**Introduction to Data Science**

**Exercise 6**

**Milan Sherman**

K-means clustering is an algorithm for performing unsupervised data mining by finding groups of observations that are similar in some regard. The process starts with grouping observations randomly (unless otherwise specified) and computing the centroid, or mean, of each of those groups (or clusters). Once each cluster has a centroid defined, the observations that are closest to the centroid define the new clusters, and centroids are recalculated for these clusters. The process iterates until the centroids and the clusters around the centroids no longer change. As the results of the algorithm can vary depending on the initial clusters, the entire process is repeated multiple times to find an optimal solution. The “k” in k-means refers to the number of clusters, and is an important decision that a data scientist must make when using this method. The Elbow Method allows a data scientist to compare the relative performance of various values of k. By plotting the cumulative distance of all the data points to their respective centroids, also known as distortion, versus the number of clusters, or k, one can visually inspect the point at which the improvement in reducing distortion tapers off to a point of diminishing returns. Visually, this is the elbow of the plot.

A scenario in my own job where I might use k-means clustering is to find customers who are good candidates to use our ecommerce platform. By using k-means clustering to group our customers based on various features, we can explore whether any of these clusters has a high proportion of our current ecommerce customers, and target those customers in those clusters who are not currently ecommerce customers with marketing for our ecommerce platform. Of course, we would want to leave out whether a customer has used the ecommerce platform from our list of features so that we don’t end up with clusters with only current ecommerce customers, thus leaving us with no one to target.

K-means is needed to help with this problem because it can help to identify customers who are similar along a large variety and combination of attributes. While I would have to define the features that are included for each customer, it seems that including a large number of features would be desirable in order to uncover unexpected similarities. Finding similar customers with so many features would be difficult or impossible to do without k-means.

The Elbow Method, as noted above, would help me to determine the optimal number of clusters. Our company has millions of customers, and thus the potential for thousands of clusters. The Elbow Method is an efficient way to choose among those thousands of options. In addition, the visual nature of the Elbow Method allows for multiple candidates for the ideal k, which would be useful for such a large dataset.

References

Bento, C. (2018, December 03). K-Means in real Life: Clustering workout sessions. Retrieved April 25, 2021, from https://towardsdatascience.com/k-means-in-real-life-clustering-workout-sessions-119946f9e8dd