Distributed Logistic Regression

Author's Name: Wu Yidi Supervisor's Name: James Cheng
the Chinese University of Hong Kong
1155047054@link.cuhk.eud.hk

The Husky Platform

- Husky [2] is developed by professor James Cheng' team aiming at developing a more expressive and most importantly, more efficient systems for distributed data analytics.
- The most impressive feature of Husky is that it gains a great balance between high performance and low development cost.

Logistic Regression

- The central equation of the logistic regression is the sigmoid function

$$h_\theta(x) = \frac{1}{1+e^{-\theta x}} \tag{1}$$

- In order to estimate the value of θ, a cost function is constructed [1]

$$l\left(\vec{Y}\middle|X;\theta\right) = \sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + (1-y^{(i)}) \log(1-h_\theta\left(x^{(i)}\right)) \tag{2}$$

- After optimizing this cost function with respect to θ, we can classify new coming data using following criteria:

$$y = \begin{cases} 0 & \text{if } h_\theta(x) >= 0.5 \\ 1 & \text{else} \end{cases}$$

Gradient Descent

- This cost function can be optimized using gradient descent algorithm which repeatedly steps towards the opposite direction of the gradient of the function at current point.

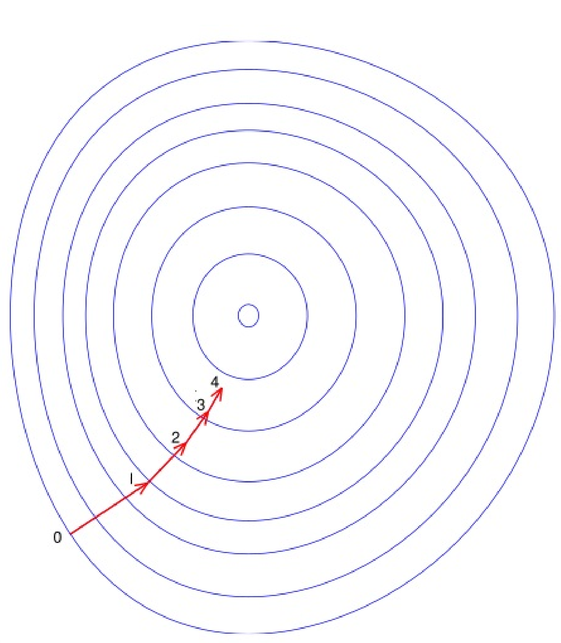

Figure 1. Graphical view of Gradient descent

- The updating rule for logistic regression is given by
For i = 1 to m,

$$\theta := \theta + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right) * x^{(i)} \tag{3}$$

- α is called learning rate which is set by the user.

Implementation strategy

- First we partition the data into several parts and assigns each part to a worker. The parmeter θ is stored globally.
- In every iteration, each worker aggregates the updates locally and send the aggregated result to the global parameter. Then Husky updates the global parameter by summing all the updates.

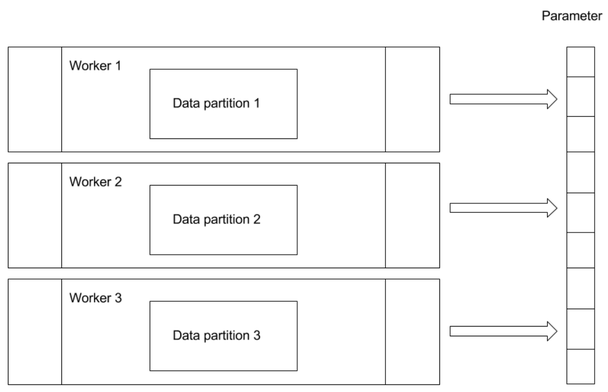

Figure 2. The distribution strategy

Performance

- The scalability can be measured by the relationship between time and number of workers.

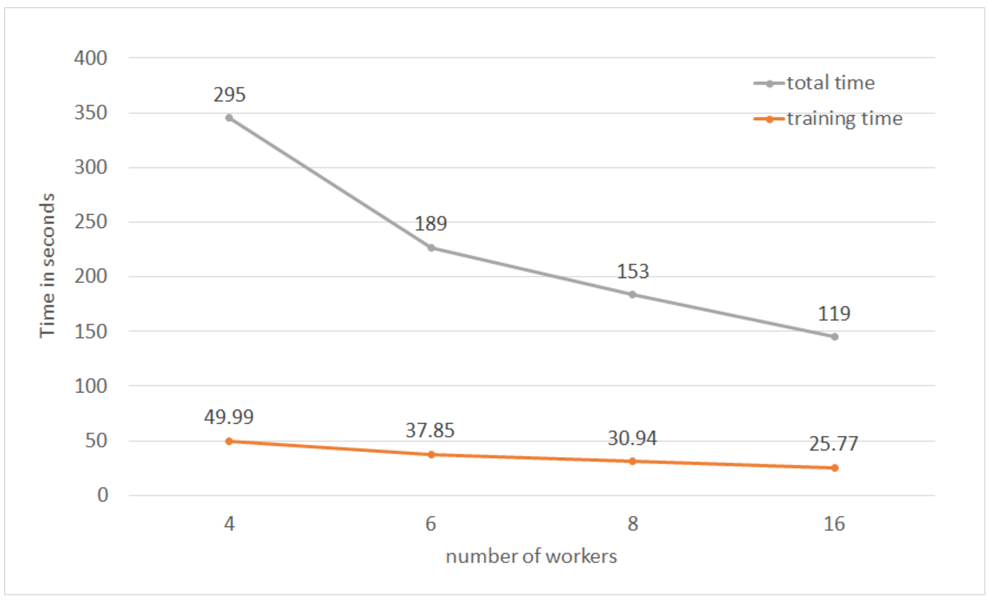

Figure 3. The scalability of logistic regression

- Webspam contains 350000 records and each with 16609143 features. The comparison of convergence time between this projehct and LogisticRegressionWithSGD on Spark [3] for webspam is shown below

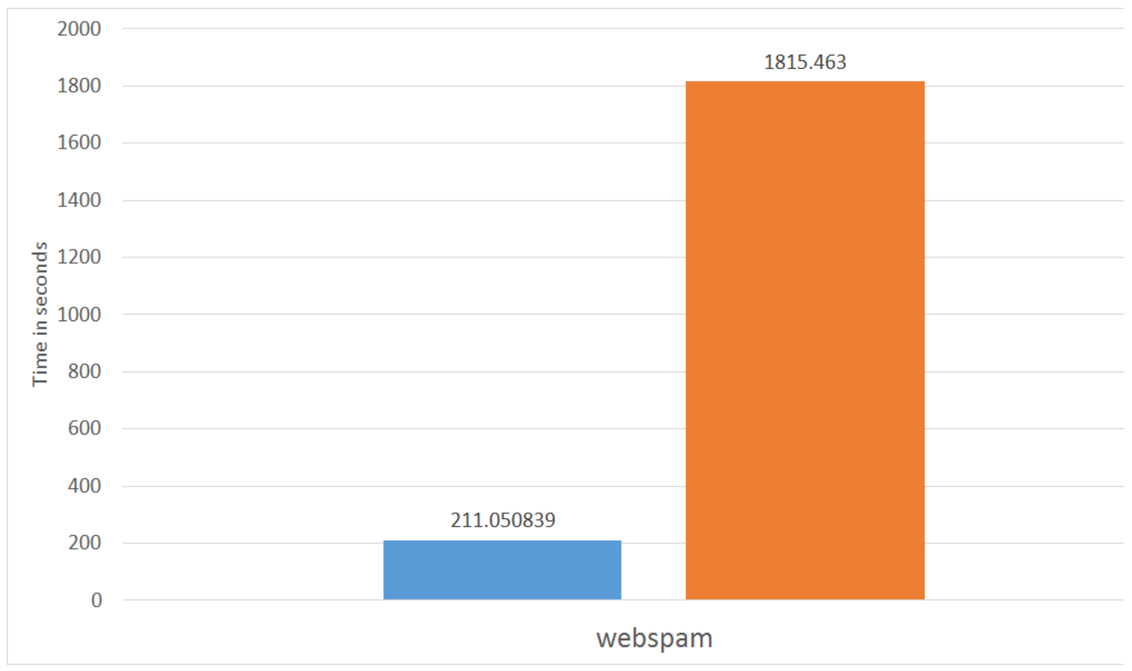

Figure 4. Comparison of convergence time for webspam

Conclusion

- In this project, a distributed logistic regression algorithm is developed on Husky platform. In terms of speed and quality, this implementation can outperform LogisticRegressionWithSGD of MLlib on Spark platform.

Acknowledgements

- I would like to express my special thanks to Fan Yang, Yunjian, Jinfeng, and Kelvin for their generous help. I would also like to thank professor James Cheng for giving me such great opportunity to do the wonderful project on machine learning.

References

[1] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.

[2] F. Yang, J. Li, and J. Cheng. Husky: Towards a more efficient and expressive distributed computing framework. *PVLDB*, 9(5):420–431, 2016.

[3] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *NSDI*, pages 15–28, 2012.