Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Marjan Celikik

Search Engines WS 09/10

http://ad.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 12 — Solutions

Exercise 1 (Marjan)

Link to the solution in C++

Exercise 2 (Marjan)

M is set to 100 (the speed of the implementation above is independent of the value of M). The clustering algorithm is initialized with 5 random titles per cluster.

Exercise 3 (Marjan)

Run No.	*RSS*	*Precision*	*Recall*
1.	6001.9	59.0%	61.6%
2.	6045.6	57.4%	59.8%
3.	6051.0	51.1%	52.0%

Table 1: RSS, Precision and Recall for three different runs of the k-means ($k = 3$) clustering algorithm from above on the provided DBLP dataset.

Exercise 4 (Hannah)

Consider an input set consisting of two subsets X and Y. Let X have m points, and let Y have $\gamma \cdot m$ points, for some $\gamma < 1/2$. We assume the following distances between points: any two points in X have distance r from each other; any two points in Y have distance r from each other, and any point from X has distance $C \cdot r$ from any point in Y, where $C > 1$.

There is no embedding of such points into Euclidean space, such that we would have such distances. However, we can modify the distances such that such an embedding exists, and the proof that follows still goes through in the same way. Anyway, the exercise did not ask for points in Euclidean space, and note that k-means works for any set of points with given distance matrix, however weird that distance matrix is.

We will do the proof for the variant of k-means (here: 2-means), where a centroid is always one of the points from the input set, that is, from $X \cup Y$ in our case. The following argument also works without this assumption, but the proof becomes more cumbersome.

Under these assumptions, we can distinguish the following three possibilities where the optimal

centroids of a 2-clustering lie: both centroids in X, one centroid in X and one in $\mathcal{Y}$, and both centroids in Y. If one centroid lies in X and one in Y, the RSS value is

$$\mathrm{RSS}_1 = m \cdot r^2 + \gamma \cdot m \cdot r^2 = (1+\gamma) \cdot m \cdot r^2.$$

If both centroids lie in X, the RSS value is

$$\mathrm{RSS}_2 = m \cdot r^2 + \gamma \cdot m \cdot (Cr)^2 = (1+C^2\gamma) \cdot m \cdot r^2.$$

If both centroids lie in Y, the best achievable RSS value is

$$\mathrm{RSS}_3 = m \cdot (Cr)^2 + \gamma \cdot m \cdot r^2 = (C^2+\gamma) \cdot m \cdot r^2.$$

Clearly, RSS_1 is the best if only $C > 1$. For the corresponding best clustering one centroid lies in X and one in Y, and the clusters are X and Y.

Now split the set X into two equal-sized subsets X_1 and X_2, and shrink the intra-X_1 and the intra-X_2 distances to 0. With respect to the optimal clustering into X and Y from the previous paragraph, this shrinks some of the intra-X distances and leaves all inter-cluster distances. Hence, if consistency would hold for 2-means, the optimal 2-clustering should have clusters X and Y.

We will show that this is not the case. If one centroid lies in X and one in Y, the RSS value is now

$$\mathrm{RSS}'_1 = m/2 \cdot r^2 + \gamma \cdot m \cdot r^2 = (0.5+\gamma) \cdot m \cdot r^2.$$

If both centroids lie in X, the optimal solution has one centroid in X_1 and one in X_2, and the RSS value for that is

$$\mathrm{RSS}'_2 = 0 + \gamma \cdot m \cdot (Cr)^2 = (C^2\gamma) \cdot m \cdot r^2.$$

If both centroids lie in Y, the RSS value remains

$$\mathrm{RSS}'_3 = m \cdot (Cr)^2 + \gamma \cdot m \cdot r^2 = (C^2+\gamma) \cdot m \cdot r^2.$$

We now see that if only $\gamma < 1/(2 \cdot (C^2-1))$, we have $C^2\gamma < 0.5 + \gamma$, and then the best RSS value is achieved when both of the cluster centroids are in X, one in X_1 and one in X_2.

We have thus shown, that shrinking some of the intra-cluster distances can change the optimal clustering of 2-means, and therefore 2-means does not have the consistency property.