John Baik

Assignment 4: Clustering

Brief Report

a. I looked around the UCI datasets that best fit the clustering model that I wanted to process. To spend less time preprocessing the data, I chose two datasets that were mostly ordinal/numerical. The poker.txt file contains comma separated values of poker hands in which clustering the data meant that the hands, consisting of 5 cards and a hand value, that were most similar to each other were clustered together.
The other dataset I used was a comma separated value dataset of forest fires which clustered similar forest fires based on location, time of year, and its intensity/destruction.

b. To process the days of the week and month of the year for the forest fires data, I converted them to numerical values in the preprocessing step with Jan-Dec being 1-12 and Mon-Sun to a 1-7 scale. This will help to see what months of the year in which forest fires occur. For the other attributes, since they were all numerical, it wasn't too necessary to change their values.

The next step was to normalize the data on a [0,1] scale so that no attribute dominates the other in scale. This was done with min-max conversion, in which every attribute was converted to a decimal value to be used in the k-means clustering algorithm.

I gave the user the option to specify the number of clusters they would like to have for the dataset as well as the threshold percentage they wanted as a necessity to count as change when finding the new mean centroid. For example, a threshold of 5 means that if a newly formed mean centroid has a percent change of less than 5% when compared to its former centroid based on the Euclidian distance of the two centroids. If one wanted to have a keener cluster set, one would set the threshold low so that the number of clustering done to the dataset using k-means would be larger than if he had chosen a larger threshold.

c. A cluster's significance can be described by the number of vectors in that cluster set. For example, a cluster of a higher count of points means that the cluster is dense and there are more vectors near that cluster than any other cluster. On the other hand, a cluster of a small size means that there are less vectors of the same time for that particular centroid and is therefore less dense when visualized. I will break down the information provided by the densest and least dense cluster sets.

Forest Fires: (ran with the following parameters: k = 7, threshold = 4%, number of vectors show per cluster = 5).
Attribute list: (pre-normalization)

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec' , (converted to 1-12)
4. day - day of the week: 'mon' to 'sun' , (converted to 1-7)
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84
(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

Largest cluster set (first 5 vectors):

```
This is the centroid for the number 6 cluster from the data set:
[2.939252336448598, 3.7757009345794397, 8.584112149532714, 3.7242990654205617,
91.65654205607477, 123.147663551402, 675.1425233644857, 9.725700934579434,
20.410280373831778, 41.92990654205609, 3.785514018691589, 0.0018691588785046728,
10.34630841121495]

The size for cluster number 6 is 179 , including the centroid.

The first 5 vectors nearest to mean centroid cluster number 6:

[6.0, 4.0, 9.0, 2.0, 91.0, 129.5, 692.6, 7.0, 18.3, 40.0, 2.7, 0.0, 0.0]
[5.0, 4.0, 9.0, 1.0, 91.8, 78.5, 724.3, 9.2, 19.1, 38.0, 2.7, 0.0, 0.0]
[6.0, 3.0, 9.0, 1.0, 88.6, 91.8, 709.9, 7.1, 11.2, 78.0, 7.6, 0.0, 0.0]
[6.0, 3.0, 9.0, 1.0, 91.8, 78.5, 724.3, 9.2, 21.2, 32.0, 2.7, 0.0, 0.0]
[6.0, 3.0, 9.0, 2.0, 90.3, 80.7, 730.2, 6.3, 18.2, 62.0, 4.5, 0.0, 0.0]
```

From the list of the densest cluster, we can see that for the following sample of attributes. I will be analyzing the most interesting ones.
From the mean-centroid:

1. X – axis = 2.94
2. Y – axis = 3.77

So most of the forest fires occurred at location (3,4) on the park map.
Map Reference: http://blog.nycdatascience.com/wp-content/uploads/2016/07/Capture.png

3. 8.58 -> August/September – occurrences for most forest fires during the year
9. 20.41 -> average temperature of the forest in Celsius
10. 41% humidity
11. 3.87 km/h
12. rain = .0018 mm
13. burned area = 10,340 square meters average

From this data, we can see that the largest cluster has a tendency of a forest fire in mid-August with each temperature near the 20.41 C (69 F). The humidity is a mere 41% average w/ low winds of 3.87 km/h and very little rain (.0018 m).

This cluster set makes sense because August/September is the time when leaves are falling and trees are drying for the preparation of winter. All the dead wood on the forest floors coupled with low humidity and low rain levels sets up for the perfect forest fire.

Smallest cluster set (first 5 vectors):

```
his is the centroid for the number 5 cluster from the data set:
[6.294117647058823, 4.784313725490194, 4.803921568627452, 2.7254901960784315,
88.65882352941175, 44.29411764705882, 192.40196078431376, 6.725490196078435,
15.503921568627455, 41.627450980392155, 4.549019607843139, .076234566232,
9.613529411764706]

The size for cluster number 5 is 31 , including the centroid.

The first 5 vectors nearest to mean centroid cluster number 5:

[6.0, 4.0, 3.0, 3.0, 89.2, 27.899999999999995, 70.8, 6.3, 15.899999999999999, 35.0,
4.0, 0.0, 0.0]
[4.0, 4.0, 3.0, 2.0, 88.1, 25.7, 67.6, 3.8, 14.099999999999998, 43.0, 2.7, 0.0, 0.0]
[4.0, 4.0, 7.0, 2.0, 79.49999999999999, 60.6, 366.7, 1.5, 23.3, 37.0, 3.1, 0.0, 0.0]
[4.0, 4.0, 3.0, 1.0, 87.2, 23.9, 64.7, 4.1, 11.8, 35.0, 1.7999999999999998, 0.0, 0.0]
[8.0, 6.0, 3.0, 5.0, 91.7, 35.8, 80.8, 7.8, 17.4, 24.0, 5.4, 0.0, 0.0]
```

From the mean-centroid:

1. X – axis = 6.29
2. Y – axis = 4.78

So least forest fires occurred near location (6,5) on the park map.
Map Reference: http://blog.nycdatascience.com/wp-content/uploads/2016/07/Capture.png

3. 4.8 -> Late April – occurrences for most forest fires during the year
9. 15.5 -> average temperature of the forest in Celsius
10. 41% humidity
11. 4.54 km/h
12. rain = .076
13. burned area = 9,614 square meters average

For the least dense cluster set, it makes sense that lesser forest fires occurred in late April because Spring usually couples with more rain (spring showers).
This is also shown with the rain in mm having a larger value than the one w/ more forest fires. The April weather is also cooler by 5 degrees C. The cool conditions, relatively low wind, and lots of rain makes it harder for forest fires to occur with this given time of the year.

d. I wanted to give the user more freedom in how they wanted the clustering to occur. By using random points as the initial centroids, I get different cluster densities w/ each run, even with the same parameters. Users have the option of picking k, the threshold for stopping in percentage, and the display of the vectors for ease of readability. Overall, it was a successful clustering as my hypothesis for forest fires throughout the year held up. So we have proven that most forest fires occur in early fall and least occur in mid-spring with this data.