Machine Learning Engineer Nanodegree

Capstone Proposal

Thiago Junqueira Vilarinho
March 12th, 2018

Proposal

Domain Background

Since taxi companies use vehicle's allocated time and trip distance to charge the passenger's trip, be able to predict the price of trip based on this information helps in the audit process, estimates of the company's profit or even charge the passenger before the trip, like Uber.

With information of every trip about when and where the passenger were picked up and droped off and how many passengers triped will allow the company to estimate how many time vehicle will be allocated and, based on that information, how much passenger should pay.

Problem Statement

The problem to be solved is based on the distance between pick up and drop off points, pick up time and the number of passengers, estimate travel time using neural network.

The proposal is to define weigths to calculate ETT as:
$ETT = W1 * D + W2 * T + W3 * P$

Where:
D is th distance between pick up and drop off points
T is the pick up time
P is the number of passengers

The evaluation metric to this solution is Root Mean Squared Logarithmic Error, calculated as:
$E=\sqrt{1 / n * \sum (\log(pi + 1) - \log(ai + 1))^2}$

Where:
E is the RMSLE value (score)
n is the total number of observations in the (public/private) data set
pi is your prediction of trip duration
ai is the actual trip duration for i
log(x) is the natural logarithm of x

Datasets and Inputs

This project is based on 2016 NYC Yellow Cab trip record data and relates to the problem because it has the trip distance, trip duration and number of passengers. Since we are trying to estimate

travel time on New York City, this is an appropriate dataset for the problem.

This dataset was obtained on Kaggle Competition New York City Taxi Trip Duration (https://www.kaggle.com/c/nyc-taxi-trip-duration) and is splitted on train and test datasets. Train dataset contains 1.458.644 trip records and Test dataset contains 625.134 trip records.

Data fields:
id - an unique identifier for each trip
vendor_id - a code indicating the provider associated with the trip record
pickup_datetime - date and time when the meter was engaged
dropoff_datetime - date and time when the meter was disengaged
passenger_count - the number of passengers in the vehicle (driver entered value)
pickup_longitude - the longitude where the meter was engaged
pickup_latitude - the latitude where the meter was engaged
dropoff_longitude - the longitude where the meter was disengaged
dropoff_latitude - the latitude where the meter was disengaged
store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
trip_duration - duration of the trip in seconds

This dataset will be used to train neural network with backpropagation to define weights on each perceptron to output the estimate trip duration. After, will be used Test Dataset to evaluate performance of this neural network on estimate trip duration using Root Mean Squared Logarithmic Error as evaluate metric.

Solution Statement

The solution to the problem is using MLPRegressor that implements a multi-layer perceptron that trains using backpropagation without activation function on the output layer. This neural network should be trained to output the estimate time travel (in seconds) based on input set as: pick up and drop off location, pick up time and number of passengers.

This regressor will use Adam Stochastic Gradient-Based optimizer as loss function to update weights and Root Mean Squared Logarithmic Error as evaluate metric.

Benchmark Model

The benchmark model that relates to the problem statement is the kernel published by Danijel Kivaranovic for the same Competition (https://www.kaggle.com/danijelk/beat-the-benchmark), where the score reached was 0.39016 with the same metric (Root Mean Squared Logarithmic Error).

Evaluation Metrics

The evaluation metric to this solution is Root Mean Squared Logarithmic Error, calculated as:

$$E=\sqrt{1 / n * \sum (\log(pi + 1) - \log(ai + 1))^2}$$

Where:
E is the RMSLE value (score)
n is the total number of observations in the (public/private) data set
pi is your prediction of trip duration
ai is the actual trip duration for i
log(x) is the natural logarithm of x

Project Design

The strategy to be employed on the analysis of this problem is:
1 - Read data
2 - Search for duplicated records and clean data
3 - Adjust data type for each column
4 - Create a new column with distance between pick up and drop off using geopy
5 - Normalize pick up time to consider just the hour
4 - Relate distance, hour and number of passenger with time duration using histograms
5 - Train and Test MLPRegressor to find the best configuration
6 - Take the evaluate metric to compare with benchmark