MONASH UNIVERSITY

Faculty of Information Technology

CSE5230 Data Mining

Semester 2, 2004

K-means Clustering

Purpose

The purpose of today's tutorial exercise is for you to gain understanding of how the K-means clustering algorithm works (see notes for lecture 4 for details of the algorithm). You will do this by using Microsoft Excel to build a semi-automated version of the algorithm.

Data Set

First, we need a set of data to cluster. We will consider a set of 20 data points, each with two attributes, X and Y. This corresponds to 20 records from a database, where each record has two numerical fields, X and Y. The data values are:

X	Y
1	2
2	4
1	3
2	8
2	3
5	7
3	3
4	6
2	1
2	2
4	7
4	8
5	6
1	4
6	6
6	5
5	3
6	9
5	8
7	7

Open an Excel spreadsheet, and enter these values in columns A and B, including the column headers in cells A1 and B1 respectively.

Cluster Centroids

The K-means algorithm computers cluster centroids. In order to get started, it needs initial guesses. We are going to try the algorithm with K = 2, i.e. two clusters. Enter the following initial guesses in cells L24:O25.

K1_X	K1_Y	K2_X	K2_Y
3	5	5	6

Distances to Centroids

The K-means algorithm calculates the distance from each data point to each cluster centroid. We can do this by entering the formula

=SQRT(($A2-$L$25)^2+($B2-$M$25)^2)

in cells C2:C21 to calculate the Euclidean distances from each point to centroid K1, and

=SQRT(($A2-$N$25)^2+($B2-$O$25)^2)

in cells D2:D21 to calculate the Euclidean distances from each point to centroid K2.

Note the absolute and relative cell references. Make sure that you understand how this works. Ask your tutor if necessary.

Assigning points to clusters

The K-means algorithm assigns each point to the cluster for which it is closer to the centroid. We can create columns showing the data points separated into two sets, one for each cluster. We will uses column E to show the cluster label for each row. Enter this formula in cell E2:

=IF((C2<D2),"K1","K2")

Note that this is an “If… Then… Else” structure. It is equivalent to:

```
IF (C2 < D2) THEN
     “K1” /* put “K1” in this cell */
ELSE
     “K2” /* put “K2” in this cell */
ENDIF
```

Now copy it to cells E3:E21. Ask your tutor if you do not know how to do this.

You should now see a cluster label for each point, either “K1” or “K2” in column E.

Now we will use columns F to I to show the data points separated into two sets. Enter the formula

=IF($E2="K1",$A2,"")

in cell F2. Copy it to cells F3:F21. Enter appropriate formulae in cells G2:I21 to copy the other data points across. Ask your tutor if you don’t know how to do this.

Enter column headings “cluster” in E1, “K1” in F1:G1 (merge the cells), and “K2” in H1:I1.

Cells E1:I21 should now show:

cluster	**K1**		**K2**	
K1	1	2		
K1	2	4		
K1	1	3		
K1	2	8		
K1	2	3		
K2			5	7
K1	3	3		
K2			4	6
K1	2	1		
K1	2	2		
K2			4	7
K2			4	8
K2			5	6
K1	1	4		
K2			6	6
K2			6	5
K1	5	3		
K2			6	9
K2			5	8
K2			7	7

New Cluster Centroids

The K-means algorithm computes new cluster centroids using the averages of the coordinates of the points assigned to each cluster. To do this, enter the formula

=AVERAGE(F2:F21)

in cell F22, and copy it to cells G22:I22. These cells should now show:

2.1	3.3	5.2	6.9

The new estimates for the cluster centroids are thus (2.1, 3.3) for cluster K1, and (5.2, 6.9) for cluster K2.

Updating the Cluster Centroid Cells

We have to be careful when updating the cluster centroid estimates in cells L25:O25, so that we do not create a circular reference. Select cells with F22:I22 with the mouse (they will be highlighted). Now choose “Copy” from the Edit menu. We now need to paste the *values* of these cells in as our new centroid estimates. Select cells L25:O25 with the mouse, and the select “Paste Special” from the Edit menu. From the dialog box, select “Values” from the “Paste” panel, and click OK. These new centroid estimates should now have caused all the values to be recalculated. This step (Updating the Cluster Centroid Cells) can now be repeated until the cluster centroids no longer change.

Questions and Exercises

- How many iterations are required for the centroids to converge to stable values?
- What happens if different initial guesses are used? Try:

  K1 = (1, 10) and K2 = (10, 10)
  K1 = (10, 10) and K2 = (11, 11)
  K1 = (3, 3) and K2 = (4, 4)

  Explain what you observe in each case.
- Use the chart tool to create an XY (scatter) graph of the data and cluster centroid estimates. This will allow you to watch the centroids move as the estimates are updated.