Title: Contrasting Weather Predictions using ANN, RNN, and Gradient Boosting

Catherine Chu, Paul Chirkov, Divik, Verma, Henry Moor

Summary of Project:

Our project uses 3 neural networks: Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) to capture the relationships in Weather data. We use a dataset found from Kaggle, listed in more detail in the sources section that details time series data for Weather. Since the data is time series dependent, we used RNNs to accurately reflect the temporal dependencies and patterns in weather over time. Additionally, we also wanted to contrast the use of ANNs, which can capture relationships in tabular data by learning from the input features to predict a target variable (we chose max_temperature). Finally, after multiple attempts to train the ANN, we tried using Gradient boosting, closely following the example file given in Math76: Mathematics and AI in the file problem set.

Abstract: We use RNNs to predict three critical features for each city in the dataset, based on the values of all the other features. We wanted to, given cloud cover, humidity, pressure, etc., predict the maximum, minimum, and mean temperature on a given day. RNNs are typically the preferred method to capture time series-dependent data. We also use ANNs just to compare the accuracy differences.

We then train an ANN and gradient boosting model on the data from Heathrow from 2000 to 2010. There were a total of 3654 daily observations from 18 places. We chose to train the data on Heathrow because Heathrow has similar fluctuation and temperature levels that correspond and can be extrapolated to places such as Stockholm, Oslo, and Kassel. Figure 1 demonstrates how the maximum temperatures of these 4 locations are all generalized. Since I had just visited London Heathrow, I chose it although all the predictions should be well able to generalize to the additional locations.

Dataset:

https://www.kaggle.com/datasets/thedevastator/weather-prediction/data?select=weather_prediction_bbq_labels.csv

We found this dataset on Kaggle and chose it for 2 main reasons. The first reason is because it included time series data from the longest time frame (2000 to 2010) with consistent data. It had aggregate data from 18 different locations which we found to be comprehensive. The second reason was because it included the most features. For each of the 18 locations, there are a variety of features that it examines such as cloud cover, humidity, pressure, global radiation, precipitation, sunshine, temperature mean, temperature min, and temperature max. Naturally, we select the features that we find most important to model or predict, but nonetheless, many features give us a larger dataset to choose from.

The original dataset is meant to model the BBQ rates/frequency in different locations throughout different points in the year.
Our project has a very different focus, however, we try to recreate a chart to produce the BBQ rates at the beginning of the Weather.ipynb file to ensure reproducibility.