

Brian Tusi
October 30, 2008
Comp 578

Assignment 9
Chapter 8, #1-6

1. Consider a data set consisting of 2^{20} data vectors, where each vector has 32 components and each component is a 4-byte value. Suppose that vector quantization is used for compression and that 2^{16} prototype vectors are used. How many bytes of storage does that data set take before and after compression and what is the compression ratio?

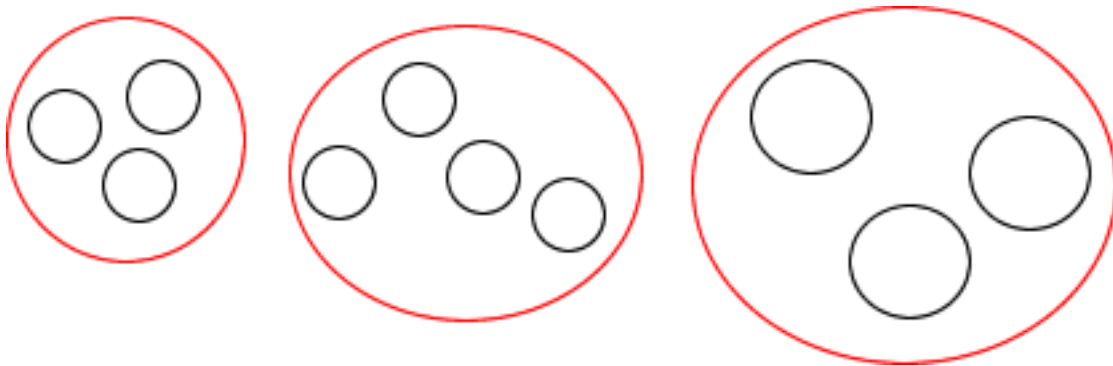
Before: $2^{20} * 32 * 4 = 134,217,728$ bytes

After: $2^{16} * 32 * 4 = 8,388,608$ bytes

This leaves us with a 16:1 compression ratio.

2. Find all well-separated clusters in the set of points shown in Figure 8.35.

Unfortunately, I don't have a scanner available to grab the exact original image. So, the following image is a mockup I drew myself.



Each black circle contains 3 points laid out as shown in the figure in the book. These clusters are well-separated from each other. Once we add in hierarchical clustering, we get the red circles – these groups contain the original black-outlined groups, and each is well-separated from the other.

3. Many partitioning clustering algorithms that automatically determine the number of clusters claim this is an advantage. List two situations in which this is not the case.

Sometimes, a particular number of clusters is needed, during which time an automatically-chosen cluster count may not be exactly what we want. For example, if a school is attempting to group children into classes based on a particular trait (such as their skill in reading), the school may already know in advance that they have exactly 5 classes to fill. Therefore, they know that exactly 5 groups are necessary – too many and there won't be enough room or teachers, and too few groups will make the classes too large.

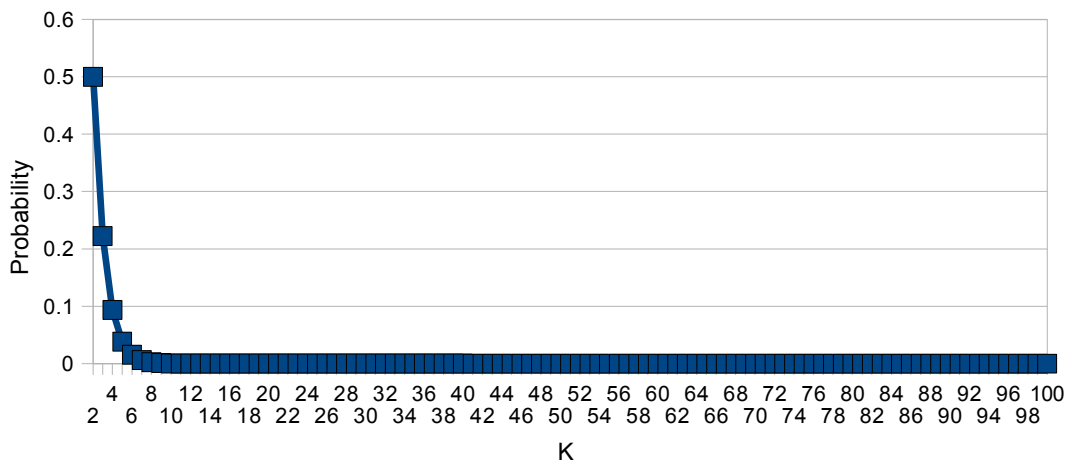
Other times, we might want to increase our efficiency, such as in embedded applications. When an automatically-chosen quantity of groups is used, the algorithm must calculate the best possible scenario

– is 3 groups ideal, or is 4? The algorithm needs to check these, taking up valuable CPU cycles. A finite number allows the algorithm to simply place them and calculate the distances to the centroid to create groups.

4. Given K equally-sized clusters, the probability that a randomly chosen initial centroid will come from any given cluster is 1/K, but the probability that each cluster will have exactly one initial centroid is much lower. (It should be clear that having one initial centroid in each cluster is a good starting situation for K-means.) In general, if there are K clusters and each cluster has n points, then the probability, p, of selecting in a sample size K one initial centroid from each cluster is given by Equation 8.20. (This assumes sampling with replacement.) From this formula we can calculate, for example, that the chance of having one initial centroid from each of four clusters is $4!/4^4 = 0.0938$.

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.

Unfortunately, my graphing utility's precision isn't too high – the vast majority of the points were rounded down to 0. However, the trend is obvious here – the probability quickly drops, and the limit is 0 as $K \rightarrow \text{infinity}$.



b) For K clusters, K = 10, 100, and 1000, find the probability that a sample of size 2K contains at least one point from each cluster. You can use either mathematical methods or statistical simulation to determine the answer.

If there are twice as many items being chosen, the probability is twice as high.

For K=10, $(2 * 10!) / (10^{10}) = 0.000726$

For K=100, $(2 * 100!) / (100^{100}) = 1.86 \times 10^{-42}$

For K=1000, $(2 * 1000!) / (1000^{1000}) = 0$ (my calculator cannot go this small...)

5. Identify the clusters in Figure 8.36 (p.560) using the center-, contiguity-, and density-based calculations. Also indicate the number of clusters for each case and give a brief introduction of

your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN.

(a) This appears to be an example of a density-based cluster, such as that with DBSCAN, or a combination of it combined with a type of agglomerative hierarchical clustering. DBSCAN throws out extraneous data, such as those outside the circles, and treats them as noise. Contiguity-based clustering, however, seems prevalent as well – note how the entire region is inside a rectangle, while the two more-dense areas have been surrounded with their own regions.

(b) This one can't be center-based, as even though the middle region is a centroid, the outer ring is not due to its hollow center. Therefore, it appears to be a density-based calculation – the outer ring is more dense than the sparse middle white ring, and DBSCAN is the only one which allows this nested ring formation.

(c) This diagram appears to be a center-based clustering method, such as K-means. The three triangular areas all can represent a 3-region, centroid-like area.

(d) This diagram appears to be contiguity-based, as each of the two parts has line segments which are formed by closely-related points.

6. 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. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 8.37 matches the corresponding part of this question; e.g., Figure 8.37(a) goes with part (a).

(a) K=2. Assuming that 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? (Again, you don't need to provide exact centroid locations, just a qualitative description.)

This figure would have two centers equidistant from the center of the circle. There are an infinite number of ways to place the regions around the center, but since the points are equally distributed, we would want them to be equidistant in order to minimize error.

(b) K=3. The distance between the edges of the circles is slightly greater than the radii of the circles.

I'm not too sure on this one, but perhaps we could have one centroid in one circle of the diagram, and another circle in another, but this doesn't seem to be like an appropriate solution.

(c) K=3. The distance between the edges of the circles is much less than the radii of the circles.

This *might* be able to be accomplished by using 3 circles side-by-side. Though the middle would be empty, the problem description seems to say that this will not be a problem since the distance is very (relatively) small.

(d) $K=2$.

One centroid in the center of each circle would do the trick here. Error would be minimized by containing our centroids to each be a circle completely contained inside the oval.

(e) $K=3$. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.

Imagine if there were 3 centroids in the same configuration as the diagram. Then imagine that those on the top (Mickey Mouse's ears) were enlarged, and the lower was shrunk.