*ISTE-470 Assignment 4*

##### Fall 2201 Team Data Miners

Answer the following questions as a team and submit them to the Assignment 4 dropbox by the specified due date. Keep in mind that even though this assignment will be submitted as a team, each team member is individually responsible for understanding the answers to each question and can perform necessary analyses on an exam.

Scoring: Questions 1-5 are worth 5 points each. Question 6 is worth 75 points.

1. Based on the discussion about K-means clustering, what are the four major items that must be established in order to perform K-means clustering?

The four major items that must be established in order to perform K-means clustering are the first step is that the user must choose the number of clusters K. The second step is to select K random instances from the data as centroids. The third step is to assign all the points to the closest cluster centroids. The fourth step is to recompute the centroids of newly formed clusters. After that we will repeat the second, third, and fourth steps until the same points are applied to each cluster in consecutive rounds or until some stopping condition is reached.

2. What is the Initial Centroid Problem and why is it an important consideration when using K-means clustering?

The Initial Centroid problem is that the initial centroids are randomly chosen. This will result in potentially huge variability in the quality of the individual clusters formed by K-means. It is an important consideration when using K-means clustering because we should not run K-means just once. The K-means needs to be run many times because running it only once then the result will not be sufficient and it will be bad clustering. So, the K-means needs to run many times and analyze the result to get something that is decent or good.

3. Explain how SSE is used to compare between two clusterings with the same K value.

The Sum of Squared Error (SSE) is a common tool to observe the distance between data instances and the centroid of clusters. The smallest error results of SSE are preferred because it’s showing the cluster is more cohesive instead of spreading out, which is most likely to be reliable and accurate.

4. Why is a silhouette coefficient of 1 considered “ideal”?

Silhouette coefficient of 1 is considered ideal because if the K value scored very close to 1 means that the clusters are very cohesive and well-separated from each other. If any K value scored very close to 0 then the clusters are overlapping each other with data points that are unlikely to display the differences and commonalities.

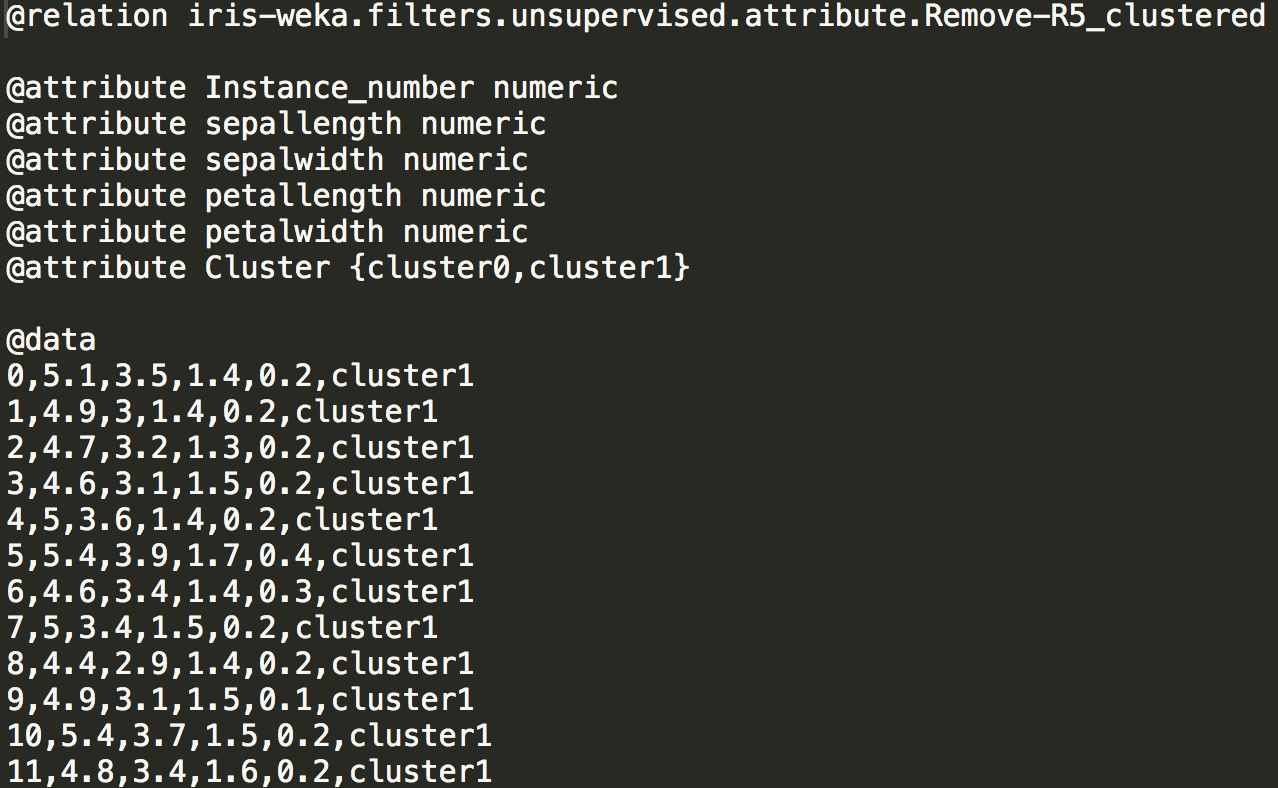
5. Describe how the silhouette coefficient may be used to help choose the K value in K-means clustering.

The silhouette coefficient method is one of the ways to find what number of clusters, K, should be, by evaluating the highest average silhouette coefficient results in both within cluster separation and between cluster separation. They are useful to consider with SSE results to tell the distinction in distance errors and average distances in each K-value.

6. Write a Python script that computes the average silhouette coefficient of unsupervised K-means clusterings for K-values of 2, 3, 5, and 7 using the iris data set and the Weka default random seed value of 10. The average silhouette coefficients should match those shown on Slide 44 of the Clustering lecture. Insert screenshots of your script’s output where indicated below and make sure you submit your code, and anything else needed to run your script, to the dropbox. Note: if you want to write a program in a language other than Python, feel free to do so, but make sure you submit everything needed to run the program to the dropbox.

Before you can write a script, however, you need the K-means clusterings from Weka! First, in the Preprocess tab, click the checkbox next to the class attribute and click the Remove button. Now you won’t have to worry about ignoring the class attribute in the Cluster tab when you run the K-means algorithm.

Once you’ve created a K-means clustering, right click on its entry in the Results List and choose “Visualize cluster assignments”. This will pop up a new window with a visualization of the clustering. Click the Save button in that window and name the ARFF file “clustering\_<K>.arff”, where K is the K-value used for that clustering. The file will look something like the screenshot below. Repeat this procedure for all clusterings.



Notice that the class attribute has been removed. Instead, you have a Cluster attribute that represents all the clusters in the clustering. In this case, this ARFF file represents a clustering for a K of 2. This is the data you’ll need to use when computing the average silhouette coefficient. **Feel free to modify the ARFF files so that they’re more convenient for your script to process**.

Insert screenshots below showing the average silhouette coefficient output for each of the four unsupervised K-means clusterings.

===== 2 K-VALUE =====

ai = [1.3477883705275213, 0.6162956218451565]

bi = [4.008245939731849, 4.183047898841672]

si = [0.6637460897377749, 0.8526682847653226]

Avg Silhouette Coefficients = 0.6808136202936816

===== 3 K-VALUE =====

ai = [1.3729921182873492, 0.5147804629130015, 1.2296490010309247]

bi = [1.448815302115151, 3.525095834563025, 1.320312880226361]

si = [0.052334610020412065, 0.8539669594608867, 0.06866848044373586]

Avg Silhouette Coefficients = 0.5525919445499757

===== 5 K-VALUE =====

ai = [1.0038317896646383, 0.5147804629130015, 0.8173290846305284, 0.6405291180888856, 0.8222356898372347]

bi = [1.3392778794598965, 3.036236799968453, 0.8413819610482947, 1.333580820317834, 1.5460491385328212]

si = [0.2504678789517055, 0.830454441854354, 0.028587345024367106, 0.5196923138589925, 0.46816975648165704]

Avg Silhouette Coefficients = 0.4885175508886279

===== 7 K-VALUE =====

ai = [0.39616028798295094, 0.7559230280884192, 0.6611394656752415, 0.6716240616412162, 0.7932367731061938, 0.8173290846305284, 0.4369985009530614]

bi = [0.8635645850888626, 1.3241131643292656, 1.3983385799922672, 0.8742322483932853, 1.4994744054348819, 0.7996855664911334, 0.6217078329961103]

si = [0.5412499599642739, 0.4291099518889424, 0.5271964350158331, 0.2317555628088922, 0.4709901214511648, 0.022063069384647527, 0.29709989522394276]

Avg Silhouette Coefficients = 0.3583580240469423

Below is a rubric for the grading on this question. This should give you some hints as to how to proceed with your script.

|  |  |
| --- | --- |
| **Grade Item** | **Point Value** |
| Reads the clustering’s data into a data structure | 5 |
| Creates a data structure containing the “names” of all clusters | 5 |
| Calculates the Euclidean distance between any pair of data instances | 10 |
| Calculates the cohesion (ai) for each data instance | 15 |
| Calculates the separation (bi) for each data instance | 30 |
| Calculates the silhouette coefficient (si) for each data instance | 5 |
| Calculates the average silhouette coefficient for the entire clustering | 5 |