Conrad Woidyla
CS 521 – Data Mining

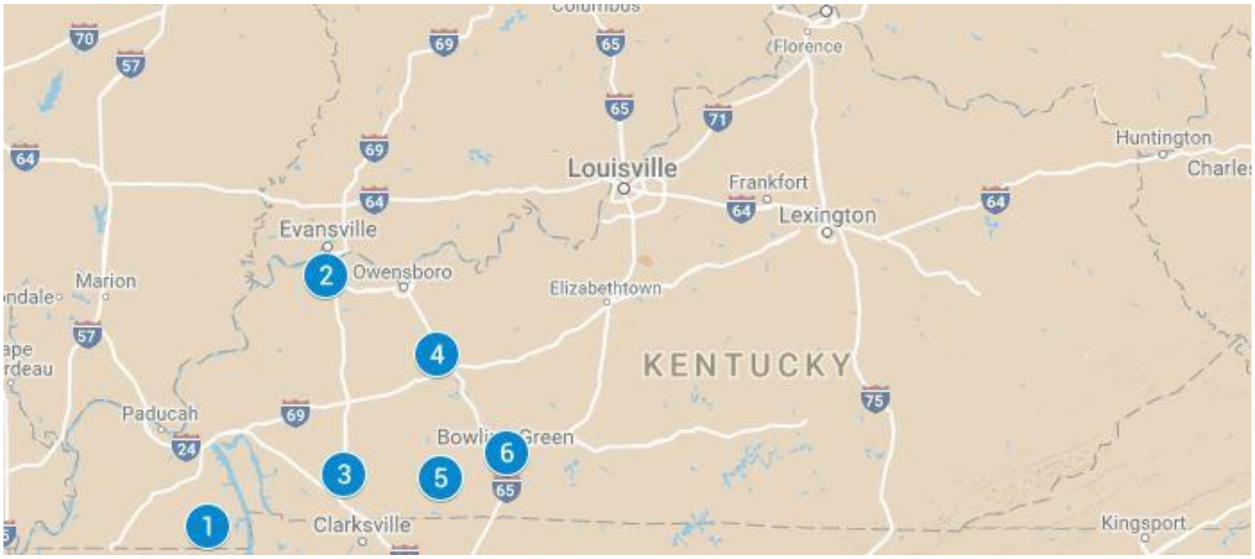
# Clustering Weather Data to Characterize Multiple Geographic Locations

This paper explores the possibility of clustering weather data to characterize a geographic location. While it's obvious that the weather in southern California will always be warmer than northern Alaska, weather differences are less obvious when geographic locations are within the same state only a few hundred miles of each other. The KY Mesonet data set will be used to determine if geographic locations can be distinguished using clustering methods. The goal is to identify six clusters from three months of data that match the labeled geographic locations. Two types of clustering algorithms are implemented: k-means and Gaussian mixture model (GMM). The normalized mutual information (NMI) scores are ~0.0075 and ~ 0.023, respectively. Due to the low NMI scores, it is likely clustering KY Mesonet's raw data is not the right approach for characterizing geographic locations.

## Initial Observations about the Data

The data is ordered and discrete. All of the data is captured every five minutes starting from March 1, 2016 to the end of May 31, 2016. There are six different sets of data from KY Mesonet. The six different sets of data are from different locations in Kentucky and are:

1. MRRY – Murray, KY
2. GRHM – Henderson, KY
3. PGHL – Hopkinsville, KY
4. HTFD – Hartford, KY
5. RSVL – Russellville, KY
6. FARM – Bowling Green, KY



Each location is described by 11 attributes. The attributes and their properties are:

| Attribute | Abbreviation | Unit | Attribute Type |
|---|---|---|---|
| Station ID | STID | Degrees | Nominal |
| Time | UTME | Time | Numeric - interval |
| Air temperature | TAIR | Celsius | Numeric - interval |
| Relative humidity | RELH | Percentage | Numeric - ratio |

| | | | |
|---|---|---|---|
| Dewpoint | TDPT | Celsius | Numeric - interval |
| Wind speed | WSPD | m/s | Numeric - ratio |
| Wind speed max gust | WSMX | m/s | Numeric - ratio |
| Wind direction | WDIR | degrees | Numeric - ratio |
| WD max | WDMX | degrees | Numeric - ratio |
| Precipitation | PRCP | mm | Numeric - ratio |
| Solar radiation | SRAD | $W/m^2$ | Numeric - ratio |

Since each location has roughly the same amount of data, the base rate for each location is about 16.67%.

### Formulating the Clustering Problem

Clustering is unsupervised, so there is not a target variable. Instead, the data is grouped according to similarity based on distances between the data points. While the data originates from different weather stations at different geographic locations in Kentucky, the regional location is similar. Since the weather data originates from a similar region, the weather data should also have a lot of similarities. If the weather data is similar, then clustering algorithms will have a difficult time distinguishing these geographic locations based on weather data. To explore this theory, the clustering algorithms k-means and GMM are used.

### Data Preprocessing

KYMesonet data is preprocessed in three ways: first, the missing values of an attribute are imputed using the attribute's mean; second, the time stamp is split into individual attributes which consist of month, day, year, hour, minute, and second; and third, the data is scaled by finding the mean and standard deviation of the entire column vector, then standardizing each data point. The range of the z-scores for each column is as follows:

| Attributes | Min Z-Score | Max Z-Score |
|---|---|---|
| Month | -1.218226 | 1.218073 |
| Day | -1.675693 | 1.712497 |
| Hour | -1.661289 | 1.661364 |
| Minute | -1.593221 | 1.593252 |
| Second | -0.9999528 | 1.0000409 |
| TAIR | -2.881572 | 2.426686 |
| RELH | -2.483487 | 1.459472 |
| TDPT | -3.287754 | 2.383150 |
| WSPD | -1.356290 | 6.970738 |
| WDIR | -1.639387 | 1.775799 |
| WSMX | -1.455247 | 7.123193 |
| WDMX | -1.846467 | 1.767372 |
| SRAD | -0.6853444 | 4.0106989 |
| PRCP | -0.1079447 | 119.0781485 |

It appears that wind speed (WSPD), wind speed max (WSMX), solar radiation (SRAD), and precipitation (PRCP) have very anomalous data points.

## Methods

A mixture of the R programming language and Python was used. R was only used for data preprocessing because R's clustering libraries were outdated, poorly documented, or did not exist. As an alternative, Python was used to perform clustering and evaluation because Python has many more up-to-date, well documented scientific computing libraries. The data does not need to be split into training and test sets since these clusters will be evaluated in two ways: the first is an extrinsic evaluation method that compares clusters against a ground truth; the second is the NMI score. The extrinsic evaluation method requires clusters to form that are similar to the class labels. Once those clusters form, calculating the F-measure is possible. The NMI score determines the dependence between the class labels' ground truth and the clustering labels. A low NMI score indicates low dependence while a high NMI score indicates high dependence.

## K-Means

K-means yielded an NMI score of approximately 0.0075. With the parameters mentioned below, k-means took approximately 16 seconds to complete. The k-means algorithm utilized originates from the sklearn Python library. The inputs for k-means were as follows: number of clusters = 6, maximum number of iterations per run = 300, number of times k-means algorithm is run with a different centroid seed = 10, and the elkan algorithm variation was used. The distribution of data points across different clusters is as follows:

| Cluster | Number of data points | Percentage of total data points |
|---|---|---|
| 0 | 22674 | 14.26 |
| 1 | 37157 | 23.37 |
| 2 | 25434 | 16 |
| 3 | 25532 | 16.06 |
| 4 | 21875 | 13.76 |
| 5 | 26289 | 16.54 |

While the distribution of the data points across different clusters is close to the base rate of 16.67% for each geographic location, further analysis shows that the clusters do not belong to an individual geographic location.

| | FARM | GRHM | HTFD | MRRY | PGHL | RSVL |
|---|---|---|---|---|---|---|
| 0 | 4669 | 4658 | 3902 | 2627 | 4894 | 1924 |
| 1 | 6156 | 5783 | 5718 | 6811 | 5115 | 7574 |
| 2 | 3611 | 3956 | 4213 | 4041 | 4974 | 4639 |
| 3 | 3779 | 4075 | 3762 | 4722 | 4425 | 4769 |
| 4 | 4055 | 4124 | 4241 | 3444 | 3462 | 2549 |
| 5 | 4214 | 3899 | 4660 | 4850 | 3625 | 5041 |

The distribution confirms that the data is too similar for a partitioning algorithm to find differences related to geographic location. Since the data points are so evenly distributed, it is possible that the weather data is organized as a high dimensional, layered sphere or ellipsoid, where the data residing in the core pertains to all of the geographical locations, but the data residing in the outer layers distinguish different geographical locations. If this is the case, describing the data points' cluster membership as probabilities could lead to a higher accuracy, which is a task best suited for the GMM method.

**Gaussian Mixture Model**

The GMM yielded an NMI score of approximately 0.023 and took about 25 seconds to compute. This score is about three times better than the k-means NMI score. GMM originates from the sklearn Python library. The algorithm uses multiple Gaussian distributions to characterize clusters and assigns probabilities to data points with respect to a cluster's mean and standard deviation. The inputs to GMM are as follows: number of clusters = 6, covariance type is 'full' meaning each component has its own covariance matrix, the conversion threshold for EM iterations = 1e-3, the max EM iterations = 100, number of initializations = 1, and the method used to initialize weights is 'kmeans'. The distribution of data points across the different clusters is as follows:

| Cluster | Number of data points | Percentage of total data points |
|---|---|---|
| 0 | 4419 | 2.78 |
| 1 | 12936 | 8.14 |
| 2 | 52640 | 33.12 |
| 3 | 68957 | 43.38 |
| 4 | 10520 | 6.62 |
| 5 | 9489 | 5.97 |

It is apparent that a majority of the data points fall into clusters 2 and 3. This is likely due to the weather data being too similar. In addition, the clusters do not characterize geographical locations as show in the table below.

| | FARM | GRHM | HTFD | MRRY | PGHL | RSVL |
|---|---|---|---|---|---|---|
| 0 | 871 | 853 | 882 | 480 | 826 | 507 |
| 1 | 2661 | 1547 | 2298 | 2026 | 1157 | 3247 |
| 2 | 8780 | 9787 | 9019 | 7443 | 10026 | 7585 |
| 3 | 12426 | 12455 | 12437 | 8646 | 12261 | 10732 |
| 4 | 813 | 792 | 811 | 4602 | 986 | 2516 |
| 5 | 933 | 1061 | 1049 | 3298 | 1239 | 1909 |

Further analysis shows that FARM, GRHM, HTFD, and PGHL have the most similar distribution of data across different clusters, where the majority of their data points are in clusters 1, 2, and 3. MRRY differs in that it has more data in clusters 4 and 5 than any other location. RSVL is in between the data characteristics for FARM, GRHM, HTFD, and PGHL and the data characteristics for MRRY. RSVL has more data in cluster 1 than any other location, and has higher than average data in clusters 4 and 5. As indicated by the map, MRRY is farthest from the other geographic locations, which could explain the difference in its clustering characteristics.

**Results**

Using clustering to determine geographic locations of weather data in a similar region is very unreliable. The resulting clusters from k-means and GMM indicate the data does not cluster based on geographic location. This is also indicated by the very low NMI scores from both clustering algorithms. That being said, the GMM algorithm performs much better than k-means, yielding an NMI score of approximately 0.023, which is about three times better than k-means NMI score of approximately 0.0075.

Since the clusters generated by k-means and GMM do no correlate to geographic locations, computing an F-measure is not possible. Computing the F-measure requires knowing which cluster belongs to which geographic location so that the number of true positives, false positives, true negatives, and false negatives can be calculated. These cannot be calculated reliably given the low NMI score between the clusters and geographic locations.

While the clusters generated by k-means and GMM do not correlate with geographic locations, the clusters generated by GMM could characterize each location based on the distribution of data points in each cluster. This may indicate that a data point is more likely to belong to a location if it falls into a certain cluster. The probability table below is generated from the GMM clusters and illustrates the percentage of data points in a cluster that belong to a certain location.

| | FARM | GRHM | HTFD | MRRY | PGHL | RSVL |
|---|---|---|---|---|---|---|
| 0 | 0.19710342 | 0.19303 | 0.199593 | 0.108622 | 0.18692 | 0.114732 |
| 1 | 0.205705 | 0.119589 | 0.177644 | 0.156617 | 0.08944 | 0.251005 |
| 2 | 0.166793 | 0.185923 | 0.171334 | 0.141394 | 0.190464 | 0.144092 |
| 3 | 0.180199 | 0.18062 | 0.180359 | 0.125382 | 0.177806 | 0.155633 |
| 4 | 0.077281 | 0.075285 | 0.077091 | 0.437452 | 0.093726 | 0.239163 |
| 5 | 0.098324 | 0.111814 | 0.110549 | 0.34756 | 0.130572 | 0.20118 |

With this table, clustering data points based on location may be a matter of evaluating into which cluster it falls, then assigning a probability to which location it belongs. For example, if a data point falls into cluster 4, there would be a 43.7% chance it belongs to MRRY. Of course, if a data point falls into cluster 2, the likelihood of assigning it to the wrong location is much higher. In light of this, GMM is a good first step towards discovering a clustering method that accurately reflects the geographical locations of the KY Mesonet dataset.

Future work on this dataset could include implementing hierarchical and density-based clustering methods. Initially, hierarchical clustering was planned for use on this dataset; however, the 64-bit sklearn Python library was not able to compute the dataset due to memory space requirements (note: the algorithm was run on a laptop with 8GB of RAM). Hierarchical clustering could distinguish between the common and extraneous data points between all of the geographic locations and find clusters in the extraneous data that characterizes each location. Density-based clustering offers the advantage of discovering non-convex clusters, which could be the reason for such poor cluster formation. It’s also possible that these clustering methods will provide just as ambiguous results since the weather data in this region is just too similar.