**Assignment 4**

By Harsh Sharma

P3. (Clustering-based outlier detection) For this practice, you will use the iris data set. By grouping data into clusters, those data not assigned to any clusters are taken as outliers you will detect outliers with k-mean algorithm.

(1) Conduct k-mean with k=3. In clustering, The clusteringbased detection is an unsupervised method.

(2) Show the centroid of each cluster

(3) Calculate distances between objects and the cluster centroids. Each object will have three distances to three centroids.

(4) Compute the local outlier factor value of each example (instance) as its outlier score. In this computation.

(5) Report the data objects with top 5 largest distances as outliers

(6) Plot the clusters and point their centroids and the five outliers. Because we cannot plot all four attributes, make a 2-dimensional plot with two attributes, PetalLength and PetalWidth.

**Solution:**

The solution is in the R file called ‘P3 Solution.R’. Here are the answers.

(1)

[1] Sepal.Length Sepal.Width Petal.Length Petal.Width Species

<0 rows> (or 0-length row.names

(2)

[1] "Centroids of each cluster:"

> print(centroids)

Sepal.Length Sepal.Width Petal.Length Petal.Width

1 6.850000 3.073684 5.742105 2.071053

2 5.006000 3.428000 1.462000 0.246000

3 5.901613 2.748387 4.393548 1.433871

(3)

[1] "Distances between objects and cluster centroids:"

> print(iris\_with\_distances[1:10, ])

Sepal.Length Sepal.Width Petal.Length Petal.Width Species Distance\_to\_Centroid\_1

1 5.1 3.5 1.4 0.2 setosa 5.328512

2 4.9 3.0 1.4 0.2 setosa 2.973420

3 4.7 3.2 1.3 0.2 setosa 5.109584

4 4.6 3.1 1.5 0.2 setosa 3.938188

5 5.0 3.6 1.4 0.2 setosa 5.328512

6 5.4 3.9 1.7 0.4 setosa 2.973420

7 4.6 3.4 1.4 0.3 setosa 5.109584

8 5.0 3.4 1.5 0.2 setosa 3.938188

9 4.4 2.9 1.4 0.2 setosa 5.328512

10 4.9 3.1 1.5 0.1 setosa 2.973420

Distance\_to\_Centroid\_2 Distance\_to\_Centroid\_3

1 3.938188 5.109584

2 5.328512 3.938188

3 2.973420 5.328512

4 5.109584 2.973420

5 3.938188 5.109584

6 5.328512 3.938188

7 2.973420 5.328512

8 5.109584 2.973420

9 3.938188 5.109584

10 5.328512 3.938188

(4)

[1] "Local Outlier Factor (LOF) values:"

> print(iris\_with\_lof[1:10, ])

Sepal.Length Sepal.Width Petal.Length Petal.Width Species LOF

1 5.1 3.5 1.4 0.2 setosa 1

2 4.9 3.0 1.4 0.2 setosa 1

3 4.7 3.2 1.3 0.2 setosa 1

4 4.6 3.1 1.5 0.2 setosa 1

5 5.0 3.6 1.4 0.2 setosa 1

6 5.4 3.9 1.7 0.4 setosa 1

7 4.6 3.4 1.4 0.3 setosa 1

8 5.0 3.4 1.5 0.2 setosa 1

9 4.4 2.9 1.4 0.2 setosa 1

10 4.9 3.1 1.5 0.1 setosa 1

(5)

> print("Top 5 outliers:")

[1] "Top 5 outliers:"

> print(outliers[1:5, ])

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

42 4.5 2.3 1.3 0.3 setosa

58 4.9 2.4 3.3 1.0 versicolor

61 5.0 2.0 3.5 1.0 versicolor

69 6.2 2.2 4.5 1.5 versicolor

88 6.3 2.3 4.4 1.3 versicolor

(6)

A graph of different colored dots

Description automatically generated with medium confidence