Machine Learning (SS2013)

Prof. Dr. Volker Sperschneider, Manuel Blum

Exercise Sheet 7

Exercise 7.1: K-means Clustering

Use the K-means algorithm and Euclidean distance to cluster the 8 data points given in Figure 1 into $K = 3$ clusters. The distance matrix based on the Euclidean distance is given in Table 1. The coordinates of the data points are:

$$x^{(1)} = (2,8), \quad x^{(2)} = (2,5), \quad x^{(3)} = (1,2), \quad x^{(4)} = (5,8),$$
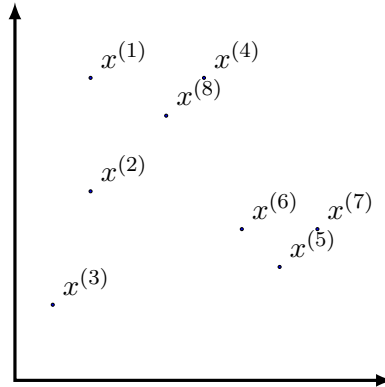$$x^{(5)} = (7,3), \quad x^{(6)} = (6,4), \quad x^{(7)} = (8,4), \quad x^{(8)} = (4,7).$$

Figure 1: Training data set for K-means clustering.

(a) Suppose you are initializing K-means using Forgy's method, that is, you initialize the cluster centers to K randomly chosen data points. Let's assume that points $x^{(3)}$, $x^{(4)}$ and $x^{(6)}$ were chosen. Perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids.

(b) Calculate the loss function $J\left(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K\right) = \frac{1}{m} \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$ before and after the first iteration of K-means using the initialization given in (a).

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
$x^{(1)}$	0	3.0000	6.0828	3.0000	7.0711	5.6569	7.2111	2.2361
$x^{(2)}$	3.0000	0	3.1623	4.2426	5.3852	4.1231	6.0828	2.8284
$x^{(3)}$	6.0828	3.1623	0	7.2111	6.0828	5.3852	7.2801	5.8310
$x^{(4)}$	3.0000	4.2426	7.2111	0	5.3852	4.1231	5.0000	1.4142
$x^{(5)}$	7.0711	5.3852	6.0828	5.3852	0	1.4142	1.4142	5.0000
$x^{(6)}$	5.6569	4.1231	5.3852	4.1231	1.4142	0	2.0000	3.6056
$x^{(7)}$	7.2111	6.0828	7.2801	5.0000	1.4142	2.0000	0	5.0000
$x^{(8)}$	2.2361	2.8284	5.8310	1.4142	5.0000	3.6056	5.0000	0

Table 1: Distance matrix for training data from Table 1.

(c) Now we will use the random partition method to initialize the cluster centroids. Suppose the initial random cluster assignments are

$$c^{(3)} = c^{(6)} = 1,$$

$$c^{(5)} = c^{(7)} = c^{(8)} = 2,$$

$$c^{(1)} = c^{(2)} = c^{(4)} = 3.$$

What are the coordinates of the initial cluster centers?

Exercise 7.2: Color Quantization

Color quantization finds a small number of representative colors within a given a picture. Each pixel yields one 3-dimensional pattern in the RGB color space. Using k-means we can cluster all the pixels of an image into k clusters and assign each pixel the color represented by its nearest cluster center. Thereby, an image containing millions of colors can be compressed to an image containing k different colors only.

(a) Implement the k-means algorithm.

(b) Load an image of your choice, treat each pixel as an individual 3-dimensional data point and cluster into k clusters (use low-resolution images to avoid long computation times).

(c) Assign each pixel the color value of its nearest cluster center.

(d) Visualize the result.