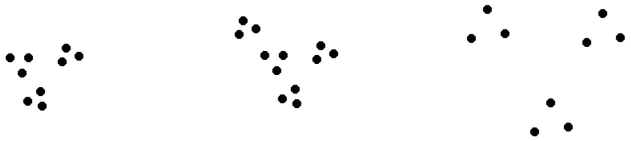
Data Mining

Week 10 Exercise Sheet

1. Find all well-separated clusters for the following sets of points:



2. Given k equally sized clusters, the probability that a randomly chosen initial centroid will come from any given cluster is $\frac{1}{k}$, but the probability that each cluster will have exactly one initial centroid is much lower.[1] In general, if there are k clusters and each cluster has n points, then the probability p of selecting in a sample of size k one initial centroid from each cluster is given as follows (this assumes sampling with replacement):

$$p = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } k \text{ centroids}} = \frac{k!n^k}{(kn)^k} = \frac{k!}{k^k}.$$

e.g. the chance of having one initial centroid from each of 3 clusters is $\frac{3!}{3^3} = 2/9$.

(a) Plot the probability of obtaining one point from each cluster in a sample of size k for values of k between 2 and 100.

(b) For k clusters where $k = 10$, 100 and 1000, find the probability that a sample of size $2k$ contains at least one point from each cluster.

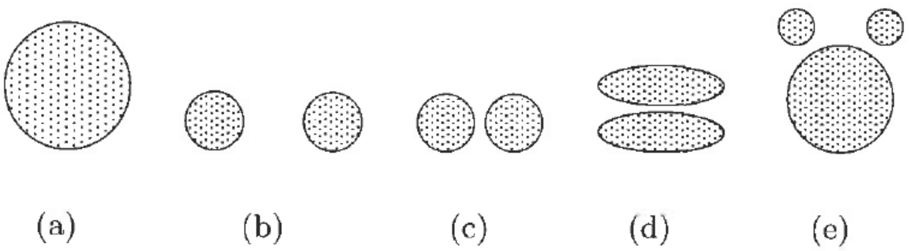
3. For the following sets of two-dimensional points,

   1. provide a sketch of how they would be split into clusters by k-means (for the given number of clusters); and
   2. indicate approximately where the resulting centroids would be.

You may assume that we are using the squared objective function. If you think there is more than one possible solution, then indicate whether each solution is a global or local minimum.

Note that the label of each diagram below corresponds to the respective part of this question (i.e. diagram (a) relates to part (a), diagram (b) to part (b), and so on.

[1] It is clear that having one initial centroid in each cluster is a good starting point for the k-means algorithm.



(a) $k = 2$

Assuming the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters?

What can you say about the positions of the two centroids? You do not need to give exact centroid locations, just a qualitative description.

(b) $k = 3$

The distance between the edges of the circles is slightly greater than the radii of the circles.

(c) $k = 3$

The distance between the edges of the circles is much less than the radii of the circles.

(d) $k = 2$

(e) $k = 3$

Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.

4. Consider the mean of a cluster of objects from a binary transaction data set.

   (a) What are the minimum and maximum values of the components of the mean?

   (b) What is the interpretation of components of the cluster mean?

   (c) Which components most accurately characterise the objects in the cluster?

5. 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.

6. Would the cosine measure be an appropriate similarity measure to use with k-means clustering for time series data? Why, or why not? If not, what similarity measure would be more appropriate?

7. Total SSE is the sum of the SSE for each attribute. What does it mean if the SSE for one variable is low for all clusters? What if it is low for just one cluster? What if it is high for all clusters? What if it is high for just one cluster? How could you use the per variable SSE information to improve clustering?