Stat 196K (Stat 129): Analyzing and Processing Big Data

California State University, Sacramento · Department of Mathematics & Statistics

The goal of this course is to teach students to apply statistics to challenging, real world data sets. Students should come away from this course with the confidence to work with nearly *any* kind of data they encounter.

CATALOG DESCRIPTION

Statistical analysis of large, complex data sets. Topics include memory efficient data processing, the split-apply-combine strategy, rewriting programs for scalability, handling complex data formats, and applications such as statistical learning, dimension reduction, and efficient data representation. Students will access data and run code on remote servers. **3.0 Units. Letter Graded**.

PREREQUISITES

(STAT 1 or STAT 50) and (MATH 26A or MATH 30) and (STAT 128 or CSC 15), or consent of the instructor.

LEARNING OUTCOMES

Students will be able to:

1. 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.
2. Summarize their approach and conclusions for a data analysis problem through technical written reports with appropriate graphics.
3. Apply standard statistical techniques suitable for data larger than memory, for example, the split-apply-combine strategy for grouped data, memory efficient streaming statistics, discretization, and dimension reduction through principal components analysis.
4. Identify, extract, and summarize elements of interest from complex data sets, including tabular, hierarchical, streaming, image, and text data.
5. Perform data analysis using remote machines, which may include databases, remote compute clusters, and cloud services.
6. Accelerate and scale data analysis programs by identifying and fixing performance bottlenecks.

SAMPLE TEXT AND MATERIALS

- Computer with internet access and administrative rights to install open source software.

METHODS OF EVALUATION

There will be midterm examination(s), a final project, and a comprehensive final examination for this course. Four to eight 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.

TIMELINE

I. Memory Efficient Data Processing (5 weeks)
   A. Streaming statistics and single pass algorithms
   B. Binning and discretization
   C. Pipelines for processing data elements in sequence
   D. Processing large data in chunks
   E. Split-apply-combine strategy for grouped data

II. Writing programs that scale to the analysis of large data sets (3 weeks)
   A. Profiling and measurements to diagnose performance problems
   B. Improving performance through parallel programming
   C. Client-server model
   D. Querying remote data sources
   E. Avoiding data movement by running code remotely

III. Analyzing Complex Data (2 weeks)
   A. Unexpected data and error handling
   B. Text processing and regular expressions
   C. Hierarchical and nested data

IV. Applications of statistical methods for big data (5 weeks)
   A. Image processing
   B. Neural networks
   C. Graph representation through sparse matrices
   D. Dimension reduction, principal components analysis (PCA)
   E. Other data analysis techniques, TBD by class interest