Julia Fairbank & Caroline Cutter

CS 311 AI

May 15, 2022

**New Final Project Proposal**

1. **The name of students in your group**

Julia Fairbank and Caroline Cutter

1. **Title of Project**

Improving the Lloyds K-Means Clustering Algorithms with different initialization strategies and the Hartigan-Wong clustering algorithm

1. **General Research Area**

Clustering Algorithms, Unsupervised Learning

1. **Summary of Related Literature & Background**

The standard K-Means Clustering algorithm is called Lloyds Algorithm, which is an easy to understand and to implement algorithm. This algorithm works by initially choosing k data points as centers. Then, each data point is assigned to the closest center, and the center of the mean is calculated again using the points assigned, then the data points are reassigned to the center. This continues until the data points no longer change centers they are assigned to. Although this algorithm is relatively effective and easy to implement, it is not guaranteed to find the optimal solution and the outcome is heavily based on the way the centers are initialized. So, for our project we wanted to research ways that this could be improved. What we found is that there is a type of optimization for initialization called kmeans++ initialization (compared to random initialization of points). This initialization chooses initial centers that are as far apart from each other as possible, which has shown to improve the ending results. We want to compare this improved Lloyds algorithm with the Hartigan-Wong k-means algorithm to see which one is most accurate, and the difference in time complexity. The Hartigan-Wong method initially assigns datapoints to random centers, which then later are calculated as the mean of their assigned data points. This algorithm has been show to reduce running into local minima than Lloyds algorithm.

**References:**

* + <https://towardsdatascience.com/three-versions-of-k-means-cf939b65f4ea>
  + <https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e>
  + <https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html#:~:text=k%2Dmeans%2B%2B%3A%20As%20spreading,probability%20proportional%20to%20the%20squared>
  + <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

1. **Specific research problem you will solve**

We are looking to try different initialization methods on Lloyds clustering algorithm as well as implementing Mean-Shift to see which gives the highest accuracy of classification.

1. **A description of the det of methods you plan to use to try to solve your problem**

First we will implement using (using Java) Lloyds K-Means Algorithm, Lloyds K-Means Algorithm with Kmeans++ initialization, and Mean-Shift Clustering. Then, we will train each of the algorithms with training data, and then run test data on it to see how accurately the algorithms classify the examples. We hope to see improved accuracy with the new methods, but want to focus on the comparison between the two improved methods, using Lloyds as a baseline. We would also like to test this data using different K-Values.

1. **A description of how you will evaluate the results of the methods**

To analyze the data, we would like to create graphs of our clusters using colors to differentiate the clusters. We would want to compare the graph that our algorithm computes with the graph of the actual classified data. We also want to calculate the accuracy of assigned points, as well the time complexity of the algorithms.