CS 4342/5342

Project on density-based clustering; points 200

Due: May 1 (submission via Canvas)

Implement the basic DBSCAN 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. Create your own points (do not use any database). Submit the source code, input data and clustering results as a single Word or pdf file. Consider three separate cases:

(a) 100 randomly generated two-dimensional real-valued data points in the square 1.0 <= x, y <= 100.0;
(b) 10 randomly generated two-dimensional real-valued data points in each of the squares 1.0 <= x, y <= 30.0 and 70.0 <= x, y <= 99.0;
(c) 10 randomly generated two-dimensional real-valued data points in the square 1.0 <= x, y <= 30.0 and 50 randomly generated two-dimensional real-valued data points in the square 70.0 <= x, y <= 99.0.

A simple implementation of the simple DBSCAN will do; if, however, you are using any clever tricks, add a little explanation. Please explain your choice of *Eps* and *MinPts*. 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.