Proposal for Big Data Course

Clark Fitzgerald

Introduction

These are working notes to develop a statistics / data science course focused on big data. Let's call it STAT 129, because STAT 128 is a prerequisite.

Broadly, the goal of this course is to teach students to apply statistics to challenging, real world data sets. For me, the goal of a certificate in data science is for the students to become increasingly independent in how they're able to compute and analyze data. This course is a step in that direction; it should provide students with the confidence to work with nearly *any* kind of data they encounter.

Long term, STAT 129 can count as an elective in the following programs:

1. math BA undergraduate major for the emphasis in Applied Math or the emphasis in Statistics
2. the department's data science certificate
3. the (currently on hold) math and applied computing undergraduate major that John Ingram is working on

Feedback

Lisa Taylor asked, 'how will we distinguish this as a math/stats course?' That is, why is this course in math/stats rather than computer science?

This course belongs in Statistics because the purpose of everything we do in this course is to answer data analysis questions. To answer these questions for real data sets we have to learn about the technologies that allow us to process the data. We write our programs solely to answer these questions, rather than developing more general purpose software. The focus of this course is *not* implementation; we build on existing lower level software implementations whenever possible, rather than reimplementing common tasks. Thus, the technologies are a means to an end.

The following are examples of high level data analysis questions:

- Which customers are likely to return their products?
- Which of these grant awards are unusual?
- What are the trends in political party affiliation by location?

Syllabus

Course Description

Tools and techniques for statistical analysis of large, complex data sets. Application of statistical techniques suitable for big data, for example, dimension reduction, clustering algorithms, and text mining. Students will access data and run code on remote servers. High level parallel computing. Technologies covered may include Python, Structured Query Language (SQL), and bash programming languages. 3 units.

Learning Objectives

Upon completion of this course, students will be able to:

- Develop complete statistical computer programs based on high level directions, using standard software packages. Their programs will be complete in the sense that they start with processing raw data, and finish by producing final summaries and results necessary for reports.
- Summarize their approach and conclusions for a data analysis problem through technical written reports with appropriate graphics.
- Apply standard statistical techniques suitable for big data, for example, dimension reduction, clustering algorithms, and text mining.
- Identify and extract elements of interest from complex data sets, including tabular, hierarchical, streaming, and text data.
- Run programs on large data sets located on remote machines, which may include databases, remote compute clusters, and cloud services.
- Implement basic data parallel programs.

Prerequisites

Stat 128 or consent of the instructor. Students should be comfortable with computer programming.

Evaluation

- 60% Assignments
- 15% Midterm
- 5% Participation
- 20% Final Project

4 to 8 assignments over the course of the semester will pose challenging data analysis problems on real data sets. Some assignments will feature 'dirty' data: missing, noisy, and possibly erroneous. This will require students to make judgement calls about when and where to apply various statistical techniques, such as imputation.

Topics

STAT 129 will primarily use the Python programming language for instruction in class, but will also include content on Structured Query Language (SQL) and bash. Students are welcome to do some assignments in any programming language they are comfortable with.

Comments:

CF: STAT 128 focuses on R, so any student who takes both STAT 128 and 129 will have at least been exposed to the 4 most common languages in data science. This separation of languages also makes STAT 128 less of a 'hard' prerequisite; we'll teach Python from the beginning, assuming only that the student has learned to program in *any* language, rather than assume they know R.

MN: I like the idea of making Stat 128 a 'soft' prereq. Some CS students in particular may wish to take Stat 129 without the Stat 128 prereq.

topic	example technologies	description
streaming data	Python, bash	Students will learn to process streams of data using iterators and the UNIX pipe model.
remote computers	bash, SSH	Remote computers are necessary to work with data that's too large for a laptop. Students will interactively login and submit batch jobs.
databases	SQL	Relational databases are the ubiqutious standard for data storage. Students will write SQL queries to filter and join tables in a database.
parallel programming	multiprocessing, MapReduce	Students will learn to recognize and program the common case of applying the same function to many data elements, and collect the results.