K-Means Clustering

Konstantinos G. Derpanis

March 5, 2006

The *K-means algorithm* (Bishop, 1995) is an algorithm for identifying K groups/clusters of data points in multidimensional spaces. Suppose we have a set of N data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_i \in \mathbb{R}^D$, the basic goal is to find K groupings of the data points such that the intra-cluster distances of the points from their respective cluster centre, μ_i, is small. Formally, the objective is to recover the K class assignments that minimize the total intra-cluster squared Euclidean distance of each point to its cluster centre, μ_i:

$$J = \sum_{n=1}^{N} \sum_{i=1}^{K} o_{n,i} \|\mathbf{x}_n - \mu_i\|^2, \tag{1}$$

where $o_{n,i}$ is a binary variable that indicates the cluster assignment of the point, if $\mathbf{x}_n$ is assigned to cluster j then $o_{n,j} = 1$ and $o_{n,m} = 0$, for $m \neq j$.

In order to minimize (1) we need to find the set of cluster ownerships $\{o_{n,i}\}$ and the set of cluster centres $\{\mu_i\}$. Starting from a set of initial values of μ_i, this can be accomplished through an iterative procedure that interleaves between optimizing with respect to the cluster ownerships, keeping the cluster centres fixed and optimizing with respect to the cluster centres while keeping the ownerships fixed; see Algorithm 1 for a summary of K-means and Fig. 1 for an illustrative clustering example. This procedure is very similar to the *Expectation-Maximization (EM) algorithm* (Dempster, Laird & Rubin, 1977) for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data, however, in the case of EM soft assignments are sought as opposed to hard assignments as in K-means. Note that K-means is not guaranteed to yield a global optimum solution and the final solution is largely dependent on the initial set of cluster centres.

An inherent limitation of K-means is that the number of clusters has to be known a priori. Also, K-means is not robust to outlying data (i.e., data that does not belong to any of the clusters). Inclusion of gross outlying data may result in the estimated centres moving significantly away from the densest regions.

Algorithm 1 K-means Clustering

1: Initialize: Select a set of K candidate cluster centres.
2: Assign each data point to the closest cluster centre.
3: Set the cluster centres to the mean value of the points in each cluster.
4: Repeat Steps 2 and 3 for a fixed number of iterations or until there is no change in cluster assignments.

References

Bishop, C. (1995). *Neural Networks For Pattern Recognition, Oxford University Press.* Oxford: Oxford University Press.

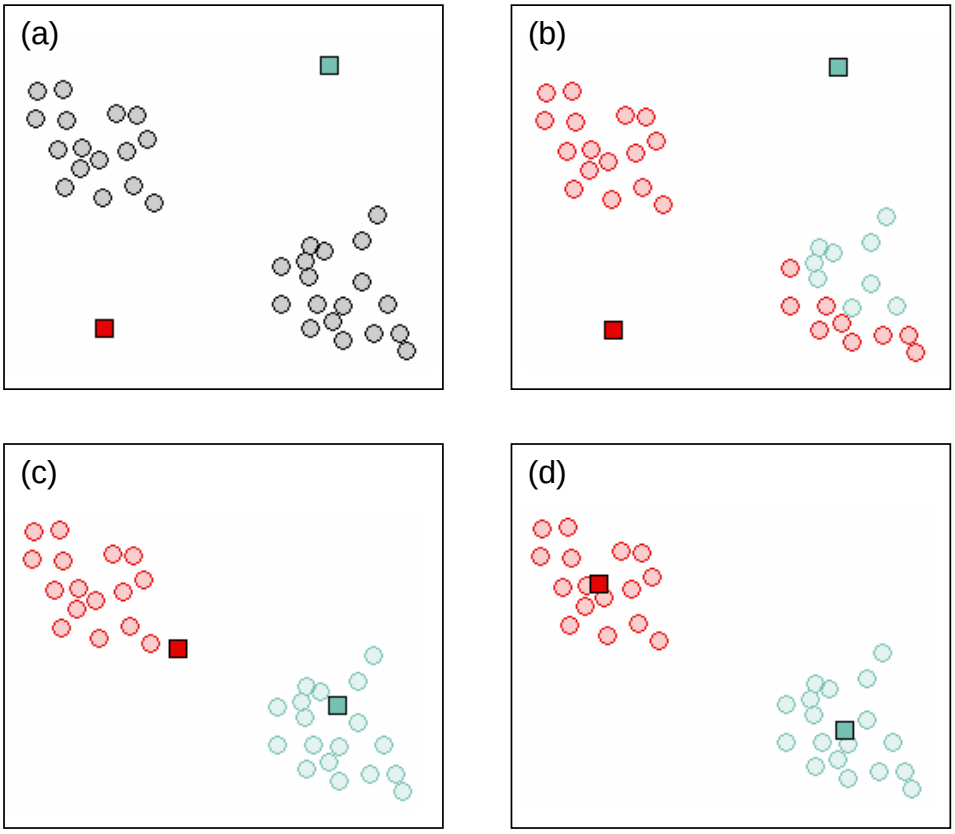

Figure 1: Illustration of the K-means algorithm. (a) Circles denote the data set and the squares the initial choices of the cluster centres. (b) Initial assignment of points to the closest cluster centre. (c) Cluster centres are recomputed. (d) Final assignment of points to clusters.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, B 39(1)*, 1–38.