Intelligent Data Analysis

Martin Russell
School of Computer Science
Thursday, 20 February 2020

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: $d_1(v, w) = |v_1 - w_1| + |v_2 - w_2|$)

(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]

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]

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

(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.

(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]

(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]

3. Suppose that X is a set of data points in 5 dimensional space. For each value of $k = 1, \ldots, 10$ 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]

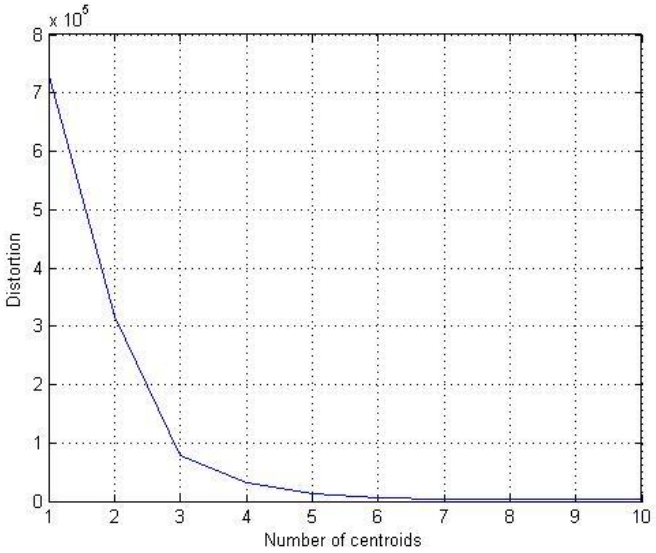

Figure 1: Distortion as a function of the number of centroids for the data set X after 10 iterations of k-means clustering.

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.

Total marks **[28]**