277B: Machine Learning Algorithms

Homework assignment #5: Clustering

Assigned February 27 and Due March 8

1. KMeans. (10pt) We will now examine unsupervised learning for classification on a data set of chemical compounds. In compounds.csv, 150 organic compounds which belong to 3 different types (phenol, ether and amide) were tested upon with 4 different testing reagents (denoted reagents A-D). We would like to cluster data points by unsupervised learning, where we would not use the true label to guide classification such as using a cost function, instead we directly learn from the given features themselves.

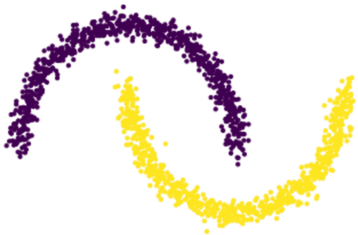
(a) (2pt) Rescale the features to a value between 0 and 1 by dividing the max of that feature. Visualize the data and comment on which features are correlated.

(b) (4pt) Do KMeans clustering with *K=2,3 and 4* clusters. Visualize your result (you can select 2 features to do visualization) and comment on which K value make the most sense to you according to the visualization you see.

(c) (2pt) For K=3 clustering result, compare it to the true data label. How good is the classification?

(d) (2pt) Comment out the part of your code that reinitialize the centroid if the initial assignment is not good. Run the KMeans algorithm multiple times with K=4, what problem do you see? Comment on how the choice of initial centroids might affect the results and what are the possible solutions.

2. DBSCAN (10 pt)



(a) (6pt) Use DBSCAN to classify compounds dataset. Adjust the Rcut and MinPts hyperparameters so that we have 3 clusters. How many core, border and noise points do you have respectively? Compared to KMeans, is DBSCAN more effective?

(b) (4pt) Let's work on the noisy moon dataset (provided in the reference code) instead. Try using DBSCAN and KMeans with K=2. Run both algorithms with different initial conditions (say 3 times each). Visualize the clustering result. Which method works better?

3. Clustering and simulated annealing (10 pt). Returning to problem (1) using the chemical compound data set, we would like to cluster N data points into K clusters by using simulated annealing (SA) and the cost function:

$$J(N,K) = \sum_{i=1}^{N}\sum_{j=1}^{K} w_{ij} d_{ij}^2$$

$$w_{ij} = \begin{cases} 1 & if\ point\ i\ is\ assigned\ to\ cluster\ j \\ 0 & otherwise \end{cases}, \quad 1 \le i \le N \quad and \quad 1 \le j \le K$$

where d_{ij} is the Euclidean distance between point i and the center of cluster j, and condition on w_{ij} ensures that a point is defined to be in one of the distinct clusters K. Use your code from HW#2

(a) (2pt) This time, normalize your chemical descriptor data for each attribute by subtracting off the mean and dividing by the standard deviation.

(b) (2pt) Given the initial categorization of the 150 organic compounds into the 3 clusters according to *Start assignment* column in the dataset, determine the centroid of each of the three clusters. The centroid for this problem is a 4-D vector where each entry is mean of a variable for the observations in that cluster.

(c) (2 pt) Given the centroid, determine the value of the cost function for this initial categorization. Check against the debugging output.

(d) (4pt) Use your code for SA with a visitation function in which a randomly chosen organic molecule i is moved from its present cluster j to another randomly chosen cluster $k \neq j$. One epoch corresponds to attempting to move all N compounds between clusters, i.e. there are N Metropolis steps, at each temperature. Use a start temperature of 500, and use a geometric cooling schedule ($T_{t+1}=\alpha T_t$) with α=0.999 and total of 5000 steps, again using at least 3 runs of CSA. Check your final temperature against debugging output. Report all 3 solutions and the members of the phenol, ether, and amide as part of each cluster. How good is the assignment compared to problem 1?