Stat243: Problem Set 6, Due Monday November 8

October 30, 2021

This covers Units 8 and 9.
It's due **as PDF submitted to Gradescope** and submitted via GitHub at 10 am on Nov. 8.
Comments:

1. The formatting requirements are the same as previous problem sets.
2. As with PS5, don't wait to get started. Problem 2involves use of the SCF cluster.
3. Please note my comments in the syllabus about when to ask for help and about working together. In particular, **please give the names of any other students that you worked with on the problem set and indicate in comments any ideas or code you borrowed from another student.**

Problems

1. Using the Stack Overflow database, write SQL code that will determine how many users have asked both R- and Python-related questions (not necessarily, but possibly, about R and Python in the same question). There are various ways to do this, of which we've only covered some approaches in the class material. Those of you with more experience with SQL might do this in a single query (for which there are various options), but it's perfectly fine to create one or more views and then use those views to get the result as a subsequent query. (Hint: if you do this using joins, you'll probably need to do the equivalent of a self join at some point.) Finally extend your query to include only users who have asked questions about R (that are not also about Python) and questions about Python (that are not also about R).
2. Consider the full Wikipedia traffic data for October-December 2008 (available in */var/local/s243/wikistats/dated_2017* on any of the low (default) partition SCF cluster nodes).
   (a) Explore the variation over time in the number of visits to Barack Obama-related Wikipedia sites, based on searching for "Barack_Obama" on English language Wikipedia pages. You should use Dask to do the reading and filtering. Then group by day-hour (it's fine to do the grouping/counting in Python in a way that doesn't use Dask data structures). You can do this either in an interactive session using *srun* or a batch job using *sbatch*. And if you use *srun*, you can run Python itself either interactively or as a background job. Time how long it takes to read the data and do the filtering to get a sense for how much time is involved working with this much data. Once you have done the filtering and gotten the counts for each day-hour, you can simply use standard R or Python code on your laptop to do some plotting to show how the traffic varied over the days of the full October-December time period and particularly over the hours of November 3-5, 2008 (election day was November 4 and Obama's victory was declared at 11 pm Eastern time on November 4).

(b) Extra credit: Carry out some analyses of how the traffic varies by language or do some other in-depth analysis of the Wikipedia data (it doesn't have to involve Barack Obama), addressing a question of interest to you.

Notes:

- Note that I'm not expecting you to know any more Python than we covered in the Unit 7/8 material on Dask and in Section, so feel free to ask for help (and for those of you who know Python to help out) on Python syntax on Piazza or in office hours.
- There are various ways to do this using Dask bags or Dask data frames, but I think the easiest in terms of using code that you've seen in Unit 8 is to read the data in and do the filtering using a Dask bag and then convert the Dask bag to a Dask dataframe to do the grouping and summarization. Alternatively you should be able to use *foldby()* from *dask.bag*, but figuring out what arguments to pass to *foldby()* is a bit involved.
- Make sure to test your code on a portion of the data before doing computation on the full dataset. **Reading and filtering the whole dataset will take something like 60 minutes with 16 cores. You MUST test on a small number of files on your laptop or on one of the stand-alone SCF machines (e.g., radagast, gandalf, arwen) before trying to run the code on the full 120 GB (zipped) of data.** For testing, the files are also available in */scratch/users/paciorek/wikistats/dated_2017*.
- When doing the full computation via your Slurm job submission:
  - You must read the data from */var/local/s243/wikistats/dated_2017* (which is available on all the machines on the SCF cluster, so you don't need to copy it, unlike in PS5).
  - Please do not use more than 16 cores in your Slurm job submissions so that cores are available for your classmates. If your job is stuck in the queue you may want to run it with 8 rather than 16 cores.
  - As discussed, when you use *sbatch* to submit a job to the SCF cluster or *srun* to run interactively, you should be using the *–cpus-per-task* flag to specify the number of cores that your computation will use. In your Python code, you can then either hard-code that same number of cores as the number of workers or (better) you can use the SLURM_CPUS_PER_TASK shell environment variable to tell Dask how many workers to start.

3. Consider a simulation study that generates $m = 100$ simulated datasets. The parameter of interest is θ and in simulating the datasets, we have $\theta = 2$. The value of 2 is included in 85 of the 95% confidence intervals.

   (a) If you're interested in the coverage of the confidence interval procedure, what is $h(Y_i)$ in this setting? What is the expectation that is of interest here?

   (b) Based on the Monte Carlo uncertainty of the expectation of interest, do you think you have simulated enough datasets? Note that this is a somewhat subjective judgment.