K-Means Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Requirements

The main requirements that a clustering algorithm should satisfy are:

- Scalability;
- Dealing with different types of attributes;
- Discovering clusters with arbitrary shape;
- Minimal requirements for domain knowledge to determine input parameters;
- Ability to deal with noise and outliers;
- Insensitivity to order of input records;
- High dimensionality;
- Interpretability and usability.

Problems

There are a number of problems with clustering. Among them:

- Current clustering techniques do not address all the requirements adequately (and concurrently);
- Dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- The effectiveness of the method depends on the definition of “distance” (for distance-based clustering);

- If an obvious distance measure doesn’t exist we must “define” it, which is not always easy, especially in multi-dimensional spaces;
- The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.