CSE 40647/60647: Data Science **Spring 2018**

Homework 3

Handed Out: March 20, 2018 *Due: April 3, 2018 11:59 pm*

- This assignment is due at **11:59 PM** on the due date. Contact TA if you have technical difficulties in submitting it on **Sakai**. We shall NOT accept any late submission!
- Homework must be submitted in ZIP format (including .pdf, .py and datafile you use). Name your ZIP file as **YourNetid-HW*x*.zip**. Programming result should be written into PDF, and handwritten answers must be scanned into PDF.
  – YourNetid-HW*x*.zip
  —- YourNetid-HW*x*.pdf
  —- YourNetid-HW*x*-Q*y*.py
  —- ... (and any supplementary materials)
  —- Please provide your python version in your PDF file. Regardless of tools you use for your python programming, please submit the code in .py format so that TA can run it via command line.
- Please use **Piazza** if you have any question about the homework.
- Specific Python packages for clustering algorithms (e.g., sklearn) are NOT allowed to be used.

Data set

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** algorithm for **team clustering** in this homework on this data set. Here we skip the step of feature normalization. Good luck!

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

1 Compare Initialized Centroids (30 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?
Please submit your code as **YourNetid-HW3-Q1.py**.
Please visualize team clusters with a scatter plot and color two clusters with RED and BLUE. Attach your plot and write down your answers in the PDF.

2 Compare Features (30 points)

Use Python to do K Means Clustering with two features (1) Ranking in 2015 and (2) Ranking in 2017. 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 Question 1, do you prefer the clustering based on these two new features more or less?
Please submit your code as **YourNetid-HW3-Q2.py**.
Please visualize team clusters with a scatter plot and color two clusters with RED and BLUE. Attach your plot and write down your answers in the PDF.

3 Choose a good K (40 points)

Use Python to do K Means Clustering with two features (1) Ranking in 2015 and (2) Ranking in 2017. Suppose the number of clusters is $\mathbf{K} = \mathbf{3}$. Use *Manhattan distance* as the distance metric.

1. Draw the teams as points in a scatter plot. If I ask you 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 initialized centroids that can generate your favoriate grouping as above. If you cannot make it, just show the best result that you can do in a figure.

3. Compared with results of $K = 2$, do you prefer $K = 3$ more and why?

Please submit your code as **YourNetid-HW3-Q3.py**.
Attach your figures and write down your answers in the PDF.