CS 4342/5342

Project on Clustering; points 200

Due: Apr 19 (submission via Canvas)

Implement the bisecting k-means clustering algorithm in a language of your choice (e.g., C, C++, Java, Python, …) (do not use any software package where this is available as a ready-to-use component). Consider only two dimensions. Submit the source code, input data and clustering results as a single Word or pdf file. Use 20 randomly generated two-dimensional real-valued data points in the square 1.0 <= x, y <= 100.0. Show results for two separate cases: k = 2 and k = 4. Show the effect on clustering of using two different distance measures (Euclidean and Manhattan). In the report, print the intra-cluster distance of each cluster, the sum of all intra-cluster distances for all the clusters, and the minimum and maximum distances between pairs of clusters (using the inter-cluster distance metric, namely min d_{ij} or max d_{ij} where points i and j belong to two different clusters) for the final solution.

A simple implementation of bisecting k-means will do; if, however, you are using any clever tricks, add a little explanation. Please prepare a very simple graphical output showing the clusters. Nothing fancy is needed – maybe you can plot the points on an MS excel sheet and just hand-draw the cluster boundaries (if you cannot put everything together in a single pdf, submit a hard copy in class).

Make sure that your results are reproducible – do not use any system-generated values for your initial (or other) seed(s) for the random number generator. Store all your seeds etc. in a file so that any single run can be replicated at will.