Winners and Losers: Examining NSF Grant Funding Trends at the Congressional District Level

Machine Learning Methods

By Erik Lopez Jr

The federal government is a major player in funding scientific research throughout the country. However, with shifts in government budgets, the distribution and amount of that funding changes year to year and could very well be affecting its various recipients unevenly. For this project we want to identify those federal funding trends at the congressional district level and identify similarly affected districts to determine if they share any common political or demographic characteristics. We will focus only on research funding distributed by the National Science Foundation (NSF) from 2010 to 2018 to avoid noise from other government departments' own research funding changes.

But how do we cluster together districts with similar funding patterns when we do not know what these patterns are? To solve this problem, we will use both KMeans and Gaussian Mixture clustering, which are unsupervised machine learning clustering algorithms that can optimally separate entities into a pre-specified number of clusters given a set of features from their data. The features for both algorithms will be the total funding that a congressional district received for a specific year between 2010 to 2018. The output of both clustering algorithms will be a label corresponding to the cluster that each congressional district belongs to.

The reason we are using both KMeans and Gaussian Mixture initially is because both algorithms have different biases for the shape of clusters they prefer. KMeans tends to prefer circular-shaped clusters, whereas Gaussian Mixture prefers axial-aligned elliptical clusters. Choosing the wrong algorithm for the shape of your clusters could lead to subpar clustering results, making our proceeding analysis less meaningful. Moreover, to improve the accuracy of our ML models we will perform hyperparameter tuning to find the best number of clusters for each algorithm. By using the Silhouette Score algorithm, we can calculate a measure of how well the algorithm performed to separate each data point from points in other clusters. We will tune both KMeans and Gaussian Mixture, testing cluster numbers from 2 to 10, keeping the model of the algorithm and cluster number that yields the highest Silhouette score for our demographic and political analysis.

Finally, since we are using unsupervised machine learning algorithms, we don't have a good measure of precision or recall because we do not know what the ground truth is. If a cluster with an overall increasing funding trend has districts with erratic funding histories, we cannot be sure that those districts truly belong there. To get around this, we can use the silhouette

score as a measure of how good our clustering method is, depending on how separable the underlying data is. We do not expect our data to be fully separable, what is ultimately important is that the ending clusters do show some distinct trends in funding throughout the past years so we can have defined groupings to answer our research question.