New York City Taxi Fare Prediction

Runsha Pan[1]

[1] Hunan University, China

Introduction

In this kaggle topic selection competition, you have to predict the taxi fare according to the known features of the passen- gers taking taxis, which including longitude coordinate and latitude co- ordinate of where the taxi ride started and so on. While you can get a basic estimate based on just the distance between the two points, this will result in an RMSE of 5–8, the challenge is to do better than this using Machine Learning techniques!

Date files The competition gives three documents train.csv, test.csv, sample_submission.csv.

Data fields There are six features in the test and training sets,which includes "pickup_datetime", "pickup_longitude", "pickup_latitude", "dropoff_longitude", "dropoff_latitude", "passenger_count".

The Target The target is predict "fare_amount " field in "test.csv" file.

The evaluation metric

- The evaluation metric for this competition is the root mean-squared error or RMSE. RMSE measures the difference between the predictions of a model, and the corresponding ground truth. A large RMSE is equivalent to a large average error, so smaller values of RMSE are better. One nice property of RMSE is that the error is given in the units being measured, so you can tell very directly how incorrect the model might be on unseen data.
- RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where y_i is the ith observation and $\hat{y}_i$ is the prediction for that observation.

Data viewing and preliminary analysis

Data Viewing The author analyzed the data through jupyter.Firstly, for the imported data, remove the first five rows of the training set and the test set to see the general situation of the data, and output the field and data type information.Among them, since the training set is too large and contains 5400W rows, the first 200W rows are selected for training in order to save running time. And calculate the number of data contained in the test set and the training set, the mean, variance, standard deviation, minimum value, maximum value, quartile and median, in order to understand the basic situation of the data, from which you can roughly understand the simple abnormal situation of the data.

Data Visualization Combined with the data description results of the training set, the histogram of passenger consumption in the training set was made, and the outliers of passenger consumption were screened and removed.

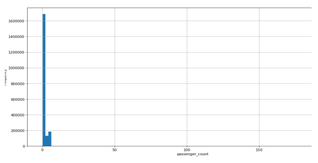

figure 1:100 groups

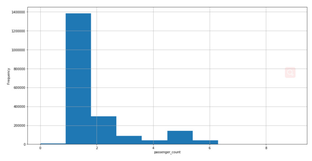

figure 2:10 groups,x<10

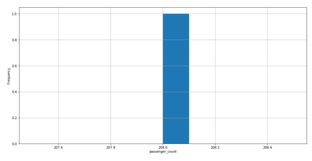

figure 3:10 groups,x>=10

Data Conversion The timestamp data types in the training set and the test set were converted into numerical data that were easy to analyze and convenient for subsequent analysis.

Data Calculation Exploring the data further. Firstly, 14 rows with missing values were deleted, resulting in 1999986 remaining rows. Secondly, handling outliers. Trim the rows with negative taxi fares, pick-up/drop-off longitude outside the range (-180-180), pick-up/drop-off latitude outside the range (-90-90).

Data Reduction and computation

Through the histogram drawing, the outliers with the number of passengers greater than 10 are found, and a record is screened out for deletion.After pruning, 1999822 pieces of training set data were obtained.

In order to make the data of the training set and the test set closer and simplify the training samples, the longitude and latitude range of the test set was found out, and the training set data was framed in this range to perform data pruning. The post-construction training dataset contains 1957,917 records.

Distance the distance between the pick-up and drop-off locations of the training set and the test set was calculated according to the Haversine Equation, and a new field was formed and added to the data set. The records with zero distance between taxi fare and pick-up and drop-off location are invalid, and 1957913 training set data are obtained after deletion.

Model Building

Model The expression of the multivariate linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{0_i} + \beta_2 X_{1_i} + \cdots + \beta_k X_{k_i} + \mu$$

where i=1, 2, ..., n.

Matrix representation:

$$Y = X\beta + \mu$$

Solution In this paper, the pick-up and drop-off distance, travel time (day of the week), and the number of passengers are used as independent variables to construct matrix X for solving:

$$\begin{bmatrix} x_0 & \dots & x_0^n & 1 \\ x_1 & \dots & x_1^n & 1 \\ & \dots & & \\ x_n & \dots & x_n^n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_n \\ \beta_0 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \dots \\ y_n \end{bmatrix}$$

It is solved using the least squares method:

$$\beta = (X^T X)^{-1} X^T Y$$

The multiple regression equation is obtained as follows:

$$Y_i = 4.46 + 2.10 X_{0_i} - 0.05 X_{1_i} + 0.04 X_{2_i}$$

where X_{0_i} is the value of "H_Distance", X_{1_i} is the value of "weekday+1", X_{2_i} is the value of "passenger_count".

Conclusion

Predicted results Putting the relevant fields of the prediction set into the regression equation yields the predicted value of "fare_amount" :

	key	fare_amount
0	2015-01-27 13:08:24.000000200	9.29
1	2015-01-27 13:08:24.000000300	9.50
2	2011-10-08 11:53:44.000000200	5.51
...	...	...

Acknowledgement

- The teacher elder sister, group leader and classmates, who gave the author great advice and tutorials.

LaTeX TikZposter