

# CSC 411

## MACHINE LEARNING and DATA MINING

Lectures: Monday, Wednesday 3-4  
Lecture Room: Bahen 1220  
Instructor: Richard Zemel <zemel@cs.toronto.edu>  
Office hours: Thursday 4-5 Pratt 290D, and by appointment  
Tutor: Daniel Tarlow  
Tutorials: Friday 3-4  
Tutorial Room: Bahen 1220  
Class URL: [www.cs.toronto.edu/~zemel/Courses/csc411.html](http://www.cs.toronto.edu/~zemel/Courses/csc411.html)

### Overview

Machine learning research aims to build computer systems that learn from experience. Learning systems are not directly programmed by a person to solve a problem, but instead they develop their own program based on examples of how they should behave, or from trial-and-error experience trying to solve the problem. These systems require learning algorithms that specify how the system should change its behavior as a result of experience. Researchers in machine learning develop new algorithms, and try to understand which algorithms should be applied in which circumstances.

Machine learning is an exciting interdisciplinary field, with historical roots in computer science, statistics, pattern recognition, and even neuroscience and physics. In the past 10 years, many of these approaches have converged and led to rapid theoretical advances and real-world applications.

This course will focus on the machine learning methods that have proven valuable and successful in practical applications. This course will contrast the various methods, with the aim of explaining the circumstances under which each is most appropriate. We will also discuss basic issues that confront any machine learning method.

### Pre-requisites

You should understand basic probability and statistics, and college-level algebra and calculus. Knowledge of linear algebra will be a big help. For the programming assignments, you should have some background in programming, and it would be helpful if you know Matlab. Some introductory material for Matlab will be available on the course website as well as in the first tutorial.

## Readings

There is no required textbook for the course. Lecture slides will be available from the course website.

I will recommend readings from this book: *Pattern Recognition and Machine Learning*, by Christopher Bishop. This book is available at the bookstore (for \$108); at amazon.ca (for \$89); and at amazon.com (for \$55 US).

Relevant readings from this book are listed in this syllabus. However, some of the topics I will cover are not included in the book; I will make supplementary readings for most of these topics available on the course website.

## Course requirements and grading

The format of the class will be lecture, with some discussion. I strongly encourage interaction and questions. There are assigned readings for each lecture that are intended to prepare you to participate in the class discussion for that day.

The grading in the class will be divided up as follows:

Assignments	45%
Mid-Term Exam	25%
Final Exam	30%

There will be three assignments, each of which will be worth 15% of your grade.

Final grades will be based on a class curve. I will give each assignment and test a numerical score, and will try to give you an idea of where you lie along the curve as the course progresses.

## Homework assignments

The best way to learn about a machine learning method is to program it yourself and experiment with it. So the assignments will generally involve implementing machine learning algorithms, and experimentation to test your algorithms on some data. You will be asked to summarize your work in brief (3-4 page) write ups. The implementations must be done in Matlab, but prior knowledge of Matlab is not required. A brief tutorial on Matlab is available from the course web-site. You may also use Octave.

Collaboration on the assignments is not allowed. Each student is responsible for his or her own work. Discussion of assignments and programs should be limited to clarification

of the handout itself, and should not involve any sharing of pseudocode or code or simulation results. Violation of this policy is grounds for a semester grade of F, in accordance with university regulations.

The schedule of assignments is included in the syllabus. Assignments are due at the beginning of class on the due date. Because they may be discussed in class that day, it is important that you have completed them by that day. Assignments handed in late but before 5 pm of that day will be penalized by 5% (i.e., total points multiplied by 0.95); a late penalty of 10% per day will be assessed thereafter. Extensions will be granted only in special situations, and you will need a Student Medical Certificate or a written request approved by the instructor at least one week before the due date.

## **Exams**

There will be a mid-term in class on October 21<sup>st</sup>, which will be a closed book exam on all material covered up to that point in the lectures, tutorials, required readings, and assignments.

The final will not be cumulative, except insofar as concepts from the first half of the semester are essential for understanding the later material.

## **Attendance**

I expect students to attend all classes, and all tutorials. This is especially important because I will cover material in class that is not included in the textbook. Also, the tutorials will not only be for review and answering questions, but new material will be covered.

## **Electronic Communication**

Feel free to email me with questions or comments about the course. I will try to respond promptly. You should include your full name in the email, and it may also be useful to include your CDF account name and/or student number.

There is a bulletin board set up for this course:

<https://csc.cdf.toronto.edu/bb/YaBB.pl?board=CSC411H1F>

This is a good place for discussion of class topics and assignments.

For questions about marks on the assignments, please first contact the TA, Danny Tarlow. Questions about the exams should be addressed to me.

## CLASS SCHEDULE (subject to change)

<b>Date</b>	<b>Topic</b>	<b>Reading</b>	<b>Assignments</b>
Sep 9	Introduction	Sections 1-1.2.4; 2-2.3	
Sep 14	Basic Methods · linear regression · $k$ -nearest neighbors	Sections 3-3.2; 2.5 8.1-8.2	
Sep 16	Probabilities and Loss Functions	Sections 2-2.4	
Sep 21	Simple Classifiers · naive Bayes · logistic regression	Sections 4-4.3	
Sep 23, 28	Decision Trees		Asst 1 Out (9/28)
Sep 30	Over-fitting		
Oct 5, 7	Neural Networks	Sections 5-5.3; 5.5	
Oct 14, 19	Bayesian Learning	Supplementary	Asst 1 In (10/14)
Oct 21	MIDTERM		
Oct 26	Ensemble Methods	Supplementary	
Oct 28 Nov 2, 4, 9	Unsupervised Learning	Sections 9-9.2	Asst 2 Out
Nov 11, 16	Support Vector Machines	Sections 7-7.2	Asst 2 In (11/11)
Nov 18, 23	Hidden Markov Models	Sections 13-13.2	Asst 3 Out (11/18)
Nov 25, 30	Reinforcement Learning	Supplementary	
Dec 2	Wrap-up		Asst 3 In