DATS-6203: Machine Learning II Final Project Group Proposal

Sequence Modeling to Predict Airline Destinations

Problem Statement:

Our group choose a problem predicting the most likely next location to be visited within a dense, complex geospatial dataset. We will take a dataset of individual flights including Airline, unique aircraft, flight number, departure airport, and arrival airport as a starting point. During our data cleaning and preprocessing phase, we will transform this dataset into a series of sequences each N number of airports long, grouped by unique aircraft. The resulting series of sequences will be segmented into test and training data and we will build a model to predict the next airport likely to be visited based on N-1 previous airports.

We believe this approach is applicable to numerous different problems beyond predicting airports. Our world is awash in devices that record their time and location, leading to an explosion of geospatial data that can inform everything from advertising to pandemic responses. Often, the data is too dense for traditional geospatial analysis methods and a common approach is to represent a dataset as a network or sequence of known locations visited. Applying a similar methodology as the one proposed here for our air tracking data will likely lead to additional insights in multiple fields and business cases.

Dataset:

The dataset was pulled from the Bureau of Transportation Statistics. We choose to only use data up to February 2020 to avoid impacts from the Coronavirus pandemic to affect the sequence patterns. Using six months of data from September 2019 to February 2020, we have about 3.7 million flights from 5,716 unique aircraft. There are 19 different airlines, but the top five represents about 2/3 of the entire data. Segmenting sequences into a length of 50 across the 5,716 unique aircraft produces about 1 million 50 element sequences across all aircraft data. If needed, we can pull data back for many more years to get more data.

Model and Framework:

Due to the sequential nature of the data, RNNs will be used to predict the next airport a plane belonging to a particular airline will go to. LSTMs and GRUs will be used on the data to evaluate which performs better. There are multiple airlines in the dataset with very different destination airport patterns. The problem is best approached on an airline by airline basis

because it is very difficult for a model to make sense of the noise in the training data using all airlines. Different RNNs will be trained for the top airlines and the different architectures will be discussed in the final report.

Keras will be used to implement these RNNs. Keras is a user-friendly and simple to understand deep learning framework that the members of our team have experience using. Additionally, it allows for easy implementation of layers with GRU and LSTM units.

Reference Materials:

Dataset from Bureau of Transportation Statistics - Link

Natural Language Generation with Keras, Jason Browniee - Link

Sequence Modeling Tutorial, Pulkit Sharma - Link

Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio - Link

Performance Metrics:

To evaluate the performance of models, we will use them to get predictions on a holdout test set. Our main performance metric will simply be accuracy, or how many next destinations the model got right, divided by the total number of sequences in the test set.

Rough Schedule:

We plan to finish this project over the course of the next month.

- Week 1
  - Data preprocessing and initial model is already on GitHub. Add callbacks, hyperparameter tuning, and model evaluation to scripts. Getting the scripts set up to maximize the number of experimental models we can try.
- Week 2
  - Pick airlines and create models for each. Iterate until satisfactory accuracy is reached.
- Week 3
  - Finalize all modeling, work on reports, and begin drafting presentation.
- Week 4
  - Finalize reports and presentation.