CS777

Big Data Analytics

Spring 2021

Professor Kia Teymourian

Assignment 3

Kunfei Chen

U15575304

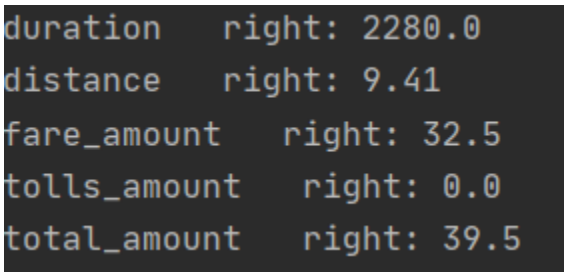
Introduction

This project aims to apply Batch Gradient Descent via PySpark to achieve Simple Linear Regression and Multiple Linear Regression in the New York City Taxi trip reports which is a dataset released by Chris Whone.

Task 1 – Simple Linear Regression

Firstly, the dataset is filtered to ensure the selected columns which consist of duration, trip distance, fare amount, tolls amount, and total amount can be converted into float type and are within the range between 0 and right outlier boundaries (See Figure 1). Since the outlier boundaries are calculated based on small data set and the right boundary of tolls amount should not be 0, the value is changed into 39.5 which is the same as that of total amount.

Figure 1



Task 1 amis to calcualate the exact answers for the parameters m and b of the simple linear regression model for varibales of trip distance and fare amount, based on the equation shown in Figure 2. In addition, the RDD of data for simple linear regression is stored in disk by the "persist ()" function, which is prepared for the training in Task 2.

Figure 2

$$Y = mX + b \tag{1}$$

$$\hat{m} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} \tag{2}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} {x_i}^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} \tag{3}$$

The calculated value or parameters are shown in Figure 3 that the m is 2.835 and b is 4.02, it also shows that the cost time of this task is about 1045 seconds.

Figure3

```
================= Task1 Simple Linear Regression ===================
21/03/22 06:03:03 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
x: 355010571.36, y : 1652514022.55, x_squared: 1255263327.72, y_squared: 21405082859.1, xy: 4986026597.72, n: 160690618.0
Equation of Simple Linear Regression : fare = 2.83506308205 * distance + 4.02037571321
Time Cost: 1044.49630308
```

Task 2 – Find the Parameters using Gradient Descent

This task will implement the gradient descent to find the optimal parameters for the simple linear regression model. The initial vale of parameters of m and b are set to be 0.1, the learning rate is set to be 100, and the maximum number of iteration is set to be 100. Since this is a batch gradient descent model, for each iteration, 10000 samples are taken from the whole dataset for the training. The costs and model parameters in each iteration is shown in Figure 4. It can be seen that after training 100 iterations, the cost decreased from 1256567 to 109314, and the trained value for m and b in the last iteration are 2.975 and 1.159 respectively.

Figure 4

```
================= Task2 Find the Parameters using Gradient Descent ===================
(0, 'weight: 0.159802503066  bias: 0.1200109778   cost: 1256566.66057 ')
(1, 'weight: 0.219414922372  bias: 0.139703338455   cost: 1221598.52705 ')
(2, 'weight: 0.277955789704  bias: 0.15907694705   cost: 1183733.27595 ')
(3, 'weight: 0.336320047047  bias: 0.178255700156   cost: 1159152.33054 ')
(4, 'weight: 0.392597956873  bias: 0.196985478616   cost: 1101553.58921 ')
(5, 'weight: 0.448220245818  bias: 0.215469384481   cost: 1070236.46812 ')
(6, 'weight: 0.502306470308  bias: 0.233657515747   cost: 1031094.32841 ')
(7, 'weight: 0.555560801321  bias: 0.251533777331   cost: 997762.069495 ')
(8, 'weight: 0.608299790681  bias: 0.269222833225   cost: 978329.105985 ')
(9, 'weight: 0.659255259257  bias: 0.286492759762   cost: 927851.97007 ')
(10, 'weight: 0.7100926516  bias: 0.303535019469   cost: 904159.287082 ')
(11, 'weight: 0.759355112001  bias: 0.320296758993   cost: 872441.20807 ')
(12, 'weight: 0.80808327338  bias: 0.33683588596   cost: 847938.381266 ')
(13, 'weight: 0.855706855344  bias: 0.353083715862   cost: 818838.622584 ')
(14, 'weight: 0.902787943118  bias: 0.369153730478   cost: 796848.067131 ')
(15, 'weight: 0.948104670084  bias: 0.384854519442   cost: 763730.504346 ')
(16, 'weight: 0.993455849836  bias: 0.400496213065   cost: 754820.015678 ')
(17, 'weight: 1.0384307633  bias: 0.415869194675   cost: 729070.809232 ')
```

```
(83, 'weight: 2.77900946328  bias: 1.06862277301   cost: 140176.855501 ')
(84, 'weight: 2.79305507926  bias: 1.07476064101   cost: 133081.591284 ')
(85, 'weight: 2.80714815192  bias: 1.08091922955   cost: 133158.411694 ')
(86, 'weight: 2.82071037268  bias: 1.08693513096   cost: 129818.665697 ')
(87, 'weight: 2.83395130828  bias: 1.09285242742   cost: 125358.057836 ')
(88, 'weight: 2.84684735873  bias: 1.09866702918   cost: 121135.365259 ')
(89, 'weight: 2.85959167777  bias: 1.10443251459   cost: 122130.859385 ')
(90, 'weight: 2.87200911563  bias: 1.11016497046   cost: 121573.272621 ')
(91, 'weight: 2.88416651109  bias: 1.11582036299   cost: 120904.249481 ')
(92, 'weight: 2.89624432809  bias: 1.12149311414   cost: 122605.509066 ')
(93, 'weight: 2.90812002195  bias: 1.12703740687   cost: 115658.73308 ')
(94, 'weight: 2.919740493  bias: 1.13246342505   cost: 114708.934888 ')
(95, 'weight: 2.93126410829  bias: 1.13789056581   cost: 112596.585248 ')
(96, 'weight: 2.94255636811  bias: 1.14319421874   cost: 109457.660947 ')
(97, 'weight: 2.95346678021  bias: 1.1484621323   cost: 108225.024473 ')
(98, 'weight: 2.9643240332  bias: 1.15371106098   cost: 107843.535041 ')
(99, 'weight: 2.97513573073  bias: 1.15892436014   cost: 109314.124075 ')
```

Task3 – Fit Multiple Linear Regression using Gradient Descent

The Task3 amis to implement the gradient descent to find the optimal parameters for the multiple linear regression for 4 variables of duration, trip distance, fare amount, tolls amount, so as to predict the total paid amounts of Taxi rides. Similarly to Task2, the initial parameters are set as 0.1, the learning rate is set to be 0.001 the maximum number of iteration is set as 100. In addition, the training of this model applys “Bold Driver” technique to dynamically change the learning rate. To be specifically, for each iteration, if the cost decrease, the learning rate will be multiplied with 1.05, otherwies, it will be multiplied with 0.5. The costs and model parameters in each iteration is shown in Figure 5. It can be seen that after training 100 iterations, the cost experienced an extreme increase during the first 20 epochs, then it started slowly decrease due to the change of learning rate. The reason of this increase is the initial learning rate is too big for this model. To solve this problem, practical method could be reducing initial learning rate or adding activation function for the input value such as sigmoid function.

Figure 5

```
================= Task3 Fit Multiple Linear Regression using Gradient Descent ===================
21/03/22 08:09:14 WARN org.apache.spark.scheduler.TaskSetManager: Stage 303 contains a task of very large size (122 KB). The maxim
(0, 'weights: [-1.02304128e+02 -2.26209559e-01 -1.36921283e+00  9.18498035e-02]  bias: -0.0116010256   cost: 43539827.1663 ')
21/03/22 08:11:50 WARN org.apache.spark.scheduler.TaskSetManager: Stage 306 contains a task of very large size (121 KB). The maxim
(1, 'weights: [1.20545074e+05 3.89850605e+02 1.74695984e+03 1.33760216e+01]  bias: 134.884568269   cost: 6.17363357952e+13 ')
21/03/22 08:14:26 WARN org.apache.spark.scheduler.TaskSetManager: Stage 309 contains a task of very large size (121 KB). The maxim
(2, 'weights: [-7.14750316e+07 -2.32160631e+05 -1.03926561e+06 -7.09519460e+03]  bias: -79656.1007506   cost: 8.63241298228e+19 ')
21/03/22 08:17:01 WARN org.apache.spark.scheduler.TaskSetManager: Stage 312 contains a task of very large size (121 KB). The maxim
(3, 'weights: [2.11250704e+10 6.94188089e+07 3.08213014e+08 1.94168237e+06]  bias: 23553319.0413   cost: 3.03072643186e+25 ')
21/03/22 08:19:37 WARN org.apache.spark.scheduler.TaskSetManager: Stage 315 contains a task of very large size (121 KB). The maxim
(4, 'weights: [-3.10416787e+12 -9.99395828e+09 -4.49152942e+10 -3.23661482e+08]  bias: -3459891227.01   cost: 2.64147012255e+30 ')
21/03/22 08:22:13 WARN org.apache.spark.scheduler.TaskSetManager: Stage 318 contains a task of very large size (121 KB). The maxim
(5, 'weights: [2.31067075e+14 7.33498205e+11 3.31855769e+12 1.94565354e+10]  bias: 2.55210688956e+11   cost: 5.81653000524e+34 ')
21/03/22 08:24:48 WARN org.apache.spark.scheduler.TaskSetManager: Stage 321 contains a task of very large size (121 KB). The maxim
(6, 'weights: [-8.26866300e+15 -2.68530723e+13 -1.19952774e+14 -8.15151926e+11]  bias: -9.27013501657e+12   cost: 3.1431032467e+38
21/03/22 08:27:25 WARN org.apache.spark.scheduler.TaskSetManager: Stage 324 contains a task of very large size (121 KB). The maxim
(7, 'weights: [1.47878515e+17 4.78715815e+14 2.14227968e+15 1.63938330e+13]  bias: 1.63133955804e+14   cost: 4.13252781847e+41 ')
21/03/22 08:30:01 WARN org.apache.spark.scheduler.TaskSetManager: Stage 327 contains a task of very large size (121 KB). The maxim
(8, 'weights: [-1.22630309e+18 -3.95537181e+15 -1.77452104e+16 -9.92872355e+13]  bias: -1.36667396195e+15   cost: 1.30084425192e+4
21/03/22 08:32:37 WARN org.apache.spark.scheduler.TaskSetManager: Stage 330 contains a task of very large size (121 KB). The maxim
(9, 'weights: [4.45793787e+18 1.44380286e+16 6.46384808e+16 4.04202588e+14]  bias: 4.95773993937e+15   cost: 8.92434640635e+45 ')
21/03/22 08:35:12 WARN org.apache.spark.scheduler.TaskSetManager: Stage 333 contains a task of very large size (121 KB). The maxim
```

```
(90, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 7.96981520554e+34 ')
21/03/22 12:06:08 WARN org.apache.spark.scheduler.TaskSetManager: Stage 576 contains a task of very large size (122 KB). The maximum re
(91, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 8.12288993109e+34 ')
21/03/22 12:08:44 WARN org.apache.spark.scheduler.TaskSetManager: Stage 579 contains a task of very large size (121 KB). The maximum re
(92, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 8.17667805096e+34 ')
21/03/22 12:11:20 WARN org.apache.spark.scheduler.TaskSetManager: Stage 582 contains a task of very large size (121 KB). The maximum re
(93, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 8.10496386125e+34 ')
21/03/22 12:13:55 WARN org.apache.spark.scheduler.TaskSetManager: Stage 585 contains a task of very large size (121 KB). The maximum re
(94, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 8.44928613146e+34 ')
21/03/22 12:16:31 WARN org.apache.spark.scheduler.TaskSetManager: Stage 588 contains a task of very large size (121 KB). The maximum re
(95, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 7.79895261797e+34 ')
21/03/22 12:19:06 WARN org.apache.spark.scheduler.TaskSetManager: Stage 591 contains a task of very large size (121 KB). The maximum re
(96, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 7.91027497084e+34 ')
21/03/22 12:21:42 WARN org.apache.spark.scheduler.TaskSetManager: Stage 594 contains a task of very large size (121 KB). The maximum re
(97, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 8.22981016286e+34 ')
21/03/22 12:24:17 WARN org.apache.spark.scheduler.TaskSetManager: Stage 597 contains a task of very large size (121 KB). The maximum re
(98, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 7.86084277542e+34 ')
21/03/22 12:26:54 WARN org.apache.spark.scheduler.TaskSetManager: Stage 600 contains a task of very large size (121 KB). The maximum re
(99, 'weights: [-1.45888330e+13  2.83440464e+14  9.31062116e+14  3.24613844e+14]  bias: 1.32584457462e+14   cost: 8.0454024764e+34 ')
21/03/22 12:26:54 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@270ac6f9{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```