Graduate Center - CUNY

Computer Science Department

Data Mining

Course Description:

In this course students will learn popular data mining methods for extracting knowledge from data. It will balance theory and practice-the principles of data mining methods will be discussed, but students will also acquire hands-on experience using state-of-the-art software (WEKA) to develop data mining solutions to scientific and business problems. Topics and related methods discussed in the class include: data preprocessing, classification and prediction, association mining, and cluster analysis. This course will also include a project, which may either involve research in data mining or may address a practical problem using the methods and tools introduced in the class.

Course Objectives:

Upon successful completion of this course, students should be able to:

- To develop familiarity with data mining techniques, to learn to apply them to real world problems, and to understand the role and impact of data mining in our society.

Software: WEKA (http://www.cs.waikato.ac.nz/ml/weka/)

Resources:

This course uses Open Education Resources. Rather than a proper textbook, since the material is new and ever-changing, we have a collection of useful resources.

On Data Science:

https://github.com/jakevdp/PythonDataScienceHandbook

(all content is available free of charge in the form of Python notebooks, and the book can be purchased on Amazon).

On programming:

No prior programming knowledge is required, although having knowledge of Python is a plus. These are good introductory lectures and we will use some of this material in the first few weeks:

https://github.com/jrjohansson/scientific-python-lectures

https://github.com/jakevdp/WhirlwindTourOfPython/

https://developers.google.com/edu/python/

On practical implementation of Machine Learning algorithms in Python:

http://scikit-learn.org/stable/user_guide.html

Additional Texts:
"Introduction to Data Mining", P. Tan, M. Steinbach & V. Kumar. Addison Wesley. 2005.
"Data Science for Business", F. Provost, T. Fawcett. 2013

Academic Integrity – Students and all others who work with information, ideas, texts, images, music, inventions, and other intellectual property owe their audience and sources accuracy and honesty in using, crediting, and citing sources. As a community of intellectual and professional workers, the College recognizes its responsibility for providing instruction in information literacy and academic integrity, offering models of good practice, and responding vigilantly and appropriately to infractions of academic integrity. Accordingly, academic dishonesty is prohibited in The City University of New York and is punishable by penalties, including failing grades, suspension, and expulsion. The complete text of the College policy on Academic Integrity may be found in the catalog.

Grading Procedure:

Test1	15%
Test2	15%
Final Exam	25%
Midterm	20%
Projects	15%
Assignments	10%
	=====
TOTAL	100%

Letter Grade	A	A-	B+	B	B-	C+	C	D	F
Numerical Grade	93-100	90-92.9	87-89.9	83-86.9	80-82.9	77-79.9	70-76.9	60-69.9	<=59.9

Course Outline:

Week	Topic Covered	Reading
Week 1	Introduction: what is data mining, supervised vs unsupervised data mining	Ch1, pp. 1 – 11
	Data: types, concepts, instances/examples, features/attributes, target/class, outliers, etc. *Other issues related to data for data mining tasks will be discussed during the course of the class.*	(Ch2, stop middle 76)
Week 2	Introduction; first steps in Python Basic Python commands and tutorials; Jupyter notebooks Introduction to classification supervised learning, features, target: linear classifier	Ch4, pp. 145 – 149 (4.1, 4.2), lecture slides
Week 3	HW1 discussion (Jan. 29) Inductive learning: Decision trees, Shannon's information gain, brief math tutorial on probabilities and entropy	Ch4, pp. 150 – 172 (4.3), lecture notes
Week 4	Overfitting, evaluation (accuracy, precision, recall, cross-validation, leave-one-out cross-validation, brief introduction to ROC curves)	Ch4, pp. 4.4, lecture notes
	Brief intro to WEKA	
Week 5	Evaluation (contd.)	
	HW2 discussion Math tutorial: Bayesian theorem	Ch.5 pp. 227 – 229
Week 6	HW2 discussion Math tutorial: Bayesian theorem	Ch.5 pp. 227 – 229
	Midterm review Rule-based classifier	Ch5, 5.1
Week 7	Midterm	
	Midterm Revision Naïve Bayesian classifier	Ch.5 pp. 229 – 238
Week 8		
	Naïve Bayesian classifier / Nearest neighbor	Ch5, 5.2

Week 9	HW3 discussion Nearest neighbor (contd.) Unsupervised learning, similarity, example-based learning, clustering	Ch8 8.1-8.4 (skip 8.2.6, 8.3.3, 8.3.4)
Week 10	Clustering (contd.)	
	Association rule mining	Browse pp. 325 – 341
Week 11	Brief introduction to Neural networks HW4 discussion	Lecture notes
	Natural Language learning	Lecture notes
Week 12	Natural Language learning (contd.)	Lecture notes
	Project presentations	Lecture notes
Week 13	Advanced Topics	Lecture notes
	Advanced Topics	Lecture notes
Week 14	Final Exams	