MSSE 277B: Machine Learning
Homework assignment #6: Clustering
Assigned Mar. 3 and Due Mar. 14

1. KMeans. (8pt) We will now examine unsupervised learning for classification for the data set of chemical compounds. In the 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) (1pt) 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 (utilize 2-3 seaborn methods as demonstrated in Tutorial 6).

(b) (3pt) 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. (Use the provided code if you are a ugrad or fill in the code for KMeans if you are a grad.)

(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 the 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. (6pt)

(a) (4pt) 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) (2pt) Let's work on the noisy moon dataset (provided in the reference code) instead. Try using DBSCAN and one of KMeans with K=2. Visualize the clustering result. This time which method works better?