CSE 40647/60647: Data Science **Fall 2019**

Homework 3: Programming Assignment

Handed Out: October 29, 2019 *Due: November 11, 2019 11:55pm*

Save your homework submission as *NETID-hw3-programming.zip*. The zip file has one pdf file *NETID-hw3-programming.pdf*, multiple code files, the Dataset-clustering.txt file, and one README file.
In the README file, please specify the python version you used and how to run your code in command line.

K-Means Clustering

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

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 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?
For each initialization, please visualize the team clusters using a scatter plot and color the two clusters with RED and BLUE.

Output: Attach the two plots and write down your answer in the pdf. Save your code as NETID-hw3-1.py.

Solution:

If choosing "Wisconsin" as group 1, we will come to following results.

For group 1 centroids, the scatter plot refers to Figure 1. Black triangles are final centroids.

Cluster 1: ("LSU", "Michigan State","Northwestern", "Notre Dame", "Oklahoma State", "Stanford", "TCU", "Wisconsin")

Cluster 2: ("Alabama", "Clemson", "Ohio State", "Oklahoma")

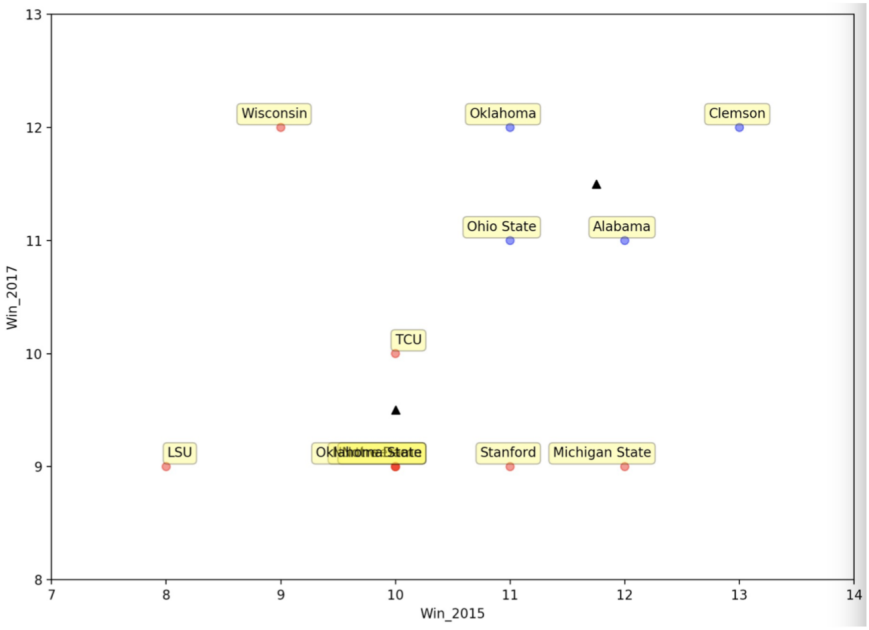

Figure 1: Clustering Scatter plot of 1st group of centroids

For group 2 centroids, the scatter plot refers to Figure 2. Black triangles are final centroids. One of the centroid is hide behind the yellow label.

Cluster 1: ("LSU", "Michigan State", "Northwestern", "Notre Dame", "Oklahoma State", "Stanford", "TCU")

Cluster 2: ("Alabama", "Clemson", "Ohio State", "Oklahoma", "Wisconsin")

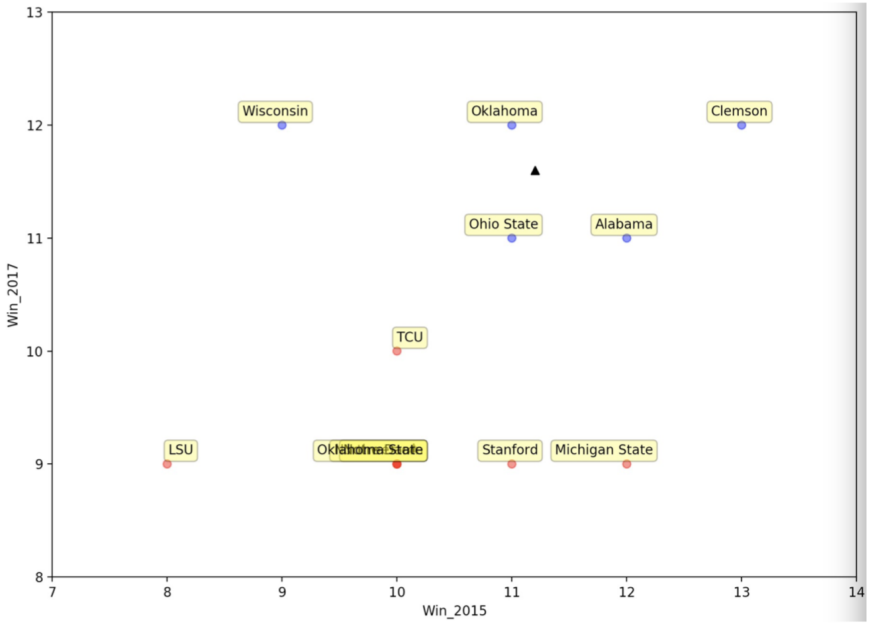

Figure 2: Clustering Scatter plot of 2nd group of centroids

If taking (7,7) and (14,14) initialization, the final centroids are (10.0, 9.5) and (11.75, 11.5). If taking (7,7) and (7,14) initialization, the final centroids are (10.14, 9.14) and (11.2, 11.6). Two groups of centroids does not generate same result because KMeans clustering is centroids sensitive. We prefer the first initialization because the SSE is smaller from group 2. SSE of two groups [21.73, 19.72].

[Code 20 points and results 10 points]

If choosing "Wisconsin" as group 2, we will come to following results.

For group 1 centroids, the scatter plot refers to Figure 1. Black triangles are final centroids.

Cluster 1: ("LSU", "Michigan State","Northwestern", "Notre Dame", "Oklahoma State", "Stanford", "TCU")

Cluster 2: ("Alabama", "Clemson", "Ohio State", "Oklahoma", "Wisconsin")

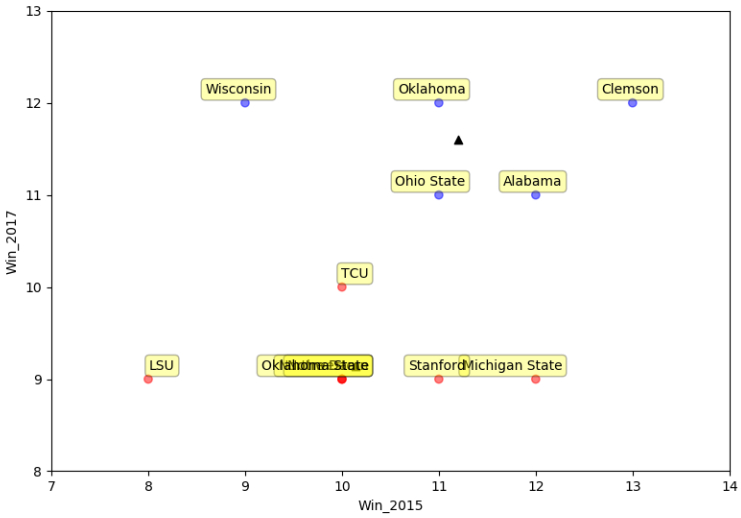

Figure 3: Clustering Scatter plot of 1st group of centroids

For group 2 centroids, the scatter plot refers to Figure 2. Black triangles are final centroids. One of the centroid is hide behind the yellow label.

Cluster 1: ("LSU", "Michigan State", "Northwestern", "Notre Dame", "Oklahoma State", "Stanford", "TCU")

Cluster 2: ("Alabama", "Clemson", "Ohio State", "Oklahoma", "Wisconsin")

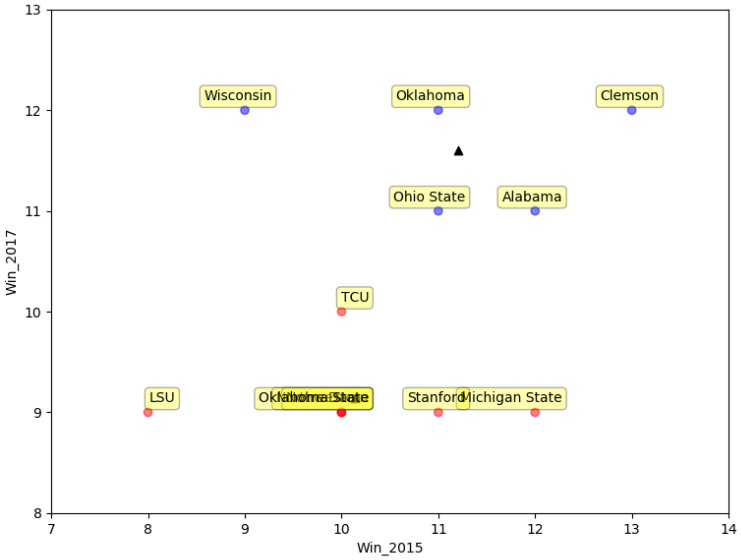

Figure 4: Clustering Scatter plot of 2nd group of centroids

If taking (7,7) and (14,14) initialization, the final centroids are (9.83, 9.16) and (11.33,

11.16). If taking (7,7) and (7,14) initialization, the final centroids are (10.14, 9.14) and (11.2, 11.6). Two groups of centroids are same and SSEs are both 19.72.

If you discuss two above different results, you will get **5 extra credits**.

Python Code was developed in python 3.6, please refer to files "HW3Functions.py" and "DataScience-HW3-Q1.py"

Q2: Compare Features (30 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.

Output: Attach the plot and write down your answer in the pdf. Save your code as NETID-hw3-2.py.

Solution:

The scatter plot refers to Figure 3

Cluster 1: ("Alabama", "Clemson", "Michigan State","Northwestern", "Notre Dame", "Ohio State", "Stanford", "Oklahoma")

Cluster 2: ("LSU", "Oklahoma State", "TCU", "Wisconsin")

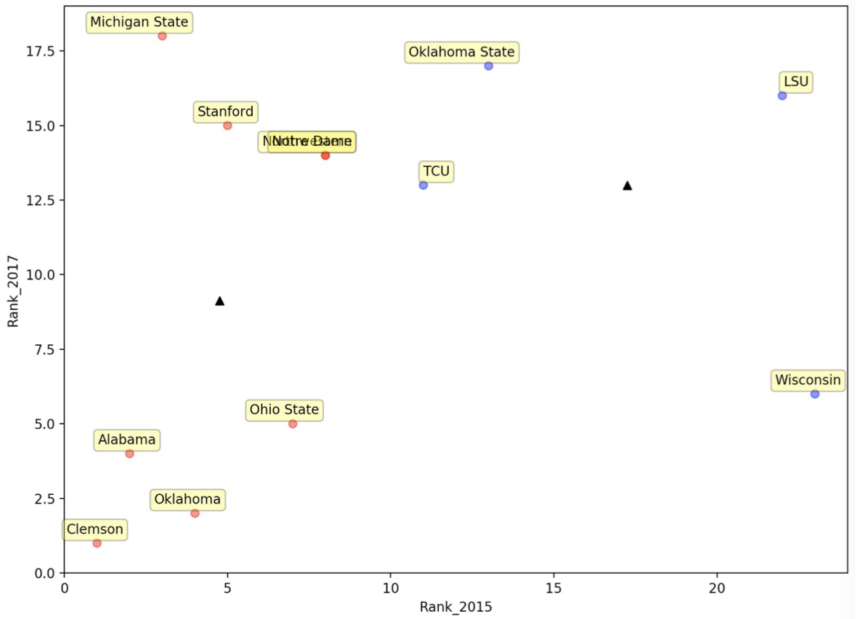

Figure 5: Clustering Scatter plot of Manhattan Distance

final centroids are (4.75, 9.125) and (17.25, 13.0). Comparing to Question 1, we prefer the clustering based on the two new features less because according to the scatter plot, distance between points within same cluster are more disperse.

Python Code was developed in python 3.6, please refer to files “HW3Functions.py” and “DataScience-HW3-Q2.py”

[Code 20 points and results 10 points]

Q3: 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. 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 initialized 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?

Output: Attach the plots and write down your answer in the pdf. Save your code as NETID-hw3-3.py.

1. The scatter plot with three group is shown in Figure 4. [15 points]

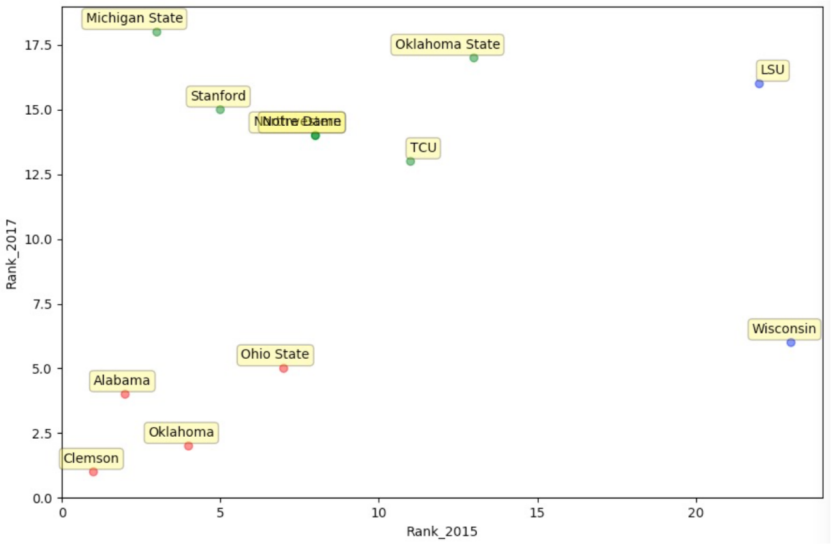

Figure 6: Raw Scatter plot

2. Using initialization centroids: (1, 1), (25, 25), (10, 10). [10 points]

The scatter plot refers to Figure 5. Black triangles are final centroids.

Cluster 1: ("Alabama", "Clemson", "Ohio State", "Oklahoma")
Cluster 2: ("LSU", "Wisconsin")
Cluster 3: ("Michigan State", "Northwestern", "Notre Dame", "Oklahoma State", "Stanford", "TCU")

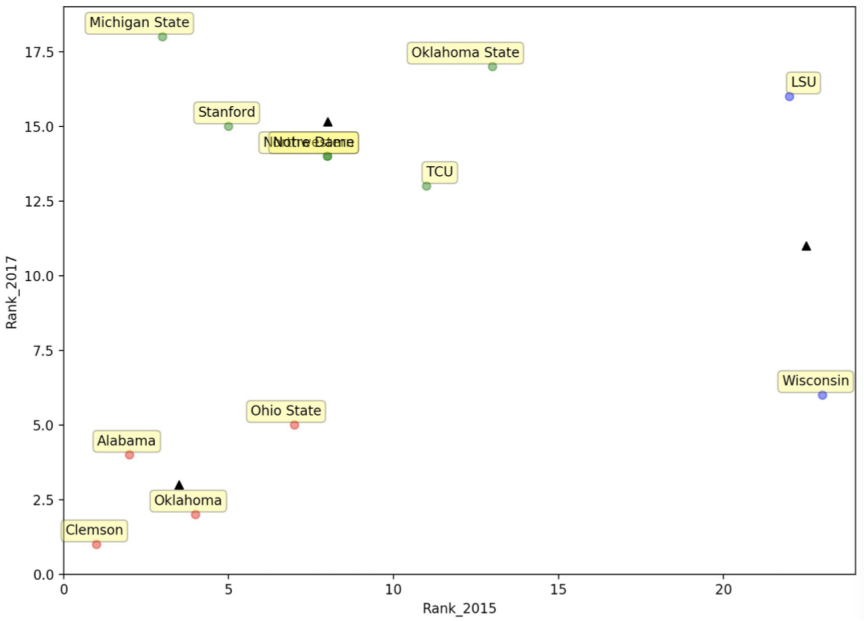

Figure 7: K Means Clustering Scatter plot of Manhattan Distance

3. Comparing with 2 clusters, we prefer the K-means (K=3) because the sse (266.97) from Q3 is much smaller than the sse (917.89) from Q2. Meanwhile, comparing the scatter plots, 3-means clustering looks more reasonable as the data is more disperse of given features. [15 points]

Python Code was developed in python 3.6, please refer to files "HW3Functions.py" and "DataScience-HW3-Q3.py"