Intelligent Data Analysis

Martin Russell
School of Computer Science
Wednesday, 17 April 2019

Exercise sheet – week 6 – k-means clustering

1. (a) Let $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ be the set of 2-dimensional vectors given by:

$$x_1 = \begin{bmatrix} 0 \\ -5 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, x_3 = \begin{bmatrix} -2 \\ 3 \end{bmatrix}, x_4 = \begin{bmatrix} -4 \\ 7 \end{bmatrix}, x_5 = \begin{bmatrix} 3 \\ 1 \end{bmatrix},$$

$$x_6 = \begin{bmatrix} 4 \\ -2 \end{bmatrix}, x_7 = \begin{bmatrix} -1 \\ 6 \end{bmatrix}, x_8 = \begin{bmatrix} 5 \\ -6 \end{bmatrix}, x_9 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}, x_{10} = \begin{bmatrix} -5 \\ 10 \end{bmatrix}.$$

Using initial centroids $c_0 = \begin{bmatrix} -3 \\ 5 \end{bmatrix}$ and $c_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and the d_1 ('city block') metric, write down the new values of c_0 and c_1 after one iteration of k-means clustering is applied to the data set X. Show your calculations. [5]

(Recall that the 'city block' d_1 distance between two vectors $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ and $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ is given by: $d_1(v, w) = |v_1 - w_1| + |v_2 - w_2|$)

Solution:

First calculate the distance between each data point and each centroid (columns 2 and 3 in the table below). Then, for each data point determine the closest centroid (columns 4 and 5). Finally, gather together the data points closest to c_0 (columns 6 & 7) and take their average – this is the new value of c_0. Similarly for c_1.

It follows that the new values are:

$$c_0^{(1)} = \begin{bmatrix} -2.6 \\ 6 \end{bmatrix}, c_1^{(1)} = \begin{bmatrix} 2.6 \\ -2 \end{bmatrix}$$

					New c_0		New c_1			
	$d(x_n, c_0)$	$d(x_n, c_1)$	c_0	c_1						
x_1	13	9		1			13	9		
x_2	7	1		1			7	1		
x_3	3	5	1		3	5				
x_4	3	11	1		3	11				
x_5	10	2		1			10	2		
x_6	14	6		1			14	6		
x_7	3	7	1		3	7				
x_8	19	11		1			19	11		
x_9	3	5	1		3	5				
x_{10}	7	15	1		7	15				

(b) Let $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_5, \mathrm{x}_6, \mathrm{x}_7, \mathrm{x}_8\}$ be the set of two-dimensional vectors defined by:

$$x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathrm{x}_2 = \begin{bmatrix} 5 \\ -1 \end{bmatrix}, \mathrm{x}_3 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \mathrm{x}_4 = \begin{bmatrix} 4 \\ -3 \end{bmatrix},$$

$$\mathrm{x}_5 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \mathrm{x}_6 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \mathrm{x}_7 = \begin{bmatrix} 7 \\ -4 \end{bmatrix}, \mathrm{x}_8 = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$$

Let $c_0 = \begin{bmatrix} 2 \\ -4 \end{bmatrix}, c_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ be initial estimates of centroids for two clusters. Write down the new values of c_0 and c_1 after one iteration of k-means clustering. Use the Euclidean distance metric and show all of your calculations. [5]

Solution:

Proceed exactly as in 1(a) but use the new data points and centroids and Euclidean distance. The new centroids are:

$$c_0^{(1)} = \begin{bmatrix} 4.75 \\ -2.5 \end{bmatrix}, c_1^{(1)} = \begin{bmatrix} -0.75 \\ 0.5 \end{bmatrix}$$

2. (a) Let $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ be a set of 3-dimensional vectors given by:

$$x_1 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, x_2 = \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix}, x_3 = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}, x_4 = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}, x_5 = \begin{bmatrix} 9 \\ 4 \\ 2 \end{bmatrix}, x_6 = \begin{bmatrix} 7 \\ 3 \\ 7 \end{bmatrix},$$

and let $c_1^{(0)}$ and $c_2^{(0)}$be initial estimates of centroids for X given by:

$$c_1^{(o)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, c_2^{(0)} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix}.$$

(i) Calculate the new values $c_1^{(1)}$ and $c_2^{(1)}$ of these centroids after one iteration of k-means clustering. All of your distance calculations should use the “city block” (L_1) metric, given by: [5]

$$d_1(a, b) = d_1\left(\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \\ ba_3 \end{bmatrix}\right) = |a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3|$$

Solution:

Proceed as in 1(a) but with 3-dimensional data. The correct answer is:

$$c_1^{(1)} = \begin{bmatrix} 2.5 \\ 1.5 \\ 1.5 \end{bmatrix}, c_2^{(1)} = \begin{bmatrix} 6 \\ 3.25 \\ 3.5 \end{bmatrix}$$

(ii) Calculate the distortion $D(\{c_1^{(0)} c_2^{(0)}\}, X)$ for the centroids $c_1^{(0)}$ and $c_2^{(0)}$ and the set X.

Solution:

	$d\left(x_n, c_1^{(0)}\right)$	$d\left(x_n, c_2^{(0)}\right)$
x_1	5	10
x_2	9	6
x_3	10	5
x_4	6	9
x_5	15	8
x_6	17	6

The distortion $D(\{c_1^{(0)}c_2^{(0)}\},X)$ is the average distance between each data point and its closest centroid out of $\left\{c_1^{(0)}c_2^{(0)}\right\}$. This can be calculated using the distances that you calculated in Part (i), shown in the table above. The numbers are also in the table in the excel spreadsheet that goes with this solution sheet.

The distance between each data point and its closest centroid is highlighted in yellow. Hence:

$$D\left(\left\{c_1^{(0)}c_2^{(0)}\right\},X\right)=\frac{1}{6}(5+6+5+6+8+6)=\frac{36}{6}=6.$$

(iii) Calculate the distortion $D(\{c_1^{(1)}c_2^{(1)}\},X)$ for the centroids $c_1^{(1)}$ and $c_2^{(1)}$ and the set X. [4]

Solution:

This requires more calculations because we need the distance between each data point and its closest centroid out of $\left\{c_1^{(1)}c_2^{(1)}\right\}$. The distances are in the following table:

	$d\left(x_n,c_1^{(1)}\right)$	$d\left(x_n,c_2^{(1)}\right)$
x_1	1.5	7.75
x_2	4.5	3.75
x_3	4.5	6.25
x_4	1.5	6.75
x_5	9.5	5.25
x_6	11.5	4.75

From the table, $D\left(\left\{c_1^{(1)}c_2^{(1)}\right\},X\right)=\frac{1}{6}(1.5+3.75+4.5+1.5+5.25+4.75)=3.54$

(iv) In general, the set of K centroids $C^{(\infty)}$ to which the K-means clustering algorithm converges is only locally optimal. In what sense is it locally optimal and what choice determines the value of $C^{(\infty)}$? [3]

Solution:

It is locally optimal with respect to distortion. In other words there is a neighbourhood of the centroid set $C^{(\infty)}$ such that for any other centroid set C in that neighbourhood $D\left(C^{(\infty)},X\right)\leq D(C,X)$. However, $C^{(\infty)}$ is not in general a global optimum – i.e. there might be another centroid set C, not in the neighbourhood of $C^{(\infty)}$, such that $D\left(C^{(\infty)},X\right)\geq D(C,X)$.

3. Suppose that X is a set of data points in 5 dimensional space. For each value of $k=1,\dots,10$ a set of k initial centroids is chosen, then ten iterations of k-means clustering are applied to refine the set of centroids. Figure 1 shows the distortion as a function of the number of centroids at the end of this process. [6]

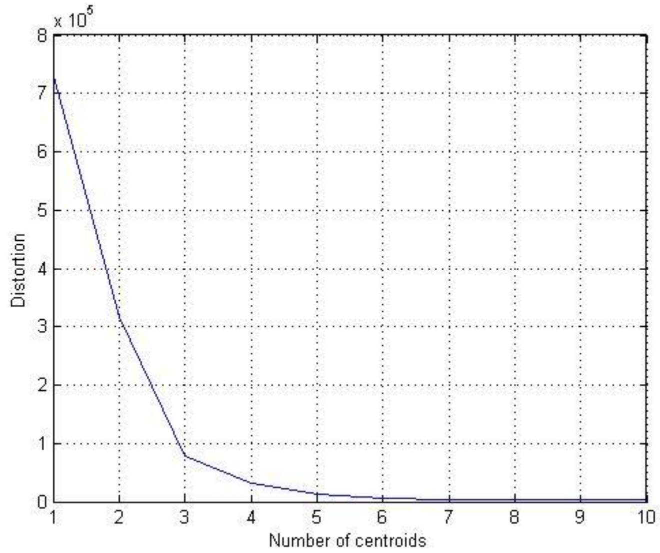

Figure 1: Distortion as a function of the number of centroids for the data set Xafter 10 iterations of k-means clustering.

Next the covariance matrix Y of X is calculated and its eigenvalue decomposition is computed as $Y = UDU^T$, where:

$$D = \begin{bmatrix} 0.0046 & 0 & 0 & 0 & 0 \\ 0 & 0.0661 & 0 & 0 & 0 \\ 0 & 0 & 0.4724 & 0 & 0 \\ 0 & 0 & 0 & 4.36 & 0 \\ 0 & 0 & 0 & 0 & 728.27 \end{bmatrix}$$

What can you conclude about the data set X? Justify your answer.

Solution:

The distortion decreases significantly as the number of centroids increases to 4, but after that the decrease is smaller. Beyond 6 centroids the decrease in distortion is very small. Therefore it seems that the data consists of a set of between 4 and 6 clusters (in fact there are 6 clusters in the data).

From PCA, the majority of the variance is in the direction of the 5th eigenvector and the next most important eigenvector is the 4th. Hence the data exists, approximately, as a 2 dimensional object embedded in 5 dimensional space.

Total marks **[28]**