CSE 40647/60647: Data Science **Fall 2020**

Homework Programming Assignment 3: Clustering

Handed Out: October 6, 2020 *Due: October 18, 2020 11:55pm*

Save and submit your solution file as *NETID-hw3-programming.zip*. The zip file has *NETID-hw3-programming.pdf* and (saving *hw3.ipynb* as) *NETID-hw3-programming.ipynb*.

K-Means Clustering (60 points)

Please use **Python** to solve the problems. You are NOT allowed to directly call any clustering functions (like k-means functions in Scikit Learn).

Can we group college football teams into clusters by their performances in 2015 and 2017? The table below collects performance data of 12 teams that were ranked at AP Top 25 in Week 14, both years. We have *number of win games* and *ranking* in each season as features. We will use **K-Means Clustering** for **team clustering** in this homework on this data set. Again, we have 12 data objects (i.e., football teams) and 4 numerical features. We may NOT have to use all the features for clustering: actually in this homework, we are often required to use only two of the four features. We **skip** the step of feature normalization.

College	#Wins in 2015	#Wins in 2017	Ranking in 2015	Ranking in 2017
Alabama	12	11	2	4
Clemson	13	12	1	1
LSU	8	9	22	16
Michigan State	12	9	3	18
Northwestern	10	9	8	14
Notre Dame	10	9	8	14
Ohio State	11	11	7	5
Oklahoma	11	12	4	2
Oklahoma State	10	9	13	17
Stanford	11	9	5	15
TCU	10	10	11	13
Wisconsin	9	12	23	6

Q1: Compare Initial Centroids (25 points)

Use Python to do K Means Clustering with two features (1) #Wins in 2015 and (2) #Wins in 2017. Suppose the number of clusters is $K = 2$. Use *Euclidean distance* as the distance metric. Initialize your algorithm with the following centroids:

1. (7,7) and (14,14).
2. (7,7) and (7,14).

Do they generate the same result? Which initialization do you prefer and why? (Hint: For K-Means, there is a measurement to compare the quality of the two clustering results.) For each initialization, please visualize the team clusters using a scatter plot and color the two clusters with RED and BLUE.

Q2: Compare Features (15 points)

Use Python to do K Means Clustering with two features (1) Ranking in 2015 and (2) Ranking in 2017. (Note that we are now using the "ranking" features, not #Wins in Q1.) Suppose the number of clusters is $K = 2$. Use *Manhattan distance* as the distance metric. Initialize your algorithm with the centroids (1,1) and (25,25). Compared with cluster results in Q1, do you prefer the clustering based on these two new features more or less? Please visualize the team clusters with a scatter plot and color the two clusters with RED and BLUE.

Q3: Choose a good K (20 points)

Use Python to do K Means Clustering with two features (1) Ranking in 2015 and (2) Ranking in 2017. Use *Manhattan distance* as the distance metric.

1. Draw the teams as points in a scatter plot. If you are asked to group them into $K = 3$ clusters and color with three different colors RED, BLUE and GREEN, how will you assign the colors to the team data points? Please visualize your coloring in a figure.
2. Find three good initial centroids that can generate your favorite grouping as given above. If you cannot make it, just show the best result that you can do.
3. Compared with results of $K = 2$ in Q2, do you prefer $K = 3$ more or less? and why?