6.2 Exercise – K-Means in Real Life  
Thip Rattanavilay, MS Data Science

Bellevue University

**Abstract**

Clustering is traditionally viewed as an unsupervised method for data analysis. However, in some cases information about the problem domain is available in addition to the data instances themselves. In this paper, we demonstrate how the popular k-means clustering algorithm can be profitably modified to make use of this information (Brodley & Danyluk, 2001).

**6.2 Exercise – K-Means in Real Life**

Say you have a lot of data points (let's say heights and weights for a group of people for an easy example) and you want to separate them into groups (in this case, men and women). K-means clustering is a *classification algorithm* used to automatically divide a large group into smaller groups.

The name comes because you choose K groups (ie, K=3 or K=4 or whatever. For our people example, K = 2). You take the average of these groups to improve the accuracy of the group (average = mean, and you do this several times, so means). The cluster is just another name for a group, but it's because different groups are said to 'cluster' together.

Let do an example. Let's say we have 10 data points, which in actuality are 5 men and 5 women, (but we don't know this) and we want to divide them into 2 groups (ie. divide by gender). For this example let's assume that all the men are taller and heavier than all the women. We select 2 random data points as a starting position. Then, we compare these points to all the other points, and find out which starting position is closest. This is our first pass at clustering (and incidentally, this is the slow part).

sum2 = 0; for (k=1; k<=n; k\*=2) for (j=1; j<=k; j++) sum2++;

(Dabbura, 2020)

We have our initial groups, but because we chose randomly, we are probably a little off. Say we got 4 men and 1 woman in one group, and 1 man and 4 women in the other. So, we take the average of all the points in one group to use as a new starting point for that group and do the same for the other group. Then we do the clustering again to get new groups.

Success! Because the average is pulled closer to the majority of each cluster, on the second go around we get all 5 men in one group and all 5 women in the other. How do we know we're done? We do the average and grouping again and see if any points changed groups. None did, so we're finished. Otherwise, we'd go again.

There are some obvious problems with k-means though. The biggest one is if you have more or fewer clusters than you tested for. For example, if our earlier group was men, women, and children, we should have used 3 clusters. With only two, the groups will bounce around a lot and the women will probably get split between the 'man' group and the 'children' group. On the other hand, if we thought we'd be clever and test for 4 groups with the man/woman example, when there's really only 2 clusters, we'd probably split both clusters into subclusters that make no sense - or worse, have some 'in between' cluster that straddles the actual groups. The usual fix for these issues is to run k-means several times above and below the expected number of groups and see how tightly the resulting clusters are created. If you get better clusters with a different number, good indication there's that number of clusters.

The other problem is you can have a set where, due to the nature of the point, you have some edge cases that keep bouncing back and forth between clusters, so the algorithm never ends. The usual fix to this is to either limits the number of clustering performing (and just say, if we're not done by clustering X we're just going with that), or making the threshold for stopping higher than zero (like, 10% of the datapoints or less.

**Conclusion**

Kmeans clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of kmeans is to group data points into distinct non-overlapping subgroups. It does a very good job when the clusters have a kind of spherical shapes. However, it suffers as the geometric shapes of clusters deviates from spherical shapes. Moreover, it also doesn’t learn the number of clusters from the data and requires it to be pre-defined (Dabbura, 2020).

**Reference:**

Brodley, C. E., & Danyluk, A. P. (2001). *Machine learning: Proceedings of the Eighteenth International Conference (ICML 2001): Williams College, June 28-July 1, 2001*. San Francisco (California): Morgan Kaufmann.

(Brodley & Danyluk, 2001)

Dabbura, I. (2020, August 10). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Retrieved October 11, 2020, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

(Dabbura, 2020)

Bento, C. (2018, December 03). K-Means in Real Life: Clustering Workout Sessions. Retrieved October 11, 2020, from https://towardsdatascience.com/k-means-in-real-life-clustering-workout-sessions-119946f9e8dd

(Bento, 2018)