Spark Program

Chapter 7: Cluster Analysis

Chapter Objectives

In this chapter, we will:

- Explore cluster analysis
- Use K-Means algorithms

Chapter Concepts

Cluster Analysis

Algorithms

Chapter Summary

Cluster Analysis

- Analysis tool to help make sense of the data before feeding it into other models
- Unsupervised
  - More about discovering patterns in data
  - Not about predicting values for unknown values
- Looks for natural groupings among the data
  - Voter groups (is it just left vs. right, or left, right, center, or more)
  - Species identification (are two groups of organisms different enough to be considered a different species or not)
  - Identify different types of customers we may have
- Often helpful as a preparatory step before classification to determine how many categories we may want to predict

Types of Cluster Analysis

- There are two main approaches to solve this
  - Top down (K-Means)
  - Bottom up (Hierarchical clustering)
- Both rely on the notion of similarity
  - Objects are similar if they share common attributes to others
  - The more similar they are, the closer they are to one another
  - If something is far away in similarity to one thing, it may be closer to something else
- Ultimately the goal is to take a large sample of data and break it up into a small number of meaningful groupings that shed insight as to what the data means

Dataset

- For this example, we have a small easy to follow dataset of the latitude and longitudes of a few Tesla superchargers

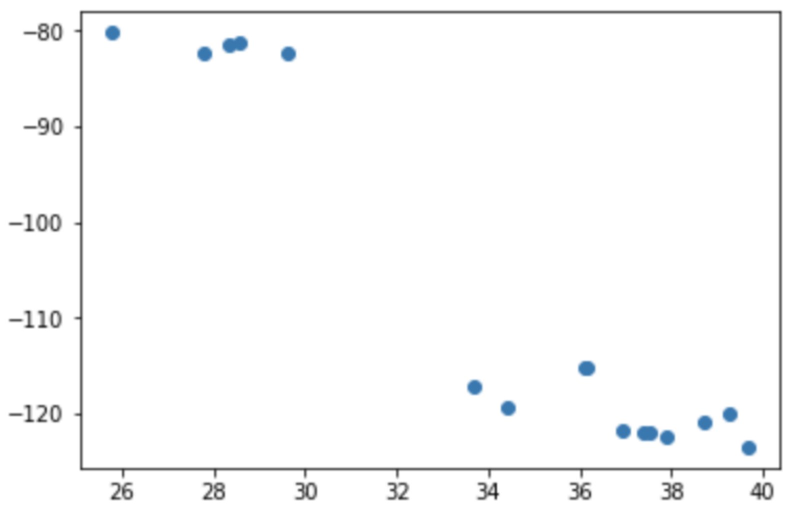
```
filename = 'superchargers.csv'
df = spark.read.csv(f'/home/student/ROI/Spark/{filename}',
header = True, inferSchema = True)
display(df)
```

	lat	lng
0	33.679646	-117.174095
1	28.331356	-81.532453
2	37.413353	-121.897995
3	37.525905	-122.006624
4	37.919969	-122.348976
5	38.730606	-120.788085
6	39.250765	-119.948927
7	36.916349	-121.773512
8	34.441994	-119.258898
9	36.116710	-115.168258

Visualize the Data

- It is often helpful to visualize the data by plotting it
  - There are only two features in this set so it's easy to plot
  - You can also plot a 3D graph for three features
  - Beyond that, it's hard to visualize more features

```
p = df.toPandas()
import matplotlib.pyplot as plt
plt.plot(p.loc[:,'lat'],p.loc[:,'lng'],'o')
```



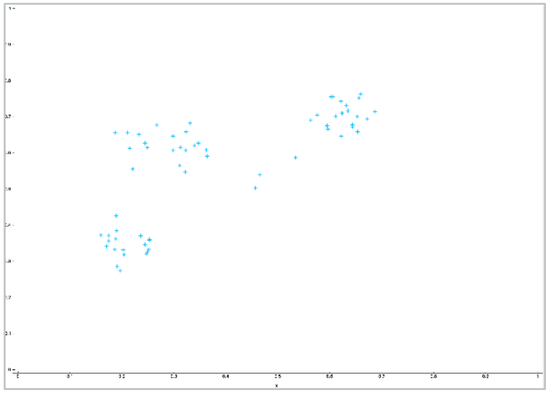
Chapter Concepts

Cluster Analysis

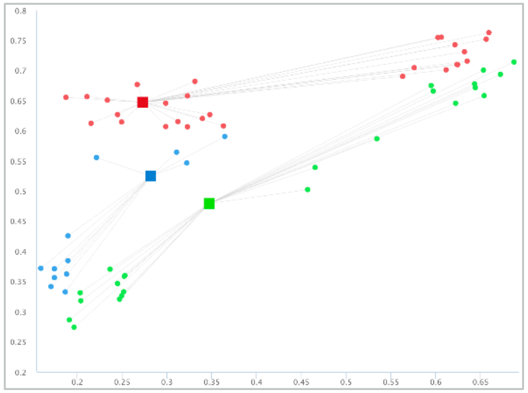
Algorithms

Chapter Summary

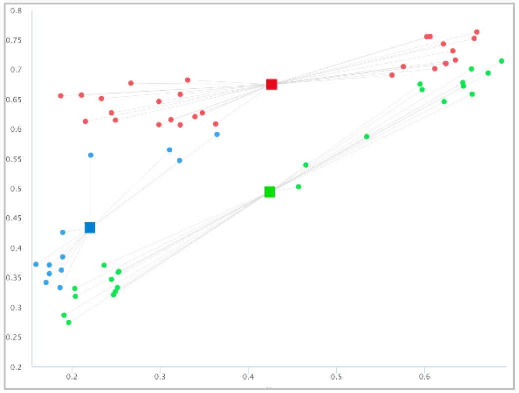
K-Means in Actions



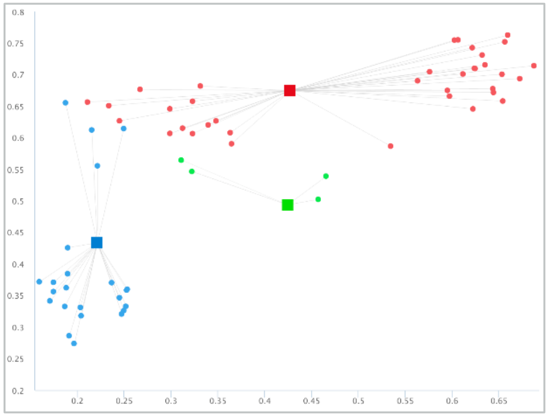
Random Data



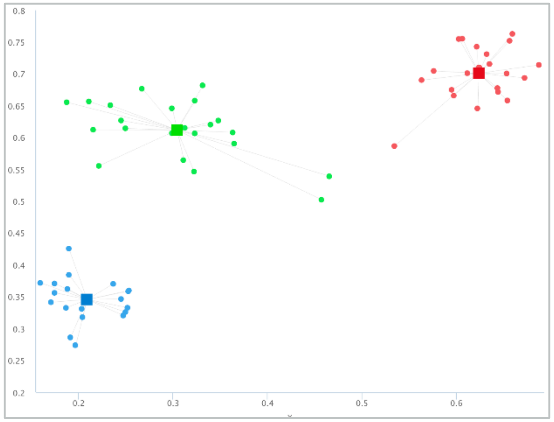
Random Centroids



Adjust Centroids



Reassign Membership



Keep Doing Until Stops Changing

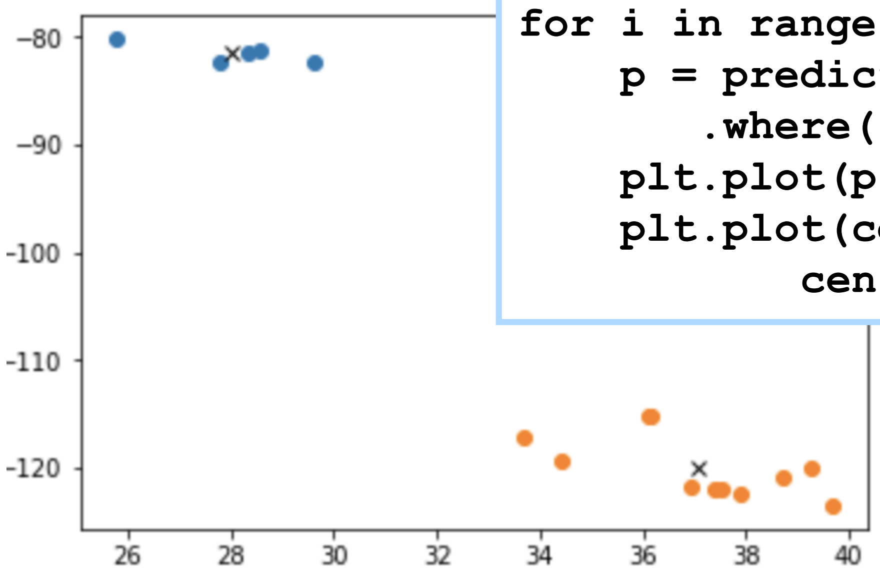
Run K-Means

- Just eyeballing it, let's try out two clusters and plot the results

```
import matplotlib.pyplot as plt

CLUSTERS = 2
kmeans = KMeans().setK(CLUSTERS).setSeed(1)
model = kmeans.fit(dfML.select('features'))
predictions = model.transform(dfML)
centroids = model.clusterCenters()

for i in range(CLUSTERS):
    p = predictions.select('lat', 'lng') \
      .where(f'prediction = {i}').toPandas()
    plt.plot(p.loc[:,'lat'],p.loc[:,'lng'],'o')
    plt.plot(centroids[i][0],
             centroids[i][1],'kx')
```

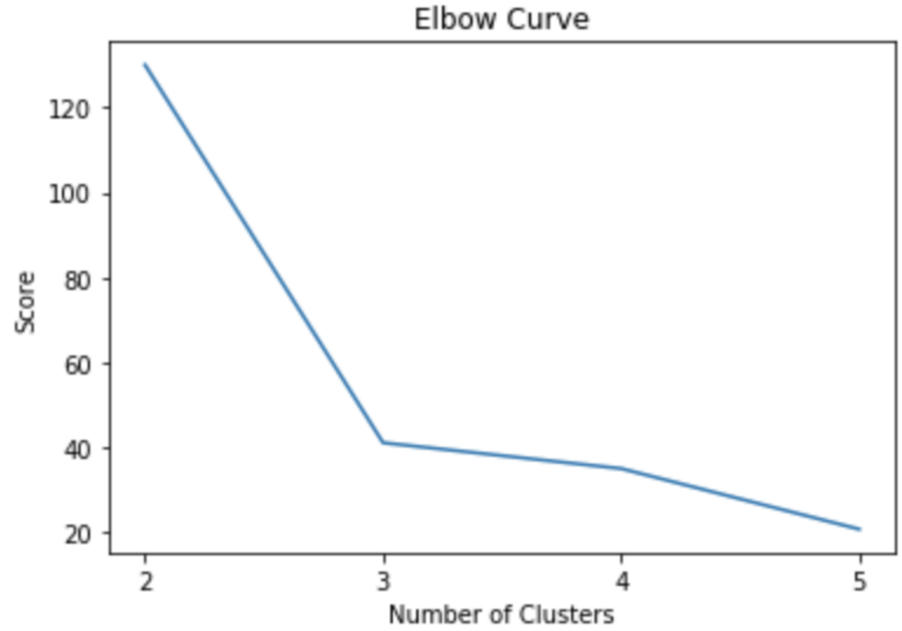


Evaluate K-Means

- With many features and rows it is difficult to visualize how many clusters if right
- There are two measures that are helpful, lower numbers are better
  - Within set sum of squared errors
  - Silhouette with squared Euclidean distance
- But it's not just about a lower number, it's about finding the marginal difference between each number until more clusters don't add any more distinctions
- Use the helper functions to view these numbers and the centroids
- For the small set of data we have, there is a big gain going from 2 to 3 clusters, but not much more going from 3 to 4, so 3 clusters is probably the right value

Elbow Chart

- Here the results are very clear cut, but sometimes the data overlap and don't fit nicely into a particular cluster
- It is often helpful to run a chart that helps figure out how many clusters is ideal
  - Too few and the items are too dissimilar
  - Too many and the additional distinctions become trivial
    - Is there much difference between a brown poodle and a chocolate poodle?



Chapter Concepts

Cluster Analysis

Algorithms

Chapter Summary

Next Steps

- The unsupervised models of clustering do not make predictions so much as they help understand the data
- Another unsupervised model to explore is association rules
  – Used to describe patterns like “people who like X also like Y”
- Principal Component Analysis
- Dimension Reduction

Chapter Summary

In this chapter, we have:

- Explored cluster analysis
- Used K-Means algorithms