• ISE 365/465: Applied Data Mining

Syllabus: Spring Semester 2020

Course Description

This course is an introduction to Data Mining. Due to the explosion of data being collected and the advancement of computer technology, the field of Data Mining has evolved to find useful patterns of information in these large collections of data. The course focuses primarily on the data, modeling, and mathematical techniques used for the application of Data Mining. Students will gain a conceptual understanding of algorithms used for Data Mining and hands-on experience with these algorithms using IBM SPSS Modeler.

Course Objectives

Upon completion of this course, students will:

- Understand the CRISP-DM and SEMMA process for applying data mining
- Understand the importance of ***data*** in the data mining process and how to collect and set up data for a successful data mining project
- Have a conceptual understanding of several common data mining algorithms drawn from the fields of statistics and machine learning. This understanding will allow the students to apply appropriate algorithms correctly to solve data mining problems
- Be able to use IBM SPSS Modeler to collect, aggregate, explore and manipulate data from multiple sources to facilitate modeling
- Be able to use IBM SPSS Modeler to model and solve data mining problems using data mining algorithms
- Be able to use IBM SPSS Modeler to evaluate and present the results from a data mining project
- Understand the wide array of fields where data mining is applied today
- Have basic skills needed to be hired for a job in the field of data mining

Prerequisites

A basic statistics course covering topics such as statistical significance and linear regression - While small samples are not a problem in data mining, basic statistical techniques will be important when using many data mining techniques. Students are expected to have basic statistical knowledge before taking this course. For advanced data mining techniques, theoretical background will be taught to allow their correct application to data mining problems.

Contact Information

Professor
Brent L. Peterson
E-mail: blp219@lehigh.edu
Office: Mohler Lab #350
Office Hours: 9:15 – 11:00, Tuesdays and Thursdays

Textbook (required)

Data Mining: Concepts and Techniques, 3rd ed. By Han, Jiawei, Kamber, Michelle, and Pei, Jian. Elsevier, 2012. ISBN 978-0-12-381479-1.

Coverage will be in chapters 1-4, chapter 6, chapters 8-10, and chapter 13 for this course.

Course Philosophy

The goal of this course is to impart knowledge through traditional lectures, in-class computer labs, readings, homework, exams, and a project. In data mining, the best way to learn is through hands-on learning. Therefore, this course will strive to give sufficient background through readings and lectures to allow the students to spend as much time as possible using IBM SPSS Modeler to develop solutions to data mining problems from a number of different problem domains. **While the course outline below is organized by solution technique, we will cover many topics throughout the course including people issues, data collection and set-up, modeling tips and tricks, and deployment that are critical to the data mining process. Students will have the opportunity to practice what they learn in labs, homework, and the project. These assignments will be heavy in hands-on use of IBM SPSS Modeler. In addition, knowledge of topics covered will be tested in written exams.**

Academic Honesty

Integrity and Honesty are vital in life, especially for data miners, since we may have access to sensitive data and the systems we design or modify can improve people's quality of life, or can do irreparable harm. Using data mining ethically requires that we state all of the facts and assumptions in as clear a manner as possible, to avoid "lying with statistics". We are also bound by honor to give credit where it is due. In this class, you might ask others for help with a homework assignment. Once you write up your answer in your own words to turn in, it is a good idea to include a mention of their help on any particular problem. It is dishonest to copy homework solutions from past years that you might obtain or have. On quizzes and exams, of course, your work should be entirely your own. **Violations of academic honesty will result in a grade of 0 on the assignment where dishonest behavior occurred and possible disciplinary proceedings.**

Here is a statement of the Lehigh Student Senate on academic integrity: We, the Lehigh University Student Senate, as the standing representative body of all undergraduates, reaffirm the duty and obligation of students to meet and uphold the highest principles and values of personal, moral and ethical conduct. As partners in our educational community, both students and faculty share the responsibility for promoting and helping ensure an environment of academic integrity. As such, each student is expected to complete all academic course work in accordance to the standards set forth by the faculty and in compliance with the University's Code of Conduct.

Grading Policy

Your final numeric score will be determined as follows:

40% : Homework
5% : Class participation
15% : Midterm Exam
20%: Project
20% : Final Exam

Late Homework Policy: Late homework will incur an immediate 20% penalty if not handed in on time and an additional 1% for each hour late after the due time.

Missed Exams: Any exam missed without a legitimate excuse will result in a score of 0 on that exam.

Plus and minus grading will be used for final grades. Final grades will be "curved".

Accommodations for Students with Disabilities: Lehigh University is committed to maintaining an equitable and inclusive community and welcomes students with disabilities into all of the University's educational programs. In order to receive consideration for reasonable accommodations, a student with a disability must contact Disability Support Services (DSS), provide documentation, and participate in an interactive review process. If the documentation supports a request for reasonable accommodations, DSS will provide students with a Letter of Accommodations. Students who are approved for accommodations at Lehigh should share this letter and discuss their accommodations and learning needs with instructors as early in the semester as possible. For more information or to request services, please contact Disability Support Services in person in Williams Hall, Suite 301, via phone at 610-758-4152, via email at indss@lehigh.edu, or online at https://studentaffairs.lehigh.edu/disabilities.

Principles of Equitable Community:

Lehigh University endorses The Principles of Our Equitable Community [http://www.lehigh.edu/~inprv/initiatives/PrinciplesEquity_Sheet_v2_032212.pdf]. We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.

Tentative Course Outline – This will almost certainly change. Official due dates and topics will be announced in class by the instructor throughout the semester. This outline is for your reference only to give an idea of topics that may be covered.

Tuesday Date	Tuesday	Thursday
Jan 21	• Course Introduction • Pre-course survey • Chapter 1 – Intro to Data Mining • **Assign Chapters 1, 4.1-2 Reading**	• Chapter 1, Ch. 4.1 and 4.2 – Intro to Data Mining and Data Warehousing • US Presidential Election Lecture • **Assign Chapter 2 &3 Reading**
Jan 28	• Chapter 2 – Getting to Know Your Data	• IBM SPSS Modeler Graphs and Exploratory Statistics • **Assign Data Mining Paper and "10 Data Mining Mistakes" Reading**
Feb 4	• Chapter 3– Data Preprocessing • Feature Selection • IBM SPSS Modeler Data Manipulation • **Assign "Getting Started with SPSS Modeler" Reading: IBM Introduction.html**	• Merge Overview Discussion • **Data Understanding and Preprocessing Lab**
Feb 11	• **Data Understanding and Preprocessing Lab** • Assign Homework 1	• Applied Linear Regression Lecture
Feb 18	• Pop Quiz Results • Top 10 Data Mining Mistakes • Linear Regression Lab	• Linear Regression Lab Continued • **Assign Chapter 8.1, 8.2, 8.5, and 8.6 Reading** • Assign Homework 2 • **Homework 1 Due**
Feb 25	• Principal Components Analysis • Model Evaluation **– Chapter 8.5** • **Chapter 8.1 and 8.2** – Decision Trees	• **Chapter 8.1 and 8.2** – Decision Trees • **Homework 2 due**
Mar 3	• **Decision Tree Lab** • Mid-class Survey and Exam Review	• **MID-TERM EXAM**
Mar 10	• ***SPRING BREAK***	• ***SPRING BREAK***
Mar 17	• Mid-term Exam results • **Chapter 8.6** - Bagging, Boosting, Ensemble Models • **Project Description**	• **Chapter 9.3** - Neural Networks • **Chapter 9.4** - Support Vector Machines • Assign Homework 3
Mar 24	• **Neural Network/Support Vector Machines/Bagging and Boosting Lab**	• **Chapter 6** - Market Basket Analysis & Association Rules - Apriori
Mar 31	• **Chapter 6** - Market Basket Analysis & Association Rules - Apriori • **MBA/Apriori Lab** • **Homework 3 Due** • Assign Homework 4	• **MBA/Apriori Lab** • **Chapter 10** - Cluster Analysis
April 7	• **Chapter 10** - Cluster Analysis • **Cluster Analysis Lab**	• **Cluster Analysis Lab** • **Assign Ch. 9.5.1 Reading – k-NN** • **Homework 4 Due**
April 14	• Project Update • k-Nearest Neighbors	• **k-Nearest Neighbor Lab**
April 21	• **Chapter 13** - Advanced Data Mining Topics	• **Chapter 13** - Advanced Data Mining Topics • Final Exam Review
April 28	• PROJECT PRESENTATIONS	• PROJECT PRESENTATIONS

	• **Project Write-up Due**	