MLND - Project Proposal

Louis Tian

August 20, 2017

1 Domain Background

Forecasting or time series prediction is an import subfield of machine learning. Since the dawn of humanity, we have been seeking the future telling crystal ball. There are countless applications ranging from traditional science, finance to robotic. While being a very important field, Time series prediction has been neglected in the recent years of up rising of machine learning and AI.

Compared to other fields of machine learning, data in time series problem are relatively scarce. One of the most successful areas in machine learning is computer vision, the prevalence of smart phone and digital camera created a huge wealth of data for computer vision, easily available to researchers. On the other hand, time series data are much harder to come by. Good quality financial market data are not only limited and prohibitively expensive to acquire.

One might argue that the time series prediction problem is an intrinsically harder problem. While in computer vision and speech recognition, the algorithms are merely catching up to human's ability, time series prediction problem has always been on the frontier of human level ability. While a three-year-old could tell the difference between a dog and a cat, only the very best of us are able to make an accurate prediction of future.

This combination of intrinsic difficulty and usefulness makes time series problem a very interesting topic to study

2 Problem Statement

The proposed problem is to predict the page visit traffic for a large quantity of Wikipedia web pages based on their historical visits data.This is based on the Kaggle Web Traffic Time Series Forecasting Competition[1].

[1] https://www.kaggle.com/c/web-traffic-time-series-forecasting

3 Datasets and Inputs

The data set for this project is available at Kaggle.com. The data set includes daily web traffic data for a total of 145,063 pages for the period from 01/07/2015 to 31/12/2016. For the purpose of this project, only a subset of 100 pages of the original data set will be used.

This data set does not distinguish between traffic values of zero and missing values. A missing value may mean the traffic was zero or that the data is not available for that day.

4 Solution Statement

This project proposed to transform this traditional time series problem into a supervised learning problem and use neural nets to forecast the web traffic for a fixed length out of sample period. Features of the supervised problem can be created using the historical web traffic. Examples of those could include lagged observation for a fixed period of windows, mean/median web traffic for lagged window period.

5 Benchmark Model

The traditional approach to time series problem is to fit the time series with ARIMA model. ARIMA model is considered as the most general form time series model in traditional statistically methods. However, the general form of ARIMA requires a significant amount of hyper-parameter tuning makes is unsuitable for our purpose as the benchmark model. For benchmarking, we will restrict to ARIMA(1,1,0), which basically assumes the time series is generated by a random walk process. Under this setting, the best prediction is simply the latest observed value in the time series.

6 Evaluation Metrics

The official evaluation criteria used in the Kaggle competition is SMAPE defined as:

$$SMAPE = \frac{1}{n}\sum_i \frac{|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i|} \tag{1}$$

7 Project Design

The overall work flow of the proposed project is summarised in this section.

1. Data Exploration and Visualisation

Calculate basic statistics about the data set and create some visualisation to understand the fundamental characteristics of the dataset.

2. Data Preparation

Reserve the last 14 days of data as the validation data set

3. Calculate the benchmark model performance metric

ARIMA(1,1,0) benchmark metrics should be calculated.

4. Build the models in Tensorflow

The model will be built in Tensorflow for performance and flexibility.

5. Hyperparameter tuning with Grid Search

Given that only 100 pages will be used, it is computationally feasible to tune our parameter by using simple grid search.

6. Model Evaluation

Train the model on the entire training set and evaluate the performance by making the 14 days out of sample prediction. Using the predictions and the validation set to calculate the SMAPE score and compare against the benchmark value.