École Polytechnique Fédérale de Lausanne

Report-223876-M3

CS-449 Milestone 3

Azouz Rami

4th June 2021

Q3.2.1

1 QUESTION

Reimplement the Predictor of Milestone 2 using the Breeze library and without using Spark. Test your implementation on the 'ml-100k/u1.base' as a training set, and the 'ml-100k/u1.test' as a test dataset, and k = 200. Ensure your implementation gives a MAE of 0.7485 (within 0.0001). Output the MAE for k = 100 and k = 200 (We will use both for grading). (Make sure to make self-similarities equal to 0, i.e. $S_{u,u} = 0$, prior to computing the k-nearest neighbours.)

2 ANSWER

The obtained MAE are:

- k = 100: 0.7561
- k = 200: 0.7485

Q3.2.2

1 QUESTION

Measure the time for computing all k-nearest neighbours (even if not all are used to make predictions) for all users on the ml-100k/u1.base' dataset. Output the min, max, average, and standard-deviation over 5 runs in your implementation.

2 ANSWER

The computation time for kNN in μs is:

- min = 451143.47
- max = 511509.50
- avg = 472911.93
- std = 20827.39

Q3.2.3

1 QUESTION

Measure the time for computing all 20,000 predictions of the'ml-100k/u1.test'. Output the min, max, average, and standard-deviation over 5 runs in your implementation.

2 Answer

The computation time for computing prediction in μs is:

- min = 6159222.75
- max = 6371626.82
- avg = 6257631.35
- std = 86250.72

Q3.2.4

1 Question

Compare the time for computing all k-nearest neighbours of 'ml-100k/u1.base' to the one you obtained in Milestone 2 (Q.2.2.7). What is the speedup of your new implementation (as a ratio $\frac{t_{old}}{t_{new}}$)? Why do you think that is the case? (Ensure you have obtained at least a 10x decrease in execution time for the k-nearest neighbours and predictions compared to Milestone 2. If you did not compute all k-NN in Milestone 2, use the time your reported for computing similarities.)

2 Answer

The computation time for similarities in the previous milestone is 3.445 * $10^7 \mu s$. In this milestone, we obtained 472911.93 μs. This represent a speedup of 72.84. This speedup might be explained by the fact that when computing the similarity between user u and user v, it is very cheap to get the values in matrix format. However, in the previous implementation, we had to look for common items between the users, get the pre-processed ratings and compute the sum. Now, this is simply done by computing the scalar product of two vectors.

Q4.1.1

1 Question

Test your spark implementation with the ml-1m/ra.train as a training set, and the ml-1m/ra.test as a test dataset, and k = 200. Output the MAE you obtain.

2 Answer

The MAE for the ml-1m dataset with k = 200 is 0.7346

Q4.1.2

1 QUESTION

Manually run your implementation on the cluster 5 times and compute the average. Compare the average kNN and prediction time to your optimized (non-Spark) version of Section 3 on the ml-1m/ra.train dataset. For the Spark version of this section, use a single executor, i.e. –master "local[1]" when using spark-submit. Are they similar? If not, how much overhead does Spark add? Answer both questions in your report.

2 ANSWER

The results are summarized in Table 1 and Table 2. The values are similar in magnitude order. However, we should expect a longer time with the spark implementation since there is additional operations to broadcast the data.

	1	2	3	4	5	Mean
Knn	18.634	18.453	19.183	18.932	19.347	18.910
Prediction	100.425	100.851	116.683	114.928	120.181	110.614

TABLE 1
Computation Time with Spark for knn and prediction in seconds

	1	2	3	4	5	Mean
Knn	23.296	24.738	25.251	28.395	25.086	25.322
Prediction	140.807	138.123	149.898	145.659	134.448	141.787

TABLE 2
Computation Time without Spark for knn and prediction in seconds

Q4.1.3

1 QUESTION

Measure and report kNN and prediction time when using 1, 2, 4, 8, 16 executors on the IC Cluster in a table. Perform each experiment 3 times and report the average, min, and max for both kNN and predictions. Do you observe a speedup? Does this speedup grow linearly with the number of executors, i.e. is the running time X times faster when using X executors compared to using a single executor? Answer both questions in your report.

2 ANSWER

#Workers	1	2	4	8	16
Min	91.139	46.152	26.411	15.209	8.126
Max	98.353	48.649	27.172	15.933	8.653
Mean	94.784	47.548	26.715	15.475	8.427

TABLE 3
Prediction time in s for different # workers

#Workers	1	2	4	8	16
Min	14.072	7.128	4.291	3.111	1.985
Max	19.830	11.049	7.592	6.173	5.368
Mean	16.788	8.869	5.589	4.176	3.217

TABLE 4
Knn time in s for different # workers

The results are summarized in Table 3 and Table 4. The speedup increasing with the number of executors. The gain is linear in the first steps but start to slowdown with more than 8 executors. This decrease can be explained by the fact that every partition is loading the required data, which at some point, worsen overall performance.

Q5.1.1

1 QUESTION

How many days of renting does it take to make buying the ICC.M7 less expensive? How many years does that represent?

2 ANSWER

It would take at most 1716 days of renting the cluster to make it less expensive to buy it, that's roughly 4.7 years.

Q5.1.2

1 QUESTION

What is the daily cost of an IC container with the same amount of RAM and CPUs (1 vCPU = 1 core) as the ICC.M7 machine? What is the ratio of $\frac{ICC.m7(CHF/day)}{Container(CHF/day)}$? Is the container cheaper? Only consider the operating costs, not the buying cost, for the ICC.M7 machine.

2 ANSWER

The daily cost of the ICCM7 is: $20.40CHF/day$

The daily cost of an equivalent container is: $20.896CHF/day$

The rartio gives = $\frac{ICC.m7(CHF/day)}{Container(CHF/day)} = 0.9763$

So the ICCM7 equivalent container is not cheaper than the ICCM7 itself.

Q5.1.3

1 QUESTION

Idem, but compared to (only) the operating costs of 4 Raspberry Pis. Consider 1vCPU = 1 Intel core = 4 RPis in throughput and configure the Container to have the same amount of RAM as the 4 Raspberry Pis combined. What is the ratio $\frac{4RPis(CHF/day)}{Container(CHF/day)}$ at max power for the RPi? at min power for RPi? Is the container cheaper than the 4RPis?

2 ANSWER

The cost of operating the 4 Raspberry Pis is:

At minimum electricity: $4 * 0.0108 = 0.0432 CHF/day$

At maximum electricity: $4 * 0.054 = 0.216 CHF/day$

The daily cost of an equivalent container of the 4 RPis: $0.472 CHF/day$

Taking the ratios of these costs:

For minimum electricty: $\frac{4RPis(CHF/day)}{Container(CHF/day)} = 0.0915$

For maximum electricity: $\frac{4RPis(CHF/day)}{Container(CHF/day)} = 0.4576$

Q5.1.4

1 QUESTION

After how many days of renting a container, is the cost higher than buying and running 4 Raspberry Pis? (1) Assuming optimistically no maintenance at minimum power usage, and (2) no maintenance at maximum power usage, to obtain a likely range. (Round up days)

2 ANSWER

The buying price of the Raspberry Pis doesn't effect the daily costs, but we have to take it into account. So considering the operating days of the Raspberry Pis and their equivalent containers we will need:

885 day minimum so that the cost of running the container becomes higher than buying and running 4 Raspberry Pis considering low energy consumption and 1482 days minimum for maximum energy consumption.

Q5.1.5

1 QUESTION

For the same buying price as an ICC.M7, how many Raspberry Pis can you get (floor the result to remove the decimal)? Assuming perfect scaling, would you obtain a larger overall throughput and RAM from these? If so, by how much (compute the ratios using the previous floored result)?

2 Answer

For the same buying price as the ICC.M7 we can get 369 Raspberry Pis.

That is we'll get a ratio of 3.2946 for throughput of these RPis compared to the ICCM7 and 1.9219 in RAM. So overall we obtain larger throughput and RAM.

Q5.1.6

1 Question

How many users can be held in memory per GB? Per RPi? Per ICC.M7? Assume a sparse matrix representation that required no space for storing indices, only space for storing similarities and ratings.

2 Answer

- users per GB: 555555 users
- users per ICC.M7: 4444444 users
- users per 4 Raspberry Pis: $8.53 * 10^8$ users

Q5.1.7

1 Question

Based on your answers, which would be your preferred option of the three? Would you buy or rent? Why?

2 Answer

We saw that given the price of the ICCM7, getting Raspberry Pis would amount for more computing power. And when comparing with an equivalent container of 4 Raspberry Pis, we have found that the Raspberry Pis are cheaper for the same throughput and RAM. Hence, the most logical solution seems to choose the Raspberry Pis. Thus we would buy 4 Raspberry and invest time in operating them than buy a cluster or rent containers.