Miles A. Peña

DSC 500 Introduction to Data Science

Matthew Metzger

10/03/2023

K-Means Clustering

A scenario in which we could use k-means clustering would be determining how to best balance study time with fun time. Some categories for measurable data would be week number in the course, time spent studying in the week, recreation time in the week, and grade points received for the week. For this example, it would be beneficial to find out what the “sweet spot” would be when it comes to how much time one needs to study and how much time one can spend having fun and doing other things in order to receive the best possible grade in a class. Utilizing the k-means clustering method, will allow us to classify data into distinct groups which guarantees that the data points in each specific group will be similar to one another.

An exploratory view using a pair-plot to get a better idea of the correlation between different features will help. First it is important to define the number of clusters. The number of clusters represents the “k” in k-means. To find the best value for K, we use the Elbow Method and look for the cluster number that reduces the summed distance only slightly; this will be the point used as the number of clusters. After this number has been set, we need to define the initial cluster centroids which are usually defined randomly. In the following step, we should assign points to the clusters. Since we have three clusters, we should have three assigned points. One by one, measure the distance from the selected point to each of the cluster centroids. Each of the points will be assigned to the cluster that is closest to it in measurement. We now can calculate the center of each cluster which will become the new centroids of the clusters (this provides a more centralized point in the cluster to act as the cluster centroid). We must again assign points to the new clusters since the centroids have shifted. The last two steps mentioned here will need to be repeated until the cluster distribution no longer changes. This will be where the analysis begins to find trends and correlations to help solve the question at hand. With the problem being finding a more efficient way to balance study time with fun time while still receiving the best possible score, the data indicating the highest grade with the most balanced times would be ideal to cluster.

*References:*

Bento, C. (2018, June 10). K-Means in Real Life: Clustering Workout Sessions. *Towards Data Science*. <https://towardsdatascience.com/k-means-in-real-life-clustering-workout-sessions-119946f9e8dd>

Gadde, M. (2021, March 1). Simple Explanation To Understand K Means Clustering. *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2021/02/simple-explanation-to-understand-k-means-clustering/#:~:text=Definition%3A%20It%20groups%20the%20data,belong%20to%20the%20same%20cluster.>