Stat 243 Exam Practice Questions

November 3, 2017

1. You have a data file that contains an n by n matrix on a remote server that you need to copy to your computer and load into R. Will you have any problems doing this? At what steps might things go wrong?

2. Suppose you want to store numbers rounded to one decimal place, like 13.4, 112.3. If you want to save on disk space, are you better off storing in ASCII or in binary using the usual double precision floating point representation? What factors would influence your decision? Ignore the effects of any compression.

3. Your friend proudly says he saved values out from R into a CSV and wrote out 40 digits per number so he would have higher accuracy for subsequent computation. What do you tell him?

4. What does this code do?

```
for file in $(ls *.csv)
do
grep 'pdf' $file >> tmp.txt
done
```

   What would happen if I changed the >> to >?

5. What is the purpose of namespaces?

   (a) Alternatively, why is it that we have to load R packages with *library()* rather than just having all the functions of all packages installed on our machine available to us automatically?

   (b) Alternatively, consider that the *lm()* function in the stats package uses the function *lm.fit()*. Where does the *lm()* function look for *lm.fit()* and how does this relate to how R handles scoping?

6. If we want to estimate a derivative of a function on a computer (often because it is hard to calculate the derivative analytically), a standard way to do this is to compute:

$$f'(x) \approx \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

   for some small ϵ. Since the limit of the expression as $\epsilon \to 0$ is exactly $f'(x)$ by the definition of the derivative, we presumably want to use ϵ very small. If we try to do this on a computer, how do the limitations of arithmetic on a computer affect our choice of ϵ?

7. When we discussed the Householder reflection approach to the QR decomposition, I said that a key part of the calculation is $\tilde{x} = Qx$ where x is $n \times 1$ and Q is $n \times n$. Naively this calculation would involve n^2 multiplications. Given that we know that $Q = I - 2uu^\top$ for a vector u that is $n \times 1$, what is the computational cost of the more efficient approach to that computation?

8. If I want to compute the trace of a matrix, $A = XY$, where the trace is $\sum_{i=1}^{n} A_{ii}$, a naive implementation is `sum(diag(X%*%Y))`. How could I (much) more efficiently compute the trace in R?

9. Explain what an index is in the context of a database and how it can improve the speed of an SQL query.

10. Consider this function:

```
f <- function(x = {y <- 1; 2}, y = 0) {
  x + y
}
f()
```

Explain fully why the result is 3, describing how the values for x and y are determined.

11. What does this SQL query do? Note that it is a "self-join" in that it joins a table to itself.

```
select displayname, userid from
(questions Q1 join questions_tags T1 on Q1.questionid = T1.questionid)
join
(questions Q2 join questions_tags T2 on Q2.questionid = T2.questionid)
on Q1.ownerid = Q2.ownerid
join users on Q1.ownerid = users.userid
where T1.tag = 'r' and T2.tag = 'python'
```

Hint: it may be helpful to think of the two chunks of code in parentheses as simply creating two views (temporary tables).

12. Suppose I am fitting eight different statistical models to a dataset and using bootstrapping to estimate uncertainty. For each model I fit $m = 500$ bootstrap samples (random samples with replacement from the n observations in the dataset). Assume that the fitting of the statistical model involves computation with large matrices.

    (a) What are the parts of the computation I could parallelize?

    (b) If I have a single machine with 20 cores, how might I parallelize this?

    (c) What if I have a single machine with 4 cores?

13. What will this R code do: `apply(X, 1, '[', 1:3)`? Why does it illustrate that R is a functional language (there are multiple aspects of this that you could comment on)?