CSC 480/680 Introduction to Data Mining

Calendar Description

Introduction to Data Mining (3) This course presents the main machine learning algorithms and evaluation methods developed to date in an intuitive way suitable for a non-specialized audience. It also introduces current research developments in the field and initiates students to the solving of applied programs in an innovative way, using existing machine learning tools. Grading: A-F only. Prerequisite: CSC-148, MATH-221, and STAT-202.

Instructor

Nathalie Japkowicz
Office: DMTI 112-B
Phone: (202) 885-6486
E-mail: japkowic@american.edu

Meeting Times and Locations

- Wednesdays: 2:30pm-5:20pm
- Online Course

Office Hours and Locations

- **Online:**

  Times: Mondays 2:30 – 4:30pm (if convenient for everyone) or by appointment, online

Overview

Data Mining (a.k.a. Machine Learning, Data Science, etc.) is the area of Artificial Intelligence concerned with the problem of building computer programs that automatically improve with experience. The intent of this course is to present a broad introduction to the principles and paradigms underlying machine learning, including presentations of its main approaches, discussions of its major theoretical issues, and overviews of its most important research themes.

Course Format

The course will consist of a mixture of lectures, labs and student presentations. The regular lectures will cover descriptions and discussions of the major approaches to Machine Learning as well as of its major theoretical issues. The student presentations will focus on recent research findings, the labs will help students familiarize themselves with the software environments and libraries typically used in machine learning experiments.

Evaluation

Students registered in **CSC 480** will be evaluated on:

- Critiques of 6 research papers (6% = 6 x 1%)
- An oral presentation of one research paper (3%)
- Two homework assignments (31%)
- A midterm exam (30%)
- A final class project of the student's choice (25%: project + 5%: poster = 30%). For the class project, students will propose their own topic in consultation with the instructor. Project proposals will be due in mid-semester.
- Extra credit: a maximum of 10% on all lab work during the sessions where the other group is meeting in person.

Students registered in **CSC 680** will be evaluated on:

- Six written group commentaries of research papers (12%)
- One oral presentation of one of these papers (8%)
- Two homework assignments (20%)
- A midterm exam (30%)
- A final class project of the student's choice (25%: project + 5%: poster = 30%). For the class project, students will propose their own topic in consultation with the instructor. Project proposals will be due in mid-semester.
- Extra credit: a maximum of 10% on all lab work during the sessions where the other group is meeting in person.

Grading Scale

In my world (and thus yours for the duration of this course), grades follow a different structure than the grades you are used to. Here is my scale:

A	➡	85% and above
A-	➡	80-84%
B+	➡	75-79%
B	➡	70-74%
B-	➡	65-69%
C+	➡	60-64%
C	➡	55-59%
C-	➡	50-54%
F	➡	less than 50

Pre-Requisites

Students should have some programming experience in a high level language, preferably Python.

Required Textbooks

- Peter Flach, Machine Learning: the Art and Science of Algorithms that Make Sense of Data (Cambridge University Press, 2012).
- Nathalie Japkowicz and Mohak Shah, Evaluating Learning Algorithms: A Classification Perspective , Cambridge University Press, 2011.

Class notes and other reading material

Class notes and other reading material will all be made available on Blackboard.

Student Learning Outcomes

In addition to introducing the students to the fascinating world of machine learning and data mining, this course will also familiarize the students to the art of scientific research. They will be exposed to research papers that they will need to present orally in their own words and critique in writing. They will also be required to come up with a research topic for their final project, research the literature relevant to their topic and expand on it. The students will need to find the balance between an interesting research topic and one that is doable in the amount of time available to them.

List of Major Approaches Surveyed

- Version Spaces
- Decision Trees
- Artificial Neural Networks
- Deep Learning
- Bayesian Learning
- Instance-Based Learning
- Support Vector Machines
- Ensemble Methods
- Unsupervised Learning/Clustering
- Association Rule Mining
- Genetic Algorithms

List of Theoretical Issues Considered

- Philosophical roots of Machine Learning
- Machine Learning versus Data Mining versus Big Data Analysis
- Experimental Evaluation of Learning Algorithms
- Computational Learning Theory

List of Practical Issues Considered

- Data Exploration
- Data Preparation
- Feature Selection
- The Class Imbalance Problem

Coursework related information:

- **Assignment 1:** Supervised Learning using Scikit-learn; Deep Learning using Colab, Tensorflow and Keras.
  - Handed out on ***Wednesday August 26*****;**
  - **Part I** due on ***Wednesday September 23***.
  - **Part II due on** ***Wednesday October 7***.

- **Assignment 2:** Supervised and Unsupervised Learning.
  - Handed out on ***Wednesday October 7***;
  - Due on ***Wednesday October 28***.

- **Midterm Exam:** The midterm exam will take place in class on ***Wednesday November 11***.

- **Research papers and presentations:** The research papers that the students will read and present to satisfy the written and oral presentation requirement in this course will be posted on Blackboard.

  - **CSC-480:**
    - **Critiques:** The students will work individually on this assignment. From each set of critiques, they will choose one research paper to read, summarize and critique. They must hand in a summary of the papers' findings and critiques on the days stated in the schedule.
    - **Presentations:** Each student is required to choose one of the research papers to present in class. Presentations will be done individually. Each student will speak for approximately 10 minutes and answer some of the other

students/instructor's questions. The presentations will take throughout the term.

- **CSC-680:**
  - **Critiques:** The students will work in groups of two to read three research papers on a regular basis and hand in a summary of the papers' findings and critiques of these papers on the days stated in the schedule. Each group is required to hand in a single set of critiques. Both students in the group must participate in a group discussion which will be summarized in the critique.
  - **Presentations:** Each student is required to choose one of the research papers to present in class. Presentations will be done individually. Each student will speak for approximately 10 minutes and answer some of the other students/instructor's questions. The presentations will take place throughout the term.

- **Final Project:** You will have to select a final project topic and hand out a project proposal by ***Wednesday October 21***. I will be available throughout the term to help you select a topic and get started on your work. On the last day of classes, ***Wednesday December 2***, there will be a session during which every student will present a 5-minute 3-slide description of their project to the class.

CSC-480 Students:

Important dates at a glance (with % of final grade represented by the student's work):

- Wednesday August 26: Assignment 1 handed out (16%), Part 1 due on September 23; Part 2 due on October 7.
- Wednesday September 9: Paper Critiques 1 due (1%)
- Wednesday September 16: Paper Critiques 2 due (1%)
- Wednesday September 23: **Assignment 1, Part 1 due (12%)**
- Wednesday September 30: Paper Critiques 3 due (1%)
- Wednesday October 7: **Assignment 1, Part 2 due (4%)**
- Wednesday October 7: Assignment 2 handed out, due on November 13 (15%)
- Wednesday October 14: Paper Critiques 4 due (1%)
- Wednesday October 21: Project proposal due
- Wednesday October 28: **Assignment 2, due (15%)**
- Wednesday November 4: Paper Critiques 5 due (1%)
- Wednesday November 4: Midterm Exam Review
- Wednesday November 11: **Midterm Exam (30%)**
- Wednesday November 18: Paper Critique 6 due (1%)
- Wednesday November 25: No Class – Thanksgiving Holiday
- Wednesday December 2: Project Presentation (5%)
- Wednesday December 2: **Final Project due (25%)**

- Student presentation on one of the presentation days (Sep 9, 16, 30 Oct 14, Nov 4 or 18): (3%)

CSC-680 Students:

Important dates at a glance (with % of final grade represented by the student's work):

- Wednesday August 26: Assignment 1 handed out (10%), Part 1 due on September 23; Part 2 due on October 7.
- Wednesday September 9: Paper Critiques 1 due (2%)
- Wednesday September 16: Paper Critiques 2 due (2%)
- Wednesday September 23: **Assignment 1, Part 1 due (7%)**
- Wednesday September 30: Paper Critiques 3 due (2%)
- Wednesday October 7: **Assignment 1, Part 2 due (3%)**
- Wednesday October 7: Assignment 2 handed out, due on November 13 (10%)
- Wednesday October 14: Paper Critiques 4 due (2%)
- Wednesday October 21: Project proposal due
- Wednesday October 28: **Assignment 2, due (10%)**
- Wednesday November 4: Paper Critiques 5 due (2%)
- Wednesday November 4: Midterm Exam Review
- Wednesday November 11: **Midterm Exam (35%)**
- Wednesday November 18: Paper Critique 6 due (2%)
- Wednesday November 25: No Class – Thanksgiving Holiday
- Wednesday December 2: Project Presentation (5%)
- Wednesday December 2: **Final Project due (20%)**
- Student presentation on one of the presentation days (Sep 9, 16, 30 Oct 14, Nov 4 or 18): 8%

Course Support:

The following documents will be posted on Blackboard to help you with the course:

- Suggested Outline for Paper Commentaries
- Midterm Exam Review Sheet

- Project Description

- Guidelines for the Final Project Report

- Additional Textbook References

Other material is available on the Web:

- **Scikit-learn**
- **TensorFlow**
- **KERAS**
- **WEKA**
- **UCI Machine Learning Repository**
- **R Code from the Japkowicz and Shah Evaluation Book**
- **KAGGLE**
- **KDNUGGETS**
- **Free Book: Information Theory, Inference, and Learning Algorithms, David MacKay**
- **David Aha's Machine Learning Resource Page**

Class Schedule:

Week	Topics	Readings
Week 1: Aug. 26	Organizational Meeting Practical Overview of Machine Learning **Homework 1 handed out: Part I Due on Sept 23. Part II Due on October 7.**	**Texts:** • **Flach:** Prologue, Chapters 1, 2 • **Japkowicz & Shah :** Chapter 1
Week 2: Sep. 2	Theoretical Overview of Machine Learning Philosophical Roots of Machine	**Texts:** • **Flach:** Prologue, Chapters 1, 2 • **Japkowicz & Shah :** Chapter 1

	Learning **Lab 1: Using Scikit-Learn**	
Week 3 Sep. 9	Decision Trees **Paper Critiques 1:** Due **Presentation Session 1**	- **Flach:** Chapters 4, 5 - **Japkowicz & Shah:** Chapter 2
Week 4: Sep. 16	Artificial Neural Networks **Paper Critiques 2:** Due **Presentation Session 2**	**Texts:** - **Flach:** Sections 7.1-2

Week 5: Sep. 23	Deep Learning **Lab 2:Using Colab, Tensorflow and Keras** **Part I of Homework 1 due.**	
Week 6: Sept. 30	Experimental Evaluation of Learning Algorithms **Paper Critiques 3** Due **Presentation Session 3**	**Texts:** - **Japkowicz & Shah:** Chapters 3-6 - **Flach:** Chapter 12
Week 7 Oct. 7	Bayesian Learning **Part II of Homework 1 due** **Homework 2: Handed out; Due on Oct 28.**	- **Flach:** Chapter 10
Week 8 Oct. 14	Instance-Based Learning **Paper Critiques 4** **Presentation Session 4**	**Texts:** - **Japkowicz & Stephen:** Class Imbalance Paper (on blackboard) - **Flach:** Chapter 9,
Week 9 Oct. 21	Support Vector Machines	**Texts:** - **Flach:** Chapter 8

	Classifier ensembles Unsupervised Learning **Project Proposal:** Due	
Week 10: Oct. 28	Data Exploration and Preparation Feature Selection; Class Imbalance Problem **Homework 2: Due**	**Texts:** • **Flach:** Sections 7.3-5
Week 11: Nov. 4	**Exam Review** **Paper Critiques 5** **Presentation Session 5**	**Texts:** • **Flach:** Chapter 11 • **Flach:** Sections 3.3, 6.3
Week 12: Nov. 11	**In-Class Exam**	
Week 13: Nov. 18	Association Rule Mining Big Data Analysis **Paper Critiques 6:** Due **Presentation Session 6**	**Texts:** • **Flach:** Sections 3.3, 6.3 • **Japkowicz & Stefanowski:** Big Data Analysis paper (on Blackboard)
Week 14: Nov. 25	**Thursday Nov 25: Thanksgiving: No Classes**	
Week 15: Dec. 2	**Project Presentations**	