CS6220 Unsupervised Data Mining

HW2B Clustering: DBSCAN, Hierarchical Clustering

Make sure you check the syllabus for the due date. Please use the notations adopted in class, even if the problem is stated in the book using a different notation.

We are not looking for very long answers (if you find yourself writing more than one or two pages of typed text per problem, you are probably on the wrong track). Try to be concise; also keep in mind that good ideas and explanations matter more than exact details.

Submit all code files Dropbox (create folder HW1 or similar name). Results can be pdf or txt files, including plots/tabels if any.

"Paper" exercises: submit using Dropbox as pdf, either typed or scanned handwritten.

DATATSET **: 20 NewsGroups : news articles**

DATATSET **: MNIST : digit images**

https://en.wikipedia.org/wiki/MNIST_database
http://yann.lecun.com/exdb/mnist/

DATATSET **: FASHION : 28x28 B/W images**

DATATSET **: UCI/Household**

PROBLEM 5: DBSCAN on toy-neighborhood data

You are to cluster, and visualize, a small dataset using DBSCAN epsilon = 7.5, MinPts = 3). You have been provided a file, dbscan.csv, that has the following columns for each point in the dataset:

- cluster originally empty, provided for your convenience pt a unique id for each data point
- x point x-coordinate
- y point y-coordinate
- num neighbors number of neighbors, according to the coordinates above neighbors the id’s of all neighbors within

As you can see, a tedious O(n^2) portion of the work has been done for you. Your job is to execute, point-by-point, the DBSCAN algorithm, logging your work.

PROBLEM 6: DBSCAN on toy raw data

Three toy 2D datasets are provided (or they can be obtained easily with scikit learn) circles; blobs, and moons. Run your own implementaion of DBSCAN on these, in two phases.

PROBLEM 7: DBSCAN on real data

Run the DBSCAN algorithm on the 20NG dataset, and on the FASHION dataset, and the HouseHold dataset (see papers), and evaluate results. You need to implement both phases (1) neighborhoods creation, (2) DBSCAN. Explain why/when it works, and speculate why/when not. You need to trial and error for parameters epsilon and MinPts

DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN
DBSCAN Revisited:Mis-Claim, Un-Fixability, and Approximation

EXTRA CREDIT: Using class labels (cheating), try to remove/add points in curate the set for better DBSCAN runs

PROBLEM 8: Hierarchical Clustering

Use a library to execute hierarchical clustering on MNIST dataset, evaluate the clusters.