

# CSC 411

## MACHINE LEARNING and DATA MINING

Lectures:	Monday, Wednesday 3-4
Lecture Room:	BA 1220
Instructor:	Richard Zemel <csc411prof@cs.toronto.edu>
Office hours:	Thursday 4-5 Pratt 290D, and by appointment
Teaching Assistants:	Hannes Bretschneider, Kevin Swersky, and Charlie Tang
TA email:	<csc411ta@cs.toronto.edu>
Tutorials:	Friday 3-4
Tutorial Room:	BA 1220
Class URL:	<a href="http://www.cs.toronto.edu/~zemel/Courses/CS411.html">www.cs.toronto.edu/~zemel/Courses/CS411.html</a>

## 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

The textbook for this year is a new book by Kevin Murphy. The book has not been published yet, but Kevin has let me print advanced copies of the book. We are getting copies printed now, and information will be posted here when we have more details on how and when they will be available.

## 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	50%
Mid-Term Exam	20%
Final Exam	30%

There will be an initial small assignment worth 5% of your grade, and then three more assignments, each of which will be worth 15% of your grade.

## 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/tutorial 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.

For one of the assignments, we will have a *bake-off*: a competition between machine learning algorithms. I will give everyone some data for training a machine learning system, and you will try to develop the best method. We will then determine which system performs best on some unseen test data.

## **Exams**

There will be a mid-term in class on October 24<sup>th</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.

The exams will cover material presented in lectures, tutorials, and assignments. You will not be responsible for topics in the reading not covered in any of these.

## **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 also 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 will be a bulletin board set up for this course. This is a good place for discussion of class topics and assignments.

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

## CLASS SCHEDULE, Part 1

Shown below are the topics for lectures and tutorials (in italics), as are the dates that each assignment will be handed out and is due. The notes from each lecture and tutorial will be available on the class web-site the evening after the class meeting. The assigned readings are specific sections from the book. All of these are subject to change.

<b>Date</b>	<b>Topic</b>	<b>Reading</b>	<b>Assignments</b>
Sep 12	Introduction	1.1	
Sep 14	Basic Methods & Concepts	1.3, 7.2-7.2.2	
<i>Sep 16</i>	<i>Matlab &amp; Linear algebra</i>		
Sep 19	Simple Classifiers	1.2-1.2.1, 1.2.3-1.2.5	
Sep 21	Nonlinear Classifiers	1.2.9-1.2.10	Asst 1 Out
<i>Sep 23</i>	<i>Optimization for ML</i>	11.1, 11.2	
Sep 26	Logistic Regression	7.4-7.4.1	
Sep 28	Neural Networks	19.6-19.6.5	Asst 1 In
<i>Sept 30</i>	<i>Neural network examples</i>		
Oct 3	Regularization & Cross-validation	1.2.6,1.2.7,16.1	
Oct 5	Kernels & Margins	18-18.1	Asst 2 Out
<i>Oct 7</i>	<i>Probability for ML</i>	2-2.5	
[Oct 10]	Thanksgiving: No class		
Oct 12	Support Vector Machines	18.4.2-18.4.5	
<i>Oct 14</i>	<i>Help with homework 2</i>		
Oct 17	Probabilities & Loss Functions	3-3.5, 8-8.2.4	
Oct 19	Probabilistic Classifiers	1.2.2, 6-6.3.3	Asst 2 In
<i>Oct 21</i>	<i>Mid-term review</i>		
Oct 24	MIDTERM		

## CLASS SCHEDULE, Part 2

<b>Date</b>	<b>Topic</b>	<b>Reading</b>	<b>Assignments</b>
Oct 26	Bayesian Reasoning		
Oct 28	<i>Intro to Graphical Models</i>	20.3	
Oct 31	Clustering	19.2, 19.5	Asst 3 Out
Nov 2	EM and mixture models	10.4-10.4.2, 19.3	
Nov 4	<i>Mixtures of Gaussians</i>		
[Nov 7]	Mid-term break: No class		
Nov 9	PCA & Factor Analysis	20.3	
Nov 11	<i>ICA</i>		
Nov 14	Ensemble Methods	18-18.1, 18.5	Asst 3 In
Nov 16	Reinforcement Learning	8.4, 12	
Nov 18	<i>Bagging &amp; Boosting</i>	18.7-18.8	Asst 4 Out
Nov 21	Q Learning		
Nov 23	Manifolds & Embeddings	11	
Nov 25	<i>RL in action</i>		
Nov 28	Bayesian Methods	8.4, 12	
Nov 30	Markov Random Fields		
Dec 2	<i>TBD</i>		Asst 4 In
Dec 5	Conditional Random Fields		