Flight Data Clustering Using EP-Means

Brandon Sams

28FEB2021

Problem

In data science, there are often problems that require clustering data points together. There are plenty of industry standard methods to accomplish this, such as K-Means clustering and hierarchical clustering. However, a limitation of these methods is that they take individual data points as the input. In some instances, distributions of data points need to be clustered, rather than individual points themselves. Traditional methods do not function well when dealing with distributions as input, at least without some modifications to the clustering method.

There are many places where a distrbution clustering method would be beneficial, but this project will focus, specifically, on flight data.

Hypothesis/Research Question

This project aims to build upon the K-Means clustering method to function on a collection of distributions of data, rather that just a collection of data.

K-Means is a method for clustering data that relies on two specific tasks to build clusters. They are:

1. Compute the average value for a cluster
2. Find the "distance" between two data points

Dataset

The United States Bureau of Transportation requires idividual flight data to be publicly available. This makes an excellent dataset, because there are many distributions that can be clustered together. For example, flight duration is not a fixed value for a flight. There is some expected value for the duration of a flight, but it comes from a distribution of possible values.

So the flight duration is a good continuous value that comes from the dataset, and there are plenty of categorical values that are also included. I ended up focusing on the route, as a categorical variable. Another possible option was to use airline. I didn't create one distribution per airline, becuase I wanted to have more distributions. That would make more fascinating clusters, in my opinion.

I collected 10 years of flight data from the years of 2009-2018. Each year's data was stored as a separate csv. I merged all ten .csv files into a single file to make it easier to load into a dataframe.

Data Preparation

The individual flight data has a lot of extra information, such as flight delays, departure/arrival time, and distance. I ended up removing most of the columns, and keeping just the following:

- FL_DATE

- OP_CARRIER
- ORIGIN
- DEST
- CRS_ELAPSED_TIME
- ACTUAL_ELAPSED_TIME
- AIR_TIME
- DISTANCE

I added a new column, called "ROUTE", which was just the "ORIGIN" and "DEST", with an arrow between them. This field served as the category for splitting the data into distributions. I also removed rows with NA values, and removed any routes that contain less than 1000 flights.

Model Fitting

(Work in Progress)

Results

(Work in Progress)

References

2015 flight delays and cancellations. (n.d.). Retrieved February 14, 2021, from https://kaggle.com/usdot/flight-delays

```
This dataset offers a large download of historical data, broken down by
flight. (600MB) I plan to start the analysis with this dataset, as it is
very comprehensive and appears complete. Data is from 2015.
```

Airline on-time performance statistics—Dataset by dot. (n.d.). Data.World. Retrieved February 14, 2021, from https://data.world/dot/airline-on-time-performance-statistics

```
This dataset is very similar to the above dataset, with the exception that
the data is from 2018. Perhaps this will give us a look into how
clustering may have changed over time.
```

Aviation data & statistics. (n.d.). [Template]. Retrieved February 14, 2021, from https://www.faa.gov/data_research/aviation_data_statistics/

```
This is where the FAA hosts a large amount of statistical data for
research, directly related to the airline industry.
```

chilamkurthy, K. (2020, October 23). Wasserstein distance, contraction mapping, and modern rl theory. Medium. https://towardsdatascience.com/wasserstein-distance-contraction-mapping-and-modern-rl-

theory-93ef740ae867

```
EP-Means relies on the ability to find the “distance” between two
distributions. This article goes into a specific way of measuring distance
called the “Wasserstein Distance”, and connects this metric to modern
Reinforcement Learning.
```

Clustering probability distributions—Methods & metrics? (n.d.). Cross Validated. Retrieved February 14, 2021, from https://stats.stackexchange.com/questions/13186/clustering-probability-distributions-methods-metrics

```
This stackexchange discussion shows some alternatives to EP-Means
clustering.
```

Epmeans: Ep-means algorithm for clustering empirical distributions in maotai: tools for matrix algebra, optimization and inference. (n.d.). Retrieved February 14, 2021, from https://rdrr.io/cran/maotai/man/epmeans.html

```
This is the documentation page for a specific R package that is used for
computing EP Means.
```

Henderson, K., Gallagher, B., & Eliassi-Rad, T. (2015). EP-MEANS: An efficient nonparametric clustering of empirical probability distributions. Proceedings of the 30th Annual ACM Symposium on Applied Computing, 893–900. https://doi.org/10.1145/2695664.2695860

```
This is the landmark paper where I originally discovered EP Means. It
highlights the efficiencies of this technique in regards to other
distribution clustering techniques, and shows how one might implement this
technique.
```

Olive, X., Strohmeier, M., & Lübbe, J. (2021). Crowdsourced air traffic data from The OpenSky Network 2020 [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.3737101

```
Another possible source of air traffic data. Data is from 2020. Data
contains firstseen and lastseen properties, so duration can be computed.
```

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of wasserstein distances. Annual Review of Statistics and Its Application, 6(1), 405–431. https://doi.org/10.1146/annurev-statistics-030718-104938

```
Detailed review of Wasserstein distances (possibly showing multiple
types?)
```

Ye, J., Wu, P., Wang, J. Z., & Li, J. (2017). Fast discrete distribution clustering using wasserstein barycenter with sparse support. IEEE Transactions on Signal Processing, 65(9), 2317–2332. https://doi.org/10.1109/TSP.2017.2659647

```
Shows how Wasserstein distance can be used to cluster distributions, but
no explicit mention of EP means appears to be present in this paper.
```