HSD5230 High performance computing Midterm Exam

Miao Cai

3/15/2019

1) What is the definition of big data? Although there are many possible answers here, I have communicated what I consider the right answer in class multiple times!
2) What is the difference between inferential modeling approaches versus predictive modeling approaches? Come up with a clarifying example for each modeling approach (one inferential, one predictive).
3) Is machine learning superior to traditional approaches to statistical inference? When would you use machine learning versus traditional approaches to statistics? Give specific reasons to support your answer.
4) Define the term 'data leakage' in the context of health care analytics (predictive models specifically). What are the symptoms of data leakage?
5) Explain 2 strengths of real world evidence compared with randomized clinical trials.
6) Explain 2 strengths of randomized clinical trials compared to real world evidence.
7) The data.table package in R allows us to work on larger datasets than using base R or dplyr. Why/how does data.table allow us to work on larger data? Isn't data.table still limited to memory?
8) What does is the difference between timing your code and profiling your code?
9) When should we start profiling (or timing) our code?
10) Define 'embarrassingly parallel' calculations.
11) Consider the following example: I have longitudinal medication data on 1 million patients in a healthcare system. I wish to calculate each patient's medication adherence for every year (a simple summary of how good each patient is at taking their medication). Is this an example of an 'embarrassingly parallel' calculation? Why or why not?
12) Consider the following example: I have longitudinal medication adherence measures on 1 million patients in a healthcare system. I wish to fit a logistic regression to see if medication adherence predicts mortality among these patients. Is this an example of an 'embarrassingly parallel' calculation? Why or why not?
13) HDF5 data stores are one method to storing data on disk when using Pandas. Name 3 advantages HDF5 data stores have over structured text (like CSV) files when using Pandas:
14) What does 'lazy evaluation' mean? Why is that advantageous?
15) Do the following frameworks generally use lazy evaluation in most of their functionality (yes or no)?

a. Base R
b. R: Data.table
c. Python: Pandas
d. Python: Dask

16) What is the smallest number of bits (not bytes!) you can use to represent the following numbers (presume you don't need negative numbers, these are unsigned)?

a. 1
b. 5
c. 55
d. 243
e. 290
f. 1025

g. 67000

17) What is the smallest number of bytes you can use to represent the following numbers (presume you don't need negative numbers, these are unsigned)?

a. 15
b. 265
c. 67000
d. 4294967298

18) Consider a dataset where I have hospital ID's that range from 1 to 200. If I am using Pandas and I wish to manually compress the data by downcasting the numeric data types, can I use int8 to represent this data? Why or why not?

19) Consider a dataset where I have hospital ID's that range from 1 to 200. If I am using Pandas and I wish to manually compress the data by downcasting the numeric data types, can I use uint8 to represent this data? Why or why not?

20) What are the two primary reasons Spark is faster than Hadoop?

21) Hadoop uses the Hadoop distributed file system (HDFS) to solve the distributed data problem, and uses MapReduce to solve the distributed analytic problem. Which of these two aspects of Hadoop does Spark replace/replicate?

22) Define vertical versus horizontal resource scaling in high performance computing.

23) Name 3 differences between SQL and NoSQL approaches to data storage.