CS 350 Computational Tools and Techniques for Data Analysis

Course Objectives

To familiarize students with a range of tools and techniques appropriate for analyzing and visualizing large amounts of experimental data.

Expected Outcomes

1. Students will be able use high level tools to visualize data to gain insight into data relationships.

2. Students will become familiar with high-level scripting languages and system utilities suitable for manipulating large data sets.

3. Students will be able to construct basic models of data for analysis.

4. Students will be introduced to basic pattern recognition and knowledge discovery techniques to discover relationships in data.

Catalog Description

Introduction to the representation, visualization, and modeling of large data sets. Data analysis using standard high level software tools. Topics include data filtering, clustering, classification, and data mining.

Course Topics

1. (1 week) An overview of the fundamental problems related to data acquisition and data analysis.

2. (2 weeks) Storage, manipulation, and visualization of large numeric data sets using spreadsheet utilities.

3. (1 week) Techniques for data normalization, handling missing or corrupted data values, and filtering large data sets. Introduction of software tools (e.g. grep, sed, cut) and scripting languages (e.g. awk, sed, Perl and Python) for data manipulation and filtering.

4. (1.5 weeks) Introduction to tools for constructing models from data and system simulation (eg. R, Matlab, and Mathematica).

5. (1.5 weeks) Creating visual representations of large data sets with an introduction to software tools for visualization.

6. (1 week) An overview of pattern recognition algorithms and techniques for analyzing data including clustering algorithm and data classification systems.

7. (1 week) Overview of data mining techniques for finding correlations or patterns in large data sets including machine intelligence techniques including artificial neural networks, evolutionary algorithms, and fuzzy sets. Introduction to software tools and packages to support data mining (e.g. matlab toolboxes, Weka)

Requirements:

Homework assignments to assess students' understanding of principles and concepts of data representation.

Assignments using the software tools for data analysis and visualization.

Midterm and final examination

Graduate Student Assignments:

The exams, homework, and reading assignments for graduate students will cover additional theoretical material not required of the undergraduate students.

Textbooks

Advanced Excel for scientific data analysis, R. D. Levie, Oxford University Press, New York, 2004.

Information Visualization - Beyond the Horizon, by Chaomei Chen, C, Springer, 2004.

Pattern Classification (2nd ed.) by Richard O. Duda, Peter E. Hart and David G. Stork Wiley Interscience, 2001, ISBN: 0-471-05669-3

CS ~~350~~/550 Computational Tools and Techniques for Data Analysis

Syllabus

This course provides an introduction to techniques used for representation, manipulation, and analysis of large datasets. Computational data analysis methods include visualization, data clustering, modeling, and data mining. This course is taught through a user's perspective. The coverage of each topic will include a presentation of the basic principles behind technique and experiments that demonstrate the potential applications of the technique. The experiments will utilize standard software packages and programming languages.

8. (1 week) An overview of the fundamental problems related to data acquisition and data analysis.

9. (2 weeks) Storage, manipulation, and visualization of large numeric data sets using spreadsheet utilities.

10. (1 week) Techniques for data normalization, handling missing or corrupted data values, and filtering large data sets. Introduction of software tools (e.g. grep, sed, cut) and scripting languages (e.g. awk, sed, Perl and Python) for data manipulation and filtering.

11. (1.5 weeks) Introduction to tools for constructing models from data and system simulation (eg. R, Matlab, and Mathematica).

12. (1.5 weeks) Creating visual representations of large data sets with an introduction to software tools for visualization.

13. (1 week) An overview of pattern recognition algorithms and techniques for analyzing data including clustering algorithm and data classification systems.

14. (1 week) Overview of data mining techniques for finding correlations or patterns in large data sets including machine intelligence techniques including artificial neural networks, evolutionary algorithms, and fuzzy sets. Introduction to software tools and packages to support data mining (e.g. matlab toolboxes, Weka)

Graduate Student Assignments:

The exams, homework, and reading assignments for graduate students will cover additional theoretical material not required of the undergraduate students.

Textbooks

Advanced Excel for scientific data analysis, R. D. Levie, Oxford University Press, New York, 2004.

Information Visualization - Beyond the Horizon, by Chaomei Chen, C, Springer, 2004.

Pattern Classification (2nd ed.) by Richard O. Duda, Peter E. Hart and David G. Stork Wiley Interscience, 2001, ISBN: 0-471-05669-3