DSGD for MF on Spark 10605 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Question 1

While both algorithms could achieve the goal of Matrix Factorization, we realized that DSGD is much more scalable than SVD. With DSGD, we could generate stratum with several independent blocks. The computation of each block is independent, which makes the algorithm very scalable to run on several cores or machines in an elegant parallel way.

2 Question 2

2.1 Question2.1

Matrix Factorization could be seen as a kind of "topic model". Basically, we could see that the factor number is the chosen topic number. For the W matrix, we could see it as the importance of doc over topics. For the H matrix, we could see it as the importance of topic over words.

2.2 Question2.2

As we mentioned above, the H matrix could be explained as the importance of topic over words. Therefore, if we want to find the top k words in a topic, we could use the H matrix to do it. We could set the ith row of a H matrix, and then find the top k words in the ith row. These top k words is the top k words of topic i.

3 Question 3

I think holding on all other conditions, the Spark implementation should be faster than the Hadoop implementation.

First of all, Spark is lazy evaluation, which uses the DAG and wait to execute after accumulating series of transformations.

Besides, the MapReduce framework is not suitable for DSGD-MF. Bascially, we know that MapReduce is based on stream and sort, which is not suitable for this iterative steps.

4 Question 4

4.1 Question 4.1

The trend is:

We could see at first, with the increase of iterations, the reconstruction error would decrease. Then it would slowly increase with more iterations.

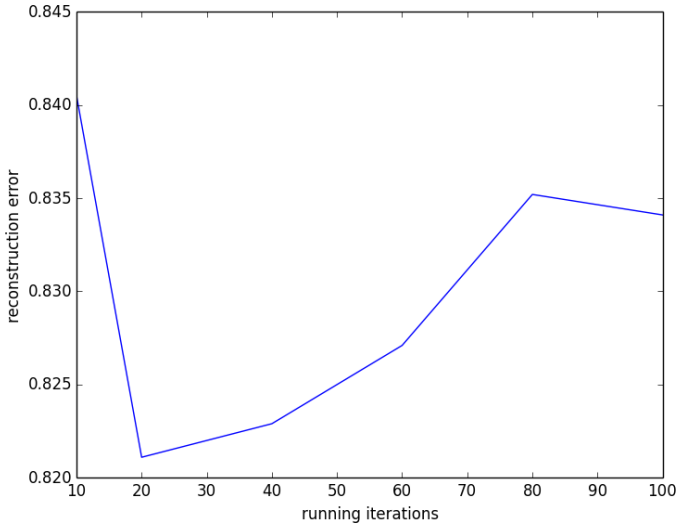

Figure 1: Error over iteration

4.2 Question 4.2

The trend is:

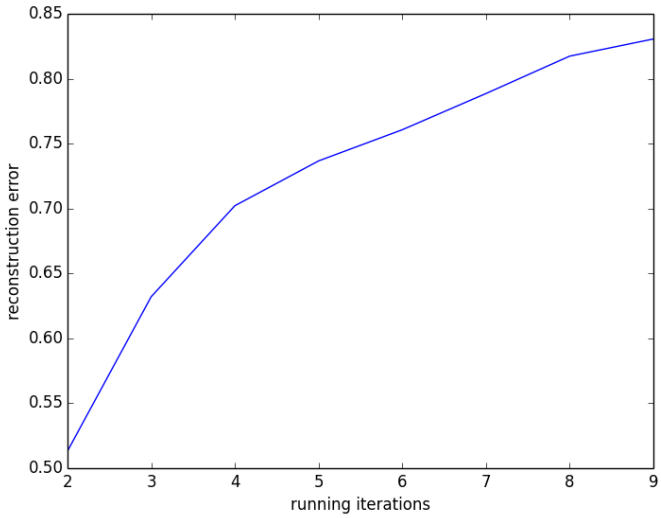

Figure 2: Error over worker

We could see that as the increase of workers, the error would increase. This is quite reasonable because with increase of workers, the error of approximation like number of iterations would be more influential.

4.3 Question 4.3

The trend is:

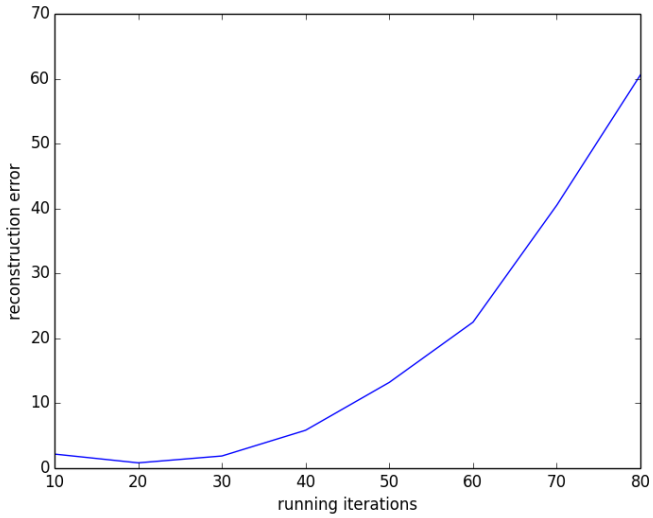

Figure 3: Error over factors

We could see at first, with the increase of topics or factors, the error would decrease. Then with the increase of factors after 20, the error would increase very fast. This is because we set too many topics to the dataset.

4.4 Question 4.4

The trend is:

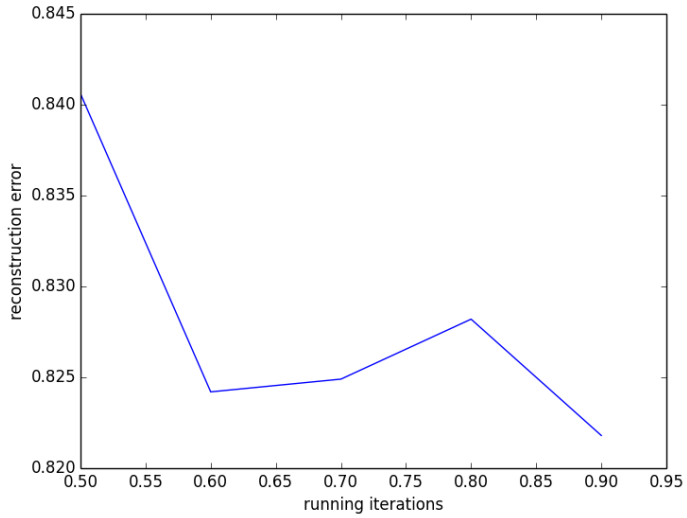

Figure 4: Error over beta

We could see the trend is at first decrease, and then is stable, not sensitive to the increase of beta

5 Question 5

I received helps from Chenran Li, whose AndrewId is chenranl. He told me how to design the blockify matrix function and how to understand the perform sgd function.