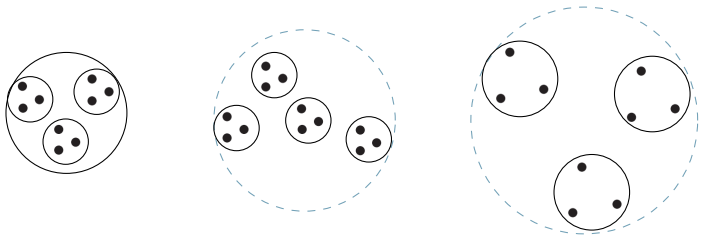
REI502M - Introduction to Data Mining

Solutions to homework 6

Elías Snorrason November 5, 2019

Problem 1

Find all well-separated clusters in the set of points shown in the figure below.



Solid circles indicate well-separated clusters. The dashed circles are *not* considered to be well-seperated clusters. This is due to their increased size (relative to the left-most cluster), as well as the distance between their edges (outermost points) is similar to their respective sizes.

Problem 2

Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

(a) Centroids should be equally distributed between more dense and less dense regions.

(b) More centroids should be allocated to the less dense region.

(c) More centroids should be allocated to the denser region.

Generally, one can more effectively cover half of the points by focusing on the dense regions, for a fixed K. **In the case of (c)**, a higher proportion of points will have lower squared errors in the dense regions and should thus minimize the SSE.

Problem 3

Consider the mean off a cluster of objects from a binary transaction data set. What are the minimum and maximum values of the components of the mean? What is the interpretation of components of the cluster mean? Which components most accurately characterize the objects in the cluster?

For a collection of N transactions, we can write the mean of the cluster component x_i:

$$\bar{x}_i = \frac{1}{N} \sum_{j}^{N} x_{ji}$$

where x_{ji} is either 0 or 1. The average value of these binary values lies somewhere between 0 and 1.
This means that each component has a minimum value of 0, and a maximum value of 1.

Since each component is an average of some binary values, they must represent occurrences of items in each transaction. E.g. support of an item (or itemsets) for transactions belonging to the cluster.

As binary data sets are normalized, the cluster centroid tends towards frequent non-zero values. On average, components with lower values have the lowest squared error with *all clusters*, and thus objects in these clusters are not characterized with low value components.

Problem 4

Give an example of a data set consisting of three natural clusters, for which (almost always) K-means would likely find the correct clusters, but bisecting K-means would not.

If the data set is symmetric, one of the natural clusters would have to lie on the symmetry "axis". This cluster will be split in the initial iteration and never combine into it's natural state.
E.g. the simplest case is one dimensional (numeric) data, such as the following set:

$$\{-10, -9 - 1, 1, 9, 10\}$$

The natural clusters in this data set are $\{-10, -9\}$, $\{-1, 1\}$ and $\{9, 10\}$.
With bisection, the data set will be split at 0, and the natural cluster $\{-1, 1\}$ will never occur.