Dalton Murray

Data Science for Business

INT 7623 - 3404-202420\_INT7623\_OL

Dr. Kamel Rushaidat

April 22, 2024

**Assignment #4 – Regression & K-Means**

**Part 1: Regression**

**1. Reposition the points so that the correlation coefficient (r) to 1. What does it mean to have r =1?**

If r = 1 it means that there is a perfect positive correlation.

A graph of a graph of a line

Description automatically generated with medium confidence

This example shows a perfect positive correlation where r = 1.

**2. Reposition the points so that the correlation coefficient (r) to -1. What does it mean to have r =-1?**

If r = -1 there is a perfect negative correlation.

A graph of a function

Description automatically generated

This example shows a perfect negative correlation where r = -1.

**3. Reposition the points so that the correlation coefficient (r) to 0 or very close to zero. What does it mean to have r =0?**

If r = 0 there is no correlation.

A graph on a grid

Description automatically generated

This example shows a correlation of r = 0, where there is no correlation between the points.

**Include screenshots for every part and make a comparison between the three different scenarios in terms of the correlation between the two variables. Discuss your results**

By reviewing all three examples, r = 1, r = -1, and r = 0 we can clearly see what it means for there to be a correlation between the points. With r = 1, it is essentially a straight line of points at an upward angle all along the same path of up 1 right 1, whereas with r = -1 this is the exact opposite, it’s a straight-line descending, all along the same path of down 1 right 1. With r = 0, it is essentially a big mess, the points are seemingly not correlated at all.

**Part 2: K-Means Clustering**

I will be using the packed circles dataset for every question.

A black and white image of small dots

Description automatically generated

**1. Choose the first strategy to initial the centroids by “choosing them by yourself”. Include screen shots for the steps. How many iterations the algorithm did till it finds the best clusters? (5 points)**

My strategy in choosing the centroids were to put the points in what seems like the most central point of each of the clusters in order to make it the easiest for the algorithm.

A colorful hexagon with many small dots

Description automatically generated

After pressing go, this is now what it looks like:

A colorful hexagon with many small dots

Description automatically generated

IA group of colorful circles

Description automatically generated with medium confidence

I then press update centroids and reassign points 3 times, on the final time the points no longer move, so by this I assume it takes 2 iterations for it to find the best clusters.

A colorful hexagons with dots

Description automatically generated

**2. Choose the second strategy to randomly choose the centroids. How many iterations the algorithm did till it finds the best clusters? (5 points)**

Choosing the same dataset as before but choosing the random strategy, I added 7 random centroids:

A colorful background with many small black dots

Description automatically generated with medium confidence

I then pressed go and it looked like this:

A colorful background with many dots

Description automatically generated with medium confidence

I then pressed update centroids and reassign points 13 times, on the 13th time it did not change from the 12th time, so it took 12 iterations.

A group of colorful circles

Description automatically generated with medium confidence

**3. Choose the third strategy by using the Farthest point as the centroids. How many iterations the algorithm did till it finds the best clusters? (5 points)**

Adding 7 centroids to the furthest points looks like this:

A colorful hexagon with black dots

Description automatically generated

I then pressed go and it looks like this:

A colorful hexagon with many dots

Description automatically generated

I then pressed update centroids and reassign points 4 times, on the fourth time there was no movement so it found the best clusters on the third iteration.

A colorful hexagons with dots

Description automatically generated

**Discuss your conclusion about using the three different strategies. Add any interesting facts/notes that you found when tried this visualization. (6 points)**

By going off of how many iterations it took, the manual method was the fastest, the second fastest was the furthest points method, and then the random method. I believe that using the furthest points method is the most efficient way to do it in a purely computational way. However, when we start adding more or less points than 7 it may be more or less difficult for a computer to calculate the best clusters, as it cannot be as easily divided. As a human I am able to view and identify easily the 7 clusters without a visualization, however, a computer can have difficulties with this, and this is where clustering algorithms come into play. In the long run, they could be able to do this faster than humans and more effectively, especially when breaking up the clusters with a different amount of centroids than there are obvious clusters.

I have neither given nor received unauthorized aid in completing this work, nor have I presented someone else's work as my own.

*Dalton Murray*