CSE 347-447 DATA MINING

Fall 2012   2:35 pm – 3:50 pm TuTh   STEPS 101

Instructor — **Professor Daniel Lopresti**
Email dal9@lehigh.edu ~ Ext 85782
Office Hours 4:00 pm – 6:00 pm Tu (or by appointment) in Packard Lab 350

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

10 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 2:35 pm (class time) on the following Tuesday. Submit your work electronically using the CourseSite Assignment feature.

The late penalty is -5 points per day or fraction thereof. The maximum penalty is -15 points. Extensions must be approved by Professor Lopresti.

Week	Topics	Readings	Other Activities
Aug. 27	Course Intro; Data Mining and Machine Learning; Simple Examples	Secs. 1.0-1.2	HW #1 out
Sept. 3	Field Applications; Statistics; Generalization as Search; Ethics Input: Concepts, Instances, and Attributes	Secs. 1.3-1.6 Ch. 2	HW #1 due HW #2 out
Sept. 10	Output: Knowledge Representation Inferring Rudimentary Rules; Missing Values; Constructing Decision Trees	Ch. 3 Secs. 4.0-4.3 Supplemental reading: Ch. 17	HW #2 due HW #3 out
Sept. 17	Covering Algorithms; Mining Association Rules; Linear Models Instance-Based Learning; Clustering; Multi-Instance Learning	Secs. 4.4-4.6 Secs. 4.7-4.9 Supplemental reading: Ch. 10; Secs. 11.0-11.2	HW #3 due HW #4 out
Sept. 24	Training and Testing; Predicting Performance; Cross-Validation; Comparing Data Mining Schemes Counting the Cost	Secs. 5.0-5.6 Sec. 5.7 Supplemental reading: Secs. 11.3-11.4	HW #4 due HW #5 out
Oct. 1	Evaluating Numeric Prediction; Minimum Description Length; MDL for Clustering Decision Trees	Secs. 5.8-5.10 Secs. 6.0-6.1 Supplemental reading: Secs. 11.6-11.7	HW #5 due
Oct. 8	*Pacing Break (no class)* Classification Rules; Association Rules	Secs. 6.2-6.3 Supplemental reading: Sec. 11.8	HW #6 out

Week	Topics	Readings	Other Activities
Oct. 15	Extending Linear Models	Sec. 6.4	HW #6 due
	Instance-Based Learning; Numeric Prediction with Local Linear Models	Secs. 6.5-6.6	
		Supplemental reading: Ch. 12; Ch. 13	
Oct. 22	*Midterm Exam (Tuesday)*		
	Return and discuss Midterm (Thursday)		HW #7 out
Oct. 29	Bayesian Networks	Sec. 6.7	HW #7 due
	Clustering	Sec. 6.8	HW #8 out
Nov. 5	Semisupervised Learning;	Secs. 6.9-6.10	HW #8 due
	Multi-Instance Learning		Final Project Proposals due
	Attribute Selection; Discretizing Numeric Attributes	Secs. 7.0-7.2	HW #9 out
Nov. 12	Projections; Sampling; Cleansing	Secs. 7.3-7.5	HW #9 due
	Transforming Multiple Classes; Calibrating Class Probabilities	Secs. 7.6-7.7	HW #10 out
Nov. 19	TBD	TBD	HW #10 due
	Thanksgiving (no class)		
Nov. 26	TBD	TBD	
	Final Project Presentations #1		
Dec. 3	*Final Project Presentations #2*		
	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.

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.