Automated Learning and Data Analysis: CSC 422 / 522

Syllabus
Spring 2013

Jon Doyle
Department of Computer Science
North Carolina State University

Instructor:

Jon Doyle
Office: EB II, Room 2298
Telephone: 919-513-0423
Email: jon_doyle@ncsu.edu
Web: http://www.csc.ncsu.edu/faculty/doyle
Office hours: see the instructor's website or office nameplate

Teaching assistants:

Andrew Beam: albeam@ncsu.edu
Lakshmi Ramachandran: lramach@ncsu.edu
Srinath Ravindran: sravind2@ncsu.edu

Time and place:

Tuesday-Thursday 9:35–10:40 AM
Engineering Building I, Room 1007
URL: http://moodle.wolfware.ncsu.edu/

Required text: (Copies available at the NCSU Bookstore)

Introduction to Data Mining
by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
Reading, MA: Addison-Wesley, 2005

Summary

This course provides an introduction to concepts and methods for extracting knowledge or other useful forms of information from data. This activity, also known under names including data mining, knowledge discovery in databases, and exploratory data analysis, plays an important role in modern science, engineering, medicine, business, and government.

Students will learn basic properties of several common types of knowledge implicit in data, along with formal representations of these types of knowledge and methods for identifying knowledge of these types contained in specific data sets. Students will also learn about the overall process of data collection and analysis that provides the setting for knowledge discovery, and concomitant issues of privacy and security. Examples and projects introduce the student to application areas including electronic commerce, information security, biology, and medicine.

Prerequisites

Students will find introductions to artificial intelligence and database management very helpful, but these are not required. The key prerequisites consist of basic knowledge of

Logic	CSC 226, LOG 201, or equivalent
Probability and statistics	ST 370 or equivalent
Linear algebra	MA 305 or equivalent

Course Outcomes

The aim of this class is to introduce the student to concepts and methods of large-scale automated data analysis. Upon completion, the student will be able to

- List and explain the major types of data and data representations;
- List and explain the problems arising in preparing data for analysis, and the methods for addressing these problems;
- List and explain representative applications of automated learning and data analysis;
- List and explain representative benefits and dangers of automated learning and data analysis;
- Identify some ethical issues in data analysis applications;
- Explain the iterative process of formulating knowledge;
- List and explain the fundamental properties of formulations of knowledge and their use in evaluating and criticizing formulations;
- List and explain some principal representations of knowledge, and compare their strengths and weakness for different representational tasks;
- List, explain, and apply the major data analysis techniques;

- Compare the strengths, weaknesses, and prerequisites of automated learning techniques;
- Design a detailed plan of analysis for a realistic data set;
- Identify contingencies occuring in a data analysis;
- Apply automated data analysis tools to carry out a data analysis plan; and
- Motivate, justify, and qualify conclusions obtained from an analysis.

Organization

The coursework consists of lectures, readings, homework assignments, tests, and a term project.

- Some of the lectures will depart from the text, either in content or in order. Some material will be covered only in lecture; other material will be covered only in assigned readings. Tests will include material from lecture and readings. Students are responsible for all material presented or discussed in lecture.
- Readings will generally be taken from the text by from Tan, Steinbach, and Kumar, with possible supplements from the literature.
- The examinations will consist of a midterm and a final exam. Statements of test objectives will be provided prior to the examinations to indicate their scope. The midterm will cover roughly the first half of the course content. The final exam will cover the entire course, but with an emphasis on the last half of the course content.
- Each student must complete an extended data analysis project. The expectation is that the project will consist of performing and reporting on a task from the Kaggle.com data analysis competition website. The task for analysis will be announced mid-March, and is expected to be either a current competition, in which case the competition sponsors offer monetary rewards, or a past competition, in which case the winners of the class competitors will be rewarded nonmonetarily with recognition and possibly extra points on grades. Small collaborations (2-3 team members) will be allowed on the analysis tasks. Each student must submit a written project report on their analysis. Guidelines for the writing of this report are presented in a separate document.

  Teams should not talk with other teams about their work, as the analysis should be done by each team alone. The project reports, in turn, should be written without collaboration between team members. Unwarranted similarity in individual written reports will constitute a serious breach of academically with severe penalties.

  Alternatively, individuals can seek permission to do individual projects that analyze data sets of special interest to the student. Students interested in such projects should, as early as possible in the semester, submit a brief proposal to the instructor describing the problem, data source and character, and results sought. Individual project reports are required in this case as well.

- Each student enrolled in CSC 522 must submit an additional report consisting of a literature summary and review. Guidelines for the writing of this additional report are presented in a separate document.

Computation

Project analyses will require the student to use data analysis software such as R, Matlab, SAS Enterprise Miner, or Weka, but the course will not teach the use of software packages in detail. Some of these systems are available through VCL (vcl.ncsu.edu).

Resources

Supplementary texts:

- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
  New York: Springer-Verlag, 2001. Available free online.
- *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, second edition, by Ian H. Witten and Eibe Frank
  San Francisco: Morgan Kaufmann Publishers, 2005.
- *Data Mining: Concepts and Techniques*, second edition, by Jiawei Han and Micheline Kamber
  San Francisco: Morgan Kaufmann Publishers, 2006
- *Machine Learning*, by Tom Mitchell
  New York: McGraw Hill, 1997, ISBN 0070428077.
- *Artificial Intelligence: A Modern Approach*, second edition, by Stuart J. Russell and Peter Norvig
  New York: Prentice Hall, 2003
- *Principles of Data Mining*, by David Hand, Heikki Mannila, and Padhraic Smyth
  Cambridge: MIT Press, 2001.
- *The Little SAS Book: a primer*, by Lora D. Delwiche and Susan J. Slaughter
  Second edition, Cary: SAS Institute Press, 1998.

Journals:

- Data Mining and Knowledge Discovery
  `http://www.kluweronline.com/issn/1384-5810`
- Journal of Machine Learning Research
  `http://www.jmlr.org/`
- Several IEEE Transaction journals

Web resources:

- UCI Knowledge Discovery in Databases Archive
  `http://kdd.ics.uci.edu/`
- KDnuggets: Data Mining, Web Mining & Knowledge Discovery News, Consulting and Recruiting
  `http://www.kdnuggets.com/`

Privacy

Do *not* include student ID numbers on papers or tests unless specifically instructed to do so by the instructor.

Grading

Clarity of writing and organization forms an important factor in grading of homework, examinations, and reports. Unclear writing can suggest a lack of understanding of the material. Examples, figures, tables, and analytical results should be accompanied by clear explanations of what lessons the reader should take away from them. Students should avoid writing in terms of bullet lists, which usually reflect a lack of thought about how to express the material. Make sure the paragraphs of the writing clearly express the main points, supporting arguments, and connections between the main points.

Homework and reports must be submitted via Moodle prior to the due date and time announced in advance. Late submissions will not be accepted apart from absences excused according to the University attendance policy. Incomplete grades will not be assigned except in cases of absences excused according to the University attendance policy.

Grades will be assigned based on a weighted combination of performance on different course activities, according to the scheme

A+	=	97 - 100	C+	=	77 - 79.9
A	=	93 - 96.9	C	=	73 - 76.9
A-	=	90 - 92.9	C-	=	70 - 72.9
B+	=	87 - 89.9	D+	=	67 - 69.9
B	=	83 - 86.9	D	=	63 - 66.9
B-	=	80 - 82.9	D-	=	60 - 62.9
			F	=	59 and below

with the weighting given by

422 & 522		**422 only**		**522 only**	
Homework	40%	422 project report	20%	522 project report	15%
Midterm	15%			522 review report	5%
Final	25%				

Attendance

This course follows the University Attendance Regulation (REG02.20.03) available at
http://policies.ncsu.edu/regulation/reg-02-20-03.

The instructor imposes no formal limits on absences, but attendance in lecture is strongly encouraged because the discussions will involve material not in any textbook. Students are responsible for all material presented or discussed in lecture.

Academic integrity

This course follows the University policy on academic integrity found in the Code of Student Conduct (POL11.35.01), available at
http://policies.ncsu.edu/policy/pol-11-35-01 and the Honor Pledge.

A student shall be guilty of a violation of academic integrity if he or she:

- Represents the work of others as his or her own;
- Obtains assistance in any academic work from another individual in a situation in which the student is expected to perform independently;
- Gives assistance to another individual in a situation in which that individual is expected to perform independently;
- Offers false data in support of laboratory or field work.

Violations will be reported to the Office of Student Conduct, which may impose penalties beyond those recommended by the instructor.

Students with disabilities

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students must register with Disability Services for Students at 1900 Student Health Center, Campus Box 7509, 919-515-7653. For more information on NC State's policy on working with students with disabilities, please see the Academic Accommodations for Students with Disabilities Regulation (REG 02.20.01) available at http://policies.ncsu.edu/regulation/reg-02-20-01.

Class evaluations

Online class evaluations will be available for students to complete during the last two weeks of the semester. Students will receive an email message directing them to a website (classeval.ncsu.edu) where they can login using their Unity ID and complete evaluations. All evaluations are confidential; instructors will never know how any one student responded to any question, and students will never know the ratings for any particular instructors. The student help desk can be reached at classeval@ncsu.edu. More information about evaluations is available at http://www.ncsu.edu/UPA/classeval/.

Schedule

This syllabus assumes a 15-week schedule with two 75 minute meetings per week. The midterm exam takes place during class time and the final follows the University final exam schedule. Reading assignments and exercises are from the text unless otherwise indicated. The order and content of lectures is subject to change.

Week	Dates	Topic	Assignments
BACKGROUND AND FOUNDATIONS			
1	Jan 8–10	Introduction and mathematical preliminaries	Read syllabus & Ch. 1
DATA ORGANIZATION AND PREPARATION			
2	Jan 15–17	Data types and data preparation	Read Ch. 2; HW 1 due 1/15
3	Jan 22–24	Visual exploration of data	Read Ch. 3; HW2 due 1/24
SUPERVISED ANALYSIS			Read Ch. 4
4	Jan 29–31	Classification and decision trees	HW3 due 1/31
5	Feb 5–7	Evaluating classifiers	Read Ch. 5
6	Feb 12–14	Ensemble methods and rule set classifiers	HW4 due 2/14
7	Feb 19	Probabilistic classification	
	Feb 21	*Midterm exam*	
8	Feb 26–28	Regression	
	Mar 5–7	Spring break, no classes ☺	
9	Mar 12–14	ANN and SVM classifiers	
10	Mar 19–21	Metric and nonparametric classifiers	HW5 due 3/21
UNSUPERVISED ANALYSIS			
11	Mar 26	Cluster analysis	Read Ch. 8
	Mar 28	Good Friday break, no classes ☹☺	
12	Apr 2–4	Cluster analysis	Read Ch. 9
13	Apr 9–11	Association analysis	Read Ch. 6; HW6 due 4/11
14	Apr 16–18	Association analysis	Read Ch. 7
15	Apr 23–25	Feature selection and dimension reduction	Reports due 4/25
	May 2	*Final exam*, 8-11AM	

