Clustering Methods

Exercises 6/7, 8.5.2017

1. K-means works better with datasets having overlap between the clusters. Can we utilize this fact by artificially introduce overlap? How exactly? Bonus points if you manage to show this in practice.

2. Select your favorite data set. Calculate how many active centroids there are at every iteration in fast K-means. Plot the results as graph. Calculate also for how many points full search is required, and for how many partial search. Estimate the amount of distance calculations saved.

3. Calculate the two possible z-values for point (11,4). Include notes on how you did it.

4. Write a program that constructs a KNN graph by brute force or any other algorithm, and uses it for the assignment step of k-means. Test with S4. (a) Did you achieve a speedup? (b) Was there a difference in quality?

5. Upload your newest version of the clustering algorithm to Sami. (samisi@cs.uef.fi) by 8.5. by 10.00 latest. In your email, have title “Clustering project work”.