Computing at Scale in Machine Learning: Distributed computing and algorithmic approaches (14038)

B-TU Cottbus-Senftenberg · WiSe 2023
Prof. Dr. Alexander Schliep, Nathalie Gocht, M.Sc. and Aleksandra Khatova, M.Sc.

Problem set 10 from January 17, 2024 · Due on January 26, 2024

Problem 1 (Pyspark and summary statistics, 8 pt)**.** : Implement computation of summary statistics—mean, standard deviation, min, max—and a histogram using `pyspark`.
Input: The input is a file containing records ⟨ `id` ⟩ \ `t`⟨ `group`⟩ \ `t` ⟨ `value` ⟩ on each line. Here ⟨`id`⟩ and ⟨`group`⟩ are integer keys, and ⟨`value`⟩ is a scalar real-valued variable. The data can be found in the file `data-assignment-8-1M.dat`.
Output: mean and standard deviation of the values, their minimal and maximal value, as well as the counts necessary for producing a histogram; i.e. for how many records does the value fall into a specific bin. (Note: you do not need to produce a graph displaying the histogram; use 10 bins). Specifically address the following tasks:

a) Implement the pyspark program for the summary statistics and histogram taking input and producing output as described above.

b) Produce a speedup plot for different number of cores using your produced script from 1a.

c) Additionally add computation of the median (or at least an approximation of the median).

Note: Please hand in either:

- A Python script named `problem1.py` and a PDF including the answer to Problem 1.

Python code must be handed in as '.py'. Code submission in PDF will be graded with 0 points. Please submit only the files specified above.

You are allowed to use the example solutions provided where they might be helpful for answering question on this problem set.

Please feel free to discuss the problems and approaches to solving them with other students. However *you* should understand results of discussions well enough to write up the solution by yourself and possibly explain the solution. When you do collaborate, please make other's contributions clear. Present your solutions concisely.