Machine Learning Engineer Nanodegree

Capstone Proposal

Gurnekpreet Sandhu

2020

Domain Background

The purpose of this capstone project is to predict the future climate of the city of Vancouver based on a dataset retrieved from kaggle which gives us Canadian weather data from the past eighty years.

Making use of the data, we have a very detailed set of data which can be used to predict the weather in the future, and also be used to see how much of an effect climate change is having on Vancouver

Problem Statement

For this project, I will look to build a weather predictor that takes weather data over a certain date range as input, and outputs the predictions for given query dates.

I wish to see for myself using a machine learning model how much the average temperature has increased since the 1940s in Vancouver and how by how much it will increase in the future if current trends continue

Datasets and Inputs

I plan to use a magnitude of data coming from a single kaggle dataset
1) Eighty years of canadian climate data by anna turner
(https://www.kaggle.com/aturner374/eighty-years-of-canadian-climate-data)

I plan to use the `LOCAL_DATE` column to find the date of when each temperature reading was recorded and the `MEAN_TEMPERATURE_VANCOUVER` column to find the actual temperature of the date in the `LOCAL_DATE` column.

Solution Statement

The solution to the issue is that if we get the machine learning models accuracy better than moving average, which has been proven to be quite capable for time series problems. We will test this on 10 years of data. I will compare the accuracy of the moving average to the machine learning model on the 10 years of testing data. If the machine learning model is better than the moving average, then the problem is solved, or else another type of model will need to be tested, until the moving average's accuracy is beat.

Benchmark Result

The benchmark result will be the data from 2010-2020. The models I use will train using the data from 1940-2009, and will test on the 2010-2020 data. I will first start with moving average to get an accuracy. It will be the benchmark for model accuracy.

I will keep trying to be as accurate as I can without overfitting. The final model will have to be as accurate as possible and more accurate than moving average to be considered production ready.

Evaluation Metrics

The model will be evaluated on the mean squared error (dividing the sum of squares of the residual error by the degrees of freedom). The smaller the mean squared error, the more accurate our model.

Project Design

#1) The full dataset which was downloaded from Kaggle contains data for all the different major weather centers in canada from the 1940s and onward. I only wish to find the temperature for Vancouver, so I will have to get rid of all the other columns but `LOCAL_DATE` and `MEAN_TEMPERATURE_VANCOUVER`.

#2) Next I will process the data. This means removing all the NAN datapoints in the dataset and replacing it with the mean of the column to get the most accurate result

#3) I will do the moving average part of this project, which is predicting 2010-2020 temperature. I will record the mse of the test set predictions and start with the next part.

#4) I will try different neural networks, starting with the sagemaker DeepAR model. I will try to train the model based on past data (1940-2009) and try to estimate the temperature for the test set. I will also record the accuracy of this model. If it beats the moving average, then the model will be considered production ready, or else I will have to try a different model or build one on my own using keras and tensorflow.

#4) if the model beats the moving average, it will be production ready. I will then create a report that explains thoroughly what happened and exactly what steps I took.