CS 340: Machine Learning
Lecture 1: Introduction

AD

January 2011

Administrivia & Announcement

- Webpage: http://www.cs.ubc.ca/~arnaud/cs340_W2010.html
- TA: Hajir (hajir@cs.ubc.ca) and Marcos (ginestra@cs.ubc.ca)
- Tutorial T2A F 3.00-4.00, Dempster 101.
- Tutorial T2B M 11.00-12.00, Dempster 101.
- Instructor: Arnaud Doucet. Office hours: Monday 5.00-6.000.
- Acknowledgments to Nando de Freitas & Kevin Murphy.

Grading

- Midterm: 25% (11th February)
- Final: 50%
- Assignments: 25%
- **Collaboration policy**: You can collaborate on homeworks (groups limited to 2 persons) if you write the name of your collaborators on what you hand in; however, you must understand everything you write, and be able to do it on your own (eg. in the exam!)
- **Sickness policy**: If you cannot do an assignment or an exam, you must come see me in person; a doctor's note (or equivalent) will be required.

Pre-requisites

- Maths
  - Multivariate calculus

$$\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial x_i} = 2x_i, \quad \frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

  - Linear algebra

$$A\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

  - Probability.

$$Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

- CS:
  - Programming skills
  - Knowledge of data structures and algorithms.

Textbook

- Textbook: K.P. Murphy, *Machine Learning: A Probabilistic Approach*. MIT Press, to appear. Available for purchase for $XX from Copiesmart in the UBC Village (next to Macdonald's).
- Optional Textbook:: C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- Optional Textbook: Hastie, Tibshirani and Friedman, *Elements of Statistical Learning - Data Mining, Inference and Prediction*, Springer-Verlag, 2001.

Matlab

- Matlab is a mathematical scripting language widely used for machine learning (and engineering and numerical computation in general).
- Everyone should have access to Matlab via their CS account. If not, you can ask the TAs for a CS guest account.
- You can buy a student version from the UBC bookstore, but you will also need the stats toolbox (and sometimes also the optimization toolbox).
- The first homework involves of some simple Matlab programming. Check you have Matlab today!

Learning Objectives

By the end of this class, you should be able to

- Understand basic principles of machine learning and its connection to other fields
- Derive, in a precise and concise fashion, the relevant mathematical equations needed for familiar and novel models/ algorithms
- Implement, in reasonably efficient Matlab, various familiar and novel ML model/ algorithms
- Choose an appropriate method and apply it to various kinds of data/ problem domains

What is Machine Learning?

- "Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time" – Herbert Simon.
- Machine Learning is an interdisciplinary field at the intersection of Statistics, CS, EE, neuroscience etc.
- At the beginning, Machine Learning was fairly heuristic (inspired by neural nets and the cortex of frogs) but it is now much closer to Statistics.

Machine Learning vs Statistics

Machine learning	**Statistics**
neural networks, graphs	logistic regression, models
neurons, weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large dataset=10^9 data points	large dataset=1000 data points
large grant = \$1,000,000	large grant = \$50,000

The Future of Machine Learning and Statistics

- "For Today's Graduate, Just One Word: Statistics", New York Times, 2009.
- "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades". — Hal Varian, McKinsey Quaterly Journal, 2009.
- Hal Varian is the chief economist of Google.

What is learning all about?

- Face & fingerprints (airport), speech/music (Shazam), handwriting recognition (Windows 7).
- Automated recommendations - movies, products to recommend based on input preferences (NetFlix, Amazon).
- Automatic translation.
- Anomaly detection - fraud detection (e.g. credit cards).
- Humans are unable/don't want to explain their expertise (speech recognition, poker).
- Biological sciences for high throughput screening - labeling biological sequences, molecules, assays.
- Solutions changes with time (tracking, robot control).
- The problem size is too vast for our limited reasoning capabilities (calculating webpage ranks and matching ads to facebook pages).

Success Stories of Machine Learning

- Deep Blue beats World Chess Champion (1997). Kasparov: "I sometimes saw deep intelligence and creativity in the machine's moves"
- Google (PageRank, YouTube indexing, Google News), Amazon, Autonomy.
- Computer vision (detection, tracking, recognition) and animation.
- Autonomous robots: self localization and mapping.
- Autonomous helicopter (Ng et al.)
- Autonomous vehicle: The DARPA challenge (Thrun et al.)
- Video games (TrueSkill ranking system).
- Automatic trading systems (automatic detection of patterns in stocks).
- Betting syndicates.

Main Learning Tasks

- *Supervised Learning*
  Predict output from input based on training data.
- *Unsupervised Learning*
  Find patterns in data
- *Reinforcement Learning*
  Learn how to behave in novel environments (eg robot navigation, poker)
- *Active Learning*
  The machine can query the environment to gather the best possible data. "Good" data is often better than a lot of data!