Clustering the Planet

CS451 HW5 Fall 2017

Henry Swaffield and Hans Goudey

The goal of this project was to apply the k-means algorithm, an unsupervised clustering algorithm, in an attempt to group countries by their cultural priorities, using UN survey data. For the first phase of the project, we implemented the k-means algorithm in Python and made preliminary visualizations using matplotlib. We also experimented by looping through various k-values. We then took our data-processing to Excel, where we produced the visualizations that are seen below. The first question considered was to determine and “ideal” number of clusters, and following that is a description of the clusters resulting from that k value.

This is a plot of the cost of the kmeans algorithm cost as a function of the number of clusters. There isn’t a clear elbow where the cost starts decreasing less, and there is a fair amount of noise in the data, because the algorithm didn’t necessarily reach the global minimum for that cluster number.

Because of this noise we decided to run the algorithm multiple times. This allows us to take away some of the per-run variation by just keeping track of the minimum over all of the runs. For smaller centroid numbers, this means that we have likely found the global minimum from the algorithm, but for larger centroid numbers it becomes increasingly less likely that we have found the global minimum. By using the minimum over 4 runs, we have effectively used 400 iterations of the kmeans algorithm.

There’s still no precise elbow, so the choice of cluster number remains somewhat arbitrary. It looks like the rate of decrease of cost lowers at around 4 to 6 clusters. We decided to use 6 clusters so that each cluster was smaller and hopefully more meaningful.

The goal of the kmeans algorithm changes as you increase the number of clusters. The question changes from “How can these countries be categorized?” to “What are a few countries similar to this country?” as the number of clusters increases from 2 to 97.

At first, we plotted the priorities of the clusters without a reference to the average. This plot shows that in general, humans across the globe tend to prioritize categories similarly. For example, all clusters tend to prioritize education over phone and internet access.

However, there is still a lot of variation when considering particular categories. To show this, we instead plotted the difference between each cluster’s opinions on the categories and the average opinion on that category. Each data point measures the difference between that cluster’s opinion about the category and the average opinion from all countries. Here we see that the opinions of the different clusters change a lot in certain categories and less in others.

We can see the abundance of North African, South Asian, and Middle Eastern countries in cluster 3, and that most rich western countries have been grouped in a separate category (cluster 6). The other categories are harder to generalize, but the algorithm has found some similarities. Category 4 primarily consists of sub-Saharan African countries.



We chose to look more closely at Afghanistan because it was part of cluster 3 which was one of the clusters that was a bit harder to explain than the others. The third cluster was wildly different in many of the categories compared to the other clusters, particularly cluster 1. Even the way Afghanistan compares to cluster 3 is not quite obvious. The prioritization for some of the categories is very similar to the cluster, but for other clusters it is totally different. Cluster 3 contains mostly Middle Eastern, North African, and South Asian countries, which seem to share similar latitudes, spanning great distances from east to west. These countries tend to be Islamic as well, which would suggest political and cultural similarities.

Some of the categories in which Afghanistan differs more from the views of its cluster could represent priorities with more variation and less correlation with other priorities. Because of this it would be harder for the algorithm to group countries based on the more variant priorities. Following from the pigeon-hole principle and the fact that we have more priorities considered than categories, it’s expected that certain categories would have more selective weight. In this example, perhaps Afghanistan was grouped with countries that agreed more so on the priority of trustworthy government and political freedoms than sanitation and internet access.

It is somewhat odd that Jamaica is also grouped in cluster 3. It’s extremely far away, and is neither Islamic nor Asian. However, the point of using an algorithm to group these countries in clusters is to find similarities that we wouldn’t necessarily have looked for or seen ourselves. Using an algorithm can eliminate our own bias and show that maybe Afghanistan and Jamaica aren’t as different as we thought.