K-Means

Ha Phuong

January 21, 2022

1 Formula

Suppose we have a data set {x1,..., xN } consisting of N observations of a random D-dimensional Euclidean variable x. We introducing a set of D-dimensional vectors µk, where k = 1,...,K, in which µk is a prototype associated with the kth cluster.

An objective function

$$J = \sum_{n=1}^{N} \sum_{K=1}^{K} r_{nk} \| x_n - \mu_k \|^2 \tag{1}$$

$$with \begin{cases} r_{n,k} \in \{0,1\}, \vee n, k \\ \sum_{j=1}^{K} r_{n,k} = 1 \vee n \end{cases}$$

Our goal is to find values for the rnk and the µk so as to minimize J.Now consider the optimization of the µk with the rnk held fixed. The objective function J is a quadratic function of µk, and it can be minimized by setting its derivative with respect to µk to zero giving

$$2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \tag{2}$$

$$\mu_k = \frac{\sum_{n=1} r_{nk} x_n}{\sum_{n=1} r_{nk}} \tag{3}$$

So µk equal to the mean of all of the data points xn assigned to cluster k