CSE 347-447 DATA MINING

Spring 2014 • 10:45 am – 12:00 noon TuTh • Packard Lab 208

Instructor **Professor Daniel Lopresti**
Email dal9@lehigh.edu ~ Ext 85782
Office Hours 3:00 pm – 5:00 pm on Tuesdays (or by appointment) in Packard Lab 350

Grader **Barri Bruno**
Email bdb214@lehigh.edu
Office Hours 2:15 pm – 3:15 pm on Mondays in Packard Lab 400 (PatRec Lab)

Text *Data Mining*, 3rd Ed., Ian H. Witten, Eibe Frank, and Mark A. Hall,
Morgan Kaufman, 2011, ISBN 978-0-12-374856-0

Software Weka 3: Data Mining Software in Java
Free download from: http://www.cs.waikato.ac.nz/ml/weka/index.html

CourseSite Lecture slides, assignments, etc. will be available @ http://coursesite.lehigh.edu/

Grading
- 8 homework assignments = 200 points (40%)
- Midterm exam = 100 points (20%)
- Final project presentation = 50 points (10%)
- Final project paper = 150 points (30%)
  (Note: Students taking CSE 447 will be required to write a more in-depth final paper.)

Notes
- Homework assignments will generally be posted on CourseSite by 9:00 am on Thursdays. Your work will be due by 9:00 pm on the following Tuesday. Submit your work electronically using the CourseSite Assignment feature.
- The late penalty is -5 points per day or fraction thereof. Homeworks turned in more than 3 days late will not be graded. Extensions must be approved by Professor Lopresti.

Week	Topics	Readings	Other Activities
Jan. 13	Course Intro; Data Mining and Machine Learning; Simple Examples	Secs. 1.0-1.2	HW #1 out
Jan. 20	Field Applications; Statistics; Generalization as Search; Ethics	Secs. 1.3-1.6	HW #1 due
	Input: Concepts, Instances, and Attributes	Ch. 2	HW #2 out
Jan. 27	Output: Knowledge Representation	Ch. 3	HW #2 due
	Inferring Rudimentary Rules; Missing Values; Constructing Decision Trees	Secs. 4.0-4.3	HW #3 out
		Supplemental reading: Ch. 17	
Feb. 3	Covering Algorithms; Mining Association Rules; Linear Models	Secs. 4.4-4.6	HW #3 due
	Instance-Based Learning; Clustering; Multi-Instance Learning	Secs. 4.7-4.9	HW #4 out
		Supplemental reading: Ch. 10; Secs. 11.0-11.2	
Feb. 10	Training and Testing; Predicting Performance; Cross-Validation;	Secs. 5.0-5.6	HW #4 due
	Comparing Data Mining Schemes Counting the Cost	Sec. 5.7	HW #5 out
		Supplemental reading: Secs. 11.3-11.4	
Feb. 17	Evaluating Numeric Prediction; Minimum Description Length; MDL for Clustering	Secs. 5.8-5.10	HW #5 due
	Decision Trees	Secs. 6.0-6.1	HW #6 out
		Supplemental reading: Secs. 11.6-11.7	

Week	Topics	Readings	Other Activities
Feb. 24	Classification Rules; Association Rules Extending Linear Models	Secs. 6.2-6.3 Sec. 6.4 Supplemental reading: Sec. 11.8; Ch. 12	HW #6 due
Mar. 3	Spring Break (no class)		
Mar. 10	Instance-Based Learning; Numeric *Midterm Exam (Thursday)*	Secs. 6.5-6.6 Supplemental reading: Ch. 13	
Mar. 17	*Return and discuss Midterm (Tuesday)* Bayesian Networks	Sec. 6.7	HW #7 out Final Project Proposals due
Mar. 24	Clustering Semisupervised Learning; Multi-Instance Learning	Sec. 6.8 Secs. 6.9-6.10	HW #7 due HW #8 out
Mar. 31	Attribute Selection; Discretizing Numeric Attributes Projections; Sampling; Cleansing	Secs. 7.0-7.2 Secs. 7.3-7.5	HW #8 due
Apr. 7	Transforming Multiple Classes; Calibrating Class Probabilities TBD	Secs. 7.6-7.7	
Apr. 14	*Final Project Presentations #1* *Final Project Presentations #2*		
Apr. 21	*Final Project Presentations #3* *Course Review and Wrap Up*		Final Project Papers due

Accommodations for Students with Disabilities

If you have a disability for which you are or may be requesting accommodations, please contact both your instructor and the Office of Academic Support Services, University Center C212 (610-758-4152) as early as possible in the semester. You must have documentation from the Academic Support Services office before accommodations can be granted.

Principles of Equitable Community

Lehigh University endorses The Principles of Our Equitable Community (http://www4.lehigh.edu/diversity/principles). 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.

Academic Integrity

The work you submit in CSE 347-447 must be entirely your own. While we encourage you to discuss basic concepts and strategies with friends and classmates, the copying or sharing of solutions to homeworks, in whole or in part, is never acceptable. Both the person receiving the copied work and the person providing the copied work are equally responsible. Such cases will be referred to the University Committee on Discipline and, if found guilty, you may be given the failing grade WF in the course.

If you have questions about this policy at any point throughout the semester, ask. It is far better to be safe than sorry when your academic career may be on the line.

Learning Outcomes

After taking CSE 347-447, you will:
(i) Understand the principles of data mining.
(ii) Be aware of the challenges that arise in data mining.
(iii) Know a range of techniques for data mining and where they can be applied.
(iv) Become aware of ethical issues that are present in data mining applications.