start with initial centroids that are real points, you will necessarily get this solution because of the limitation that the circles are more than one radius apart.

The bisector could have any angle, as above, and it could be other circle that is split.

All these solutions have the same globally minimal error.

(c) $k=3$



The distance b/w the edges of the circles is much less than the radii of the circles.

The three boxes ~~below~~ above show the three clusters that will result in the realistic case that the initial centroids are actual datapoint.

DATE:

(d) $k=2$



local minimum



Global minimum.

In both cases, the rectangles show the ~~set~~ clusters. In 1st pic, the two clusters are only a local minimum, however in second pic, the clusters represent a globally minimal solⁿ.

(e) $k=3$



Global minimum



local minimum

In the above solⁿ shown, the two top clusters are enclosed in two boxes, ~~and~~ while the third cluster is enclosed by the regions defined by a triangle and a rectangle. I believe that the second ~~set~~ solution is also possible, although it is a local minimum. While the

DATE:

two pie shaped cuts out of the larger circle are shown as meeting at the point, this is not necessarily the case. It depends on the exact positions and sizes of circles.

There could be a gap b/w the two pie shaped cuts which is filled by the third (larger) cluster. The boundary b/w two pie shaped cuts could actually be a line segment.

(11)

Answer • If the SSE of one attribute is low for all clusters then the variable is essentially a constant and of little use in dividing the data into groups.

- If the SSE for one attribute is relatively low for just one cluster, then this attribute helps define the cluster.
- If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is noise.
- If the attribute SSE is relatively high for one cluster, then it is at odds

DATE:

with the information provided by the attributes with low SSE that define the cluster. It means that this attribute does not help define the cluster.

- The idea is to eliminate attributes that have low or high SSE for all clusters, since they are useless for clustering. The attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes since they introduce lot of noise into the computation of the overall SSE.

(12) (a) Advantages & disadvantages of leader algorithm as compared to K-means —

The leader algorithm requires only a single scan of the data and therefore more computationally efficient since each is compared to the final set of centroids almost ~~once~~ once. Although the leader algorithm is order dependent, it always produces the same set of objects.

DATE:

Unlike k-means, leader algorithm is not possible to set the number of resulting clusters for the leader algorithm. Also, k-means almost always produces better quality clusters as measured by SSE.

(b) Ways in which leader algorithm might be improved.

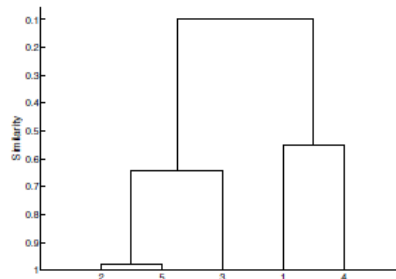
Use a sample to determine the distribution of distances between the points. The knowledge gained from this process can be used to more intelligently set the value of the threshold.

The leader algorithm could be modified to cluster for several thresholds during a single pass.

(16.)



(a) Single link.



(b) Complete link.