User Guide of `quantreg_hadoop` Package

Jiyan Yang [*]

This document provides a guide for using the `quantreg_hadoop` package, which contains the Hadoop implementation of the large-scale quantile regression algorithm appeared in the paper "Quantile Regression for Large-scale applications" (http://arxiv.org/abs/1305.0087). The algorithm computes an $(1 \pm \epsilon)$-approximation to the given quantile regression problem.

1 Setting up

Please make sure the following steps are done before running the codes.

- The codes are written in Python. To run the codes, one needs to download and install the `dumbo` (https://github.com/klbostee/dumbo/wiki) which is a Python API for writing MapReduce programs conveniently.

- To set the configuration files, copy the following contexts into `.dumborc`.

  ```
  [common]
  hadoop:  HadoopClusterName
  [hadoops]
  HadoopClusterName:  Dir
  ```

  Above, `YourHadoopClusterName` is an alias to your Hadoop cluster. It could be any text as long as it can be used to distinguish clusters. `Dir` is the directory where the Hadoop binary file is located. For example, it can be `/usr/lib/hadoop-0.20/`.

  Note here, changing the configuration file only results in a different command when calling Hadoop program. In order to run the codes, such configuring is necessary. See https://github.com/klbostee/dumbo/wiki/Configuration-files for more details.

2 Using the codes

The `zip` file contains two folders, namely, `src` and `bin`. The main script for running the algorithm is `bin/quantreg.sh`. At the top of the script, a few environment variables needed to be set. Below is an explanation of these variables.

- `DIR` is the variable specifying the absolute directory of the current folder. For example, `DIR="$HOME/quantreg"`.

- `HDFS_DIR` is the directory in HDFS used to store data and results of the experiments.

- `ORDER` is used to denote the order of the current experiment. Results (e.g., relative errors) will be stored locally in folder `$DIR/results/empirical_reuslts$ORDER`.

[*] ICME, Stanford University, Stanford, CA 94305. Email: jiyan@stanford.edu

- The data in plain text format should be stored in folder `$HDFS_DIR/data` in HDFS specified by variable `FILENAME`.
- The options for `COND_METHOD` are: `spc1, spc2, spc3, sc, noco` and `unif`. See the paper for more details.
- The source codes should be placed in folder `$DIR/src`.
- All the outputs of Hadoop will be stored in folder `$HDFS_DIR/$COND_METHOD`.
- The script will compute the relative errors. The optimal solutions and objective values should be provided in `$DIR/data` with the names `$FILENAME_x_opt.txt` and `$FILENAME_f_opt.txt`
- Number of reducers to be used, sampling size and number of independent trials for sampling can be specified in `NUM_REDUCER`, `SAMPLING_SIZE` and `NX`.

Note here, by default, in each experiment, after sampling, the solver will solve the reduced quantile regression problem for three different values of τ, namely, $0.5, 0.75$ and 0.95. This means for a fixed setting of parameters, the algorithm returns approximate solutions to the original quantile regression problem associated with $\tau = 0.5, 0.75, 0.95$, respectively. One can change such setting in the construction function of the `Solver_Reducer` class in the code `quantreg_samp_solve.py` located in the `src` folder.

3 Outputs

The outputs of each experiment will be stored in a folder in `results` specified by the `ORDER` variable in the `quantreg.sh` script. The outputs include the following.

- Basic information about the experiment, i.e., `info.txt` and `prog.log`.
- Binary files fetched back from HDFS, e.g., folders `PA, L`.
- The first and third quartiles of the relative errors on the objective, i.e., `quar_obj.txt` and solution vector, e.g., `quar_sol_l1.txt` of the original quantile regression problem for all the τ values among `NX` trials. For the latter, they are measured in three different norms, namely, ℓ_1, ℓ_2 and infinity norms.