10-605: Homework 5

Maya Tydykov

April 16, 2015

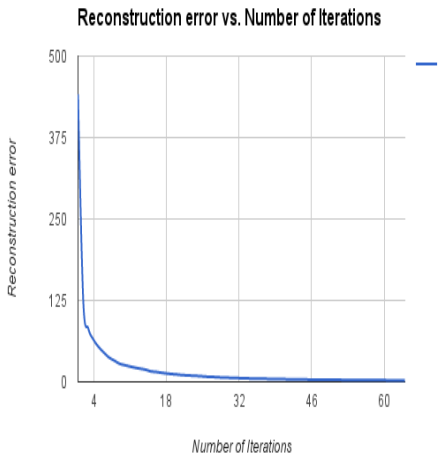
1

2

2.a Question

Set the number of workers B=10, the number of factors F = 20, and $\beta = 0.6$. Plot the reconstruction error $L_N ZSL$ vs the iteration number i = 1,2,...100. Explain the trend in your plot in the space provided below.

Answer

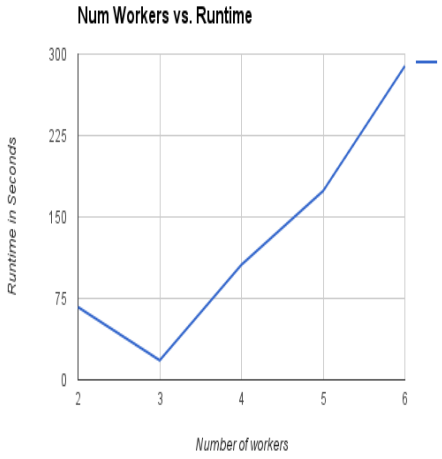


With each iteration the error decreases because of convergence. Note: I was only able to run 64 iterations because after that my code crashed due to a StackOverFlowError.

2.b Question

Set the number of iterations I = 30, the number of factors F = 20, and β = 0.6. Plot the runtime of your Spark code R versus the number of workers B = 2, 3, ..., 10 in steps of 1. Please ensure your local machine or spark cluster can support the number of parallel workers you are requesting. Explain the trend in your plot in the space provided below.

Answer

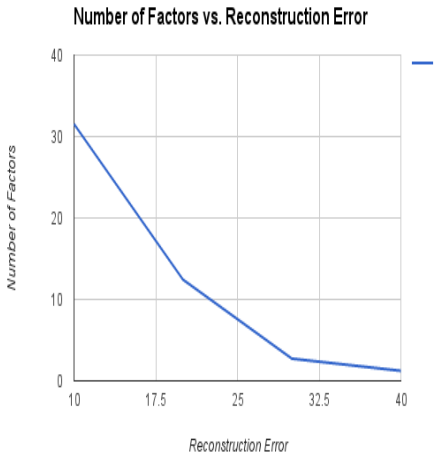


Runtime seems to be best with 3 workers. After that it starts to increase because of communication costs between workers and the master node. Note: I was only able to go up to 6 workers because after that my code crashed due to a StackOverFlowError.

2.c Question

Set the number of iterations I = 30, the number of workers B = 10, and β = 0.6. Plot the reconstruction error $L_N ZSL$ versus the number of factors F = 10, 20...100 in steps of 10. Explain the trend in your plot in the space provided below.

Answer



Reconstruction error tended to get better with a larger number of factors. However, starting with 50, the reconstruction errors were NaN.

3

3.a Question

Set the number of workers B = 10, the number of factors F = 20, and the number of iterations I = 30. Plot the reconstruction error $L_N ZSL$ versus β = 0.5, 0.6...0.9 in steps of 0.1. Explain the trend in your plot in the space provided below.

3.b Answer

4

Question

Is there any advantage to using DSGD for Matrix Factorization instead of Singular Value Decomposition (SVD) which also finds a matrix decomposition that can be used in recommendation systems?

Answer

SVD is harder to compute and parallelize relative to DSGD.

5

Question

Explain clearly and concisely your method (used in the code you have written) for creating strata at the beginning of every iteration of the DSGD-MF algorithm.

Answer

I create my strata as follows. First, I create a dictionary that stores, for each subiteration (over strata within an interation), for each worker id (since I know that worker ids will be from 0 to n, where n is the total number of workers), which movie ids it can have access to. This ensures that different workers will not work on the same range of movie ids in the same stratum. Then, I partition the data based on user ids and the number of workers. Finally, during each iteration over the data, I go through each subiteration, which selects a stratum, and then based on the selected stratum I filter the data in each worker so that only valid movie ids for that worker in that stratum can be updated.

6

Question

If you were to implement two versions of DSGD-MF using MapReduce and Spark, do you think you will find a relative speedup factor between MapReduce and Spark implementations, keeping other parameters like the total number of iterations and number of workers fixed? Which implementation do you think will be faster? Why? If your answer depends on any general optimization tricks related to MapReduce or Spark that you know, please state them as well.

Answer

I think that holding all other things constant, the Spark implementation will be faster. This is because Spark uses lazy evaluation of transformations in which it uses a DAG and waits to execute an accumulated series of transformations, which allows it to plan how it will do these transformations in an optimized manner. The MapReduce framework does not allow this.

7

Match the Spark RDD transformations to their descriptions.
1 coalesce - e
2 repartition - a
3 groupWith - d
4 cache - b
5 foldByKey - c

8 Collaboration

Did you receive any help whatsoever from anyone in solving this assignment? Yes, I discussed both the coding of the algorithm and how to solve the theory questions with David Klaper, Kavya Srinet, and Hakim Sidahmed.
Did you receive any help whatsoever from anyone in solving this assignment? Yes, I discussed both the coding of the algorithm and how to solve the theory questions with David Klaper, Kavya Srinet, and Hakim Sidahmed.