Milestone2 - Project Check-In

Brandon Sams

Flight Data Clustering Using EP-Means

Any surprises from your domain from these data?

I spent some time looking into how the data was distributed, which should be able to help guide in building a clustering algorithm. I anticipated that the data was going to distrubted in a "normal" bell-shaped fashion, but the data has some skewness to it. There is a much longer tail on right side of the duration data.

When I was doing some research on distributions, it seems that this type of distribution is typically referred to as a gamma distribution. So perhaps when attempting to cluster the distibutions, there may be some transformation on the data that would make it easier to cluster.

The dataset is what you thought it was?

Luckily, I was able to track down all the individual flight data for every flight from 2009-2018. That means that rather than trying to merge or compare flight data that I found across different sources for different years, I have two realistic options before me. Either I can merge the data into one large dataset, or handle each year separately. Personally, I think that doing comparisons year-over-year would lead to better projections for how the airline industry will shift in the future (post pandemic). That being said, keeping the data in one pooled set of all the flights would make the comparisons a lot easier to accomplish, so I think I will start with that.

Have you had to adjust your approach or research questions?

My research question is primarily concerned with seeing if there is a good way to cluster distributions of data, rather than individual data points. Due to the nature of the dataset that I have available, I think that I will be sticking with that approach.

I intend on using EP means to cluster the data, which requires creating a collection of random-ish distributions to compare each distribution against. I do not know what the correct way to make this distribution is, but I will read through the paper where I found out about EP-Means again, I know the answer is in there.

Is your method working?

I was able to spot check a few routes, and the data is very consistent with how it is distributed. There is some peak (expected) duration for every flight, and long tails on either side. So when I create the CDF (cumulative distribution function) for the EP-Means process, the distribution is quite clear. I was concered that I would need to do some kernel density estimation to compare the CDFs, but from what I can tell, that will not be necessary.

What challenges are you having?

Currently, the challeng I have ahead of me is creating a CDF function for comparison. With EP means, if you want to cluster the distributions into 5 (let k = 5) clusters, then 5 different random-ish cdfs need to be created for comparison. I am having a hard time determining what the best way to create these functions is. I don't want to just choose some CDFs randomly, because they will likely be very close to each other. There needs to be some discrepancies in the distributions to make this work properly.