Unsupervised Machine Learning

K-means Clustering with Python and ArcGIS

Data Set Description:

* Data represents Power Plants across the U.S.
* Pulled from ArcGIS Living Atlas Federal User Community
  + Regulated by federal departments and available for public consumption.
* The data is broken down by Geographic location, fuel source, Total installed/max Cap, & individual fuel source capacity by site.
* Contains 12,008 sites.

Preprocessing Techniques:

* Categorical to numerical (Label Encoding)
* Selection of relevant attributes
  + Using Install MW, Latitude, Longitude

Machine Learning Technique: K-means Clustering:

* Performed the elbow method to determine optimum values of k
  + Used values of 3, 5 & 7 to fit the data into visually correct values.

Results:

* Developed two different analyses:
  + Python
    - Generated a 3D plot using Install MW, Longitude, Latitude.
    - Utilized 3 different k values 3, 5, & 7
  + ArcGIS Pro tool called Multivariate Clustering.
    - The tool finds natural clusters of features based solely on feature attribute values.
    - The tool utilizes a process called Optimized seed locations, randomly selects the first seed, and makes sure that the subsequent seeds selected represent features that are far away from each other in data space (attribute values)
    - The number of clusters can either be determined by the tool itself or input manually.

Cluster Analysis & conclusions:

* We learned that different types of preprocessing can be utilized to perform a variety of further analysis especially depending on the type and variety of data that is collected.
  + Using power plants gives a wide range of initial attributes.
* Once K values are determined even if using the optimum value (i.e. 3) doesn’t translate to accurate clusters. So we played around with different values until we got a proper fit. The best fit from my perspective came with the value of k=7.