Weather Prediction

Anuj Tanwar

Bellevue University

Abstract

Term project I have picked for DSC630 is weather prediction system. In the project, I am using historical data of US cities that contains average temperature for each day of the year for over a century. I will also use open weathermap API to extract today's temperature to compare with mean from the past.

Keywords: Weather, Prediction, US historical data, API

Weather Prediction

Model will prompt user to enter the city and search historical data of the city for current date of the month of each historical year. Then, it will calculate the mean, minimum and maximum temperature of that date. We will then call the open weathermap API to extract today's temperature to compare with mean from the past. The difference in today's prediction from historical date and today's actual data from API will become an offset for any future date's temperature predictions.

Datasets

1. **Flat File/CSVs:**

Below are the two CSV datasets I am going to use in my term project:

- **GlobalLandTemperaturesByMajorCity:**

  https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalLandTemperaturesByMajorCity.csv

- **WeatherEvents_Jan2016-Dec2020**:

  https://www.kaggle.com/sobhanmoosavi/us-weather-events/download

GlobalLandTemperaturesByMajorCity.csv dataset is Global Land Temperatures By major city from 1743 to 2013 for each date of the year. Below are the fields:

- **Dt:** Date
- **AverageTemperature:** Average temperature on the date
- **AverageTemperatureUncertainty:** Uncertainty on the mean temperature
- **City:** City from which temperature is recorded
- **Country:** Country of the City
- **Latitude:** Latitude of the location where temperature was recorded
- **Longitude:** Longitude of the location where temperature was recorded

WeatherEvents_Jan2016-Dec2020.csv dataset is a countrywide weather events dataset that includes 6.3 million events, and covers 49 states of the United States. Examples of weather events are rain, snow, storm, and freezing condition. Some of the events in this dataset are extreme events (e.g. storm) and some could be regarded as regular events (e.g.

rain and snow). The data is from January 2016 to December 2020, using historical weather reports that were collected from 2,071 airport-based weather stations across the nation. Below are the fields in dataset:

- **EventId**: This is the identifier of a record.
- **Type**: The type of an event; examples are rain and snow.
- **Severity**: The severity of an event, wherever applicable. The severity of an event, wherever applicable.
- **StartTime(UTC)**: The start time of an event in UTC time zone.
- **EndTime(UTC)**: The end time of an event in UTC time zone.
- **TimeZone**: The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).
- **AirportCode**: The airport station that a weather event is reported from.
- **LocationLat**: The latitude in GPS coordinate of airport-based weather station.
- **LocationLng**: The longitude in GPS coordinate of airport-based weather station.
- **City**: The city in address record of airport-based weather station.
- **County**: The county in address record of airport-based weather station.
- **State**: The state in address record of airport-based weather station.
- **ZipCode**: The zipcode in address record of airport-based weather station.

2. API:

https://openweathermap.org/current

API enables user to access current weather data for any location on Earth including over 200,000 cities! Data is collected and processed from different sources such as global and local weather models, satellites, radars and a vast network of weather stations. Data is available in JSON, XML, or HTML format.

Project Execution

GlobalLandTemperaturesByMajorCity.csv will be the main driver dataset to find the historical average temperature of a given city for a date and mean of those temperature values will be used as a projected value for current year.

WeatherEvents_Jan2016-Dec2020.csv dataset will be used to extract type and severity of event of the city while displaying it for the user.

API will be used to extract current day's weather and to find the offset between value between projected value and actual value return from API. This offset will be used to determine future weather temperatures.

Relationship between various datasets will be based on City and date for which data will be extracted.

Type of Model

I am planning to use time series modeling with Auto Regressive Integrated Moving Average models (ARIMA). An ARIMA model is a class of statistical models for analyzing and forecasting time series data. Since ARIMA is used to measure events that happened over a period of time that is why I want to use it to understand past weather data and predict future weather events in a series.

How do you plan to evaluate your results?

Below are the steps I plan to execute to evaluate the results:

1. Read the dataset and filter data for the proposed city.
2. Clean the data by dropping the duplicates, find and filter out outliers using box plot, taking care of NaN
3. Splitting the data in train and test and create a model on top of it
4. Creating time series plot of the clean dataset
5. Finally predicting future values

What do you hope to learn?

I am hoping to explore and learn ARIMA modelling in detail during the building of the term project.

Assess any risks with your proposal

1. ARIMA itself cannot handle outliers
2. Finding the best values of p (number of autoregressive terms), d (number of nonseasonal differences), q (number of lagged forecast errors) parameters in ARIMA is essential else we might end up with overfitting and computations stress.

Contingency plan

If ARIMA modelling does not work then I will try to build LSTM (Long-Short Term Network) model. LSTM is able to 'remember' long-term dependencies/information from a feature. Information runs across the networks and passes down such that earlier information from a sequence could influence prediction of later input. This passing of information is crucial for time-series weather data where each input is dependent on other input from varying time points.

Potential Challenges

Main anticipated challenge will be to clean the data and join different datasets. Datasets might contain null values and outliers which will need to be cleaned. Strategy to handle those null values and outliers need to be analyzed and implemented, which could potentially impact the final prediction. Getting weather prediction is always tricky and there are external factors that impact the changing weathers hence the model is never 100% accurate.