Today

- Introduction to machine learning
- The course (syllabus)
- Math review (probability + linear algebra)
- The future of machine learning
Machine Learning (ML)

• Science that studies statistical and computational aspects of modeling data for predictive purposes

• (Mostly) Empirical science
• Task: Predict whether an image contains a malignant tumor

• Task: Predict the next movie a person should watch
THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
Subjective view of how ML relates to other fields
Historical View

- (Modern) Statistics: ~1900
- Machine Learning and Data Mining: ~1960
- Data Science: ~2000
Computer Science + Engineering

Statistics + Mathematics

Machine Learning

Data Science (BI)

Danger Zone

Traditional Research

Substantive Expertise

Inspired by Drew Conway
“Data analysis, machine learning and data mining are various names given to the practice of statistical inference, depending on the context.”

–Larry Wasserman in “All of Statistics: A Concise Course in Statistical Inference.”
Attitudes in Machine Learning and Data Mining
Versus Attitudes in Traditional Statistics

Despite these differences, there’s a big overlap in problems addressed by machine learning and data mining and by traditional statistics. But attitudes differ...

<table>
<thead>
<tr>
<th>Machine learning</th>
<th>Traditional statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>No settled philosophy or widely accepted theoretical framework.</td>
<td>Classical (frequentist) and Bayesian philosophies compete.</td>
</tr>
<tr>
<td>Willing to use <em>ad hoc</em> methods if they seem to work well (though appearances may be misleading).</td>
<td>Reluctant to use methods without some theoretical justification (even if the justification is actually meaningless).</td>
</tr>
<tr>
<td>Emphasis on automatic methods with little or no human intervention.</td>
<td>Emphasis on use of human judgement assisted by plots and diagnostics.</td>
</tr>
<tr>
<td>Methods suitable for many problems.</td>
<td>Models based on scientific knowledge.</td>
</tr>
<tr>
<td>Heavy use of computing.</td>
<td>Originally designed for hand-calculation, but computing is now very important.</td>
</tr>
</tbody>
</table>

ML vs. other fields

• Traditional statistics

  • Causality and understanding are more important than prediction. Data size is less important. Less computational aspects. Simpler, linear relationships.

• Data mining (somewhat synonymous of ML)

  • “Computational Efficiency [...] more important than statistical sophistication”. Often business-related problems.

Data View

- Statistics
  - Designs experiments to collect data

- Machine Learning
  - Uses available data but also designs collection mechanisms

- Data Science
  - Uses data made available by other processes
Applications of ML
At last — a computer program that can beat a champion Go player

ALL SYSTEMS GO
Robotic observatory makes fast work of astronomy  p. 479
A wet route to methanol  p. 522
Human noise plagues protected areas  p. 526

DIGITAL CARDS WHIZ
AI beats humans at challenging poker variant  p. 508
Progressive Growing of GANs for Improved Quality, Stability, and Variation

Karras et al., ICLR’18
Progressive Growing of GANs for Improved Quality, Stability, and Variation
Karras et al., ICLR’18
All computer generated
Deep Visual-Semantic Alignments for Generating Image Descriptions
A. Karpathy, L. Fei-Fei, CVPR’15
Deep Visual-Semantic Alignments for Generating Image Descriptions
A. Karpathy, L. Fei-Fei, CVPR'15
AI for video games

Mnih et al.  
Nature  
Volume 518,  
pages 529-533  
(26 February 2015)
• Medicine: personalized, automate diagnostics

• Social sciences: prediction problem (e.g., predict recidivism)

• Engineering: to propose new design, evaluate without building

• Finance: capture uncertainty, short-term trading

• Marketing: to understand and quantify user experience, advertising efficacy

• Many others: conservation, social projects, climate change

• Your domain of expertise...
Course Introduction & Goals
Logistics

- Course syllabus: http://www.cs.toronto.edu/~lcharlin/courses/80-629/
- (Google my name. There’s a link from my webpage.)
- 2x 10-minute breaks, one every hour (or so)
Fit with other courses

- HEC
  - PhD level (originally)
  - Computationally oriented
- Other ML courses in Montreal (U.Montreal, Polytechnique, McGill)
  - More applied (similar to COMP-551@McGill)
Short review of linear algebra, statistics, and probabilities
• Based on chapters 2 and 3 of “Deep Learning”

http://www.deeplearningbook.org/
Linear algebra

- **Scalar**: a single value.
  \[ a \in \mathbb{R}, \ a \in \mathbb{N} \quad a = 3 \]

- **Vector**: an array of values.
  \[ a \in \mathbb{R}^D, \ a \in \mathbb{N}^D \quad a = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} \]

- **Matrix**: a table of values.
  \[ A \in \mathbb{R}^{D_1 \times D_2}, \ A \in \mathbb{N}^{D_1 \times D_2} \quad A = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 2 & 9 \end{bmatrix} \]
Indexing notation

- Indexing elements of a vector: $a_i$
  
  $a = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} \leftarrow a_1$

- Indexing elements of a matrix: $a_{ij}$
  
  $A = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 2 & 9 \end{bmatrix}$

Convention:
The first element is the zero'th.
Simple operations

- Transpose

\[
a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}
\]

\[
a^\top = \begin{bmatrix} a_0 & a_1 & a_2 \end{bmatrix}
\]

\[(A_{ij})^\top = A_{ji}\]

- Addition

\[
b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}
\]

\[
a + b = \begin{bmatrix} a_0 + b_0 \\ a_1 + b_1 \\ a_2 + b_2 \end{bmatrix}
\]

\[(A + B)_{ij} = A_{ij} + B_{ij}\]

- Vectors and matrices w. the same shape
Simple operations

- Multiply by a scalar
  \[ \alpha a = \begin{bmatrix} \alpha a_0 \\ \alpha a_1 \\ \alpha a_2 \end{bmatrix} \]

- Vector product.

- The dot product
  \[ a^\top a = \sum_i a_i a_i \]
  - Note: it yields a scalar.

- Element-wise product:
  \[ a \odot a = \begin{bmatrix} a_0a_0 \\ a_1a_1 \\ a_2a_2 \end{bmatrix} \]
  - Also known as Hadamard product
Operations

- **Matrix product (dot product):**

  \[ C_{ij} = \sum_k A_{ik} B_{kj} \]

  - A's columns must equal B's rows (order is important)

  \[ A \in \mathbb{R}^{D_1 \times D_2}, B \in \mathbb{R}^{D_2 \times D_3} \]

- **Distributive:** \( A(B + C) = AB + AC \)

- **Associative:** \( A(BC) = (AB)C \)

- **Product of transpose:** \( (AB)^T = B^T A^T \)
Inverse

- We denote a matrix’s inverse as $A^{-1}$

- A matrix has an inverse iff:
  - it’s square. $D_1 = D_2$
  - its columns are linearly independent.
  - No column can be recovered using a combination of other columns

- Inverses are useful to solve systems of equations:
  $$Ax = b \quad x = A^{-1}b$$
Norms

- $L^p$ norm. Size of a vector (or matrix)

\[
\| \mathbf{a} \|_p = \left( \sum_i |a_i|^p \right)^{1/p}
\]

- Standard norms in ML:

  - Euclidean norm ($p=2$)

  \[
  \| \mathbf{a} \|_2 = \sqrt{ \left( \sum_i |a_i|^2 \right) } 
  \]

  - Dot product w. 2-norm:

  \[
  \mathbf{a}^\top \mathbf{b} = \| \mathbf{a} \|_2 \| \mathbf{b} \|_2 \cos \theta_{\mathbf{a}\mathbf{b}}
  \]

  - Frobenius norm (matrix):

  \[
  \| \mathbf{A} \|_2 = \sqrt{ \left( \sum_i \sum_j |a_{ij}|^2 \right) } 
  \]
Special matrices & vectors

- Identity. Denoted $I_n$.
  - All zeros except for ones on the main diagonal.

- Symmetric: $A = A^\top$

- Unit vector: $\|a\|_2 = 1$

- Orthogonal vectors: $a^\top b = 0$

- Orthonormal vectors: unit and orthogonal $A^\top A = AA^\top = I$

- Orthogonal matrix: Orthonormal rows & columns

$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
• Skip eigendecomposition, SVD, pseudo-Inverse, determinants (Sections 2.7–2.11).

• We will get back to them if/when needed in the course.
• On to probabilities

• Chapter 3 of “Deep Learning”
  
  • I’ve adapted some of the lecture slides from the book.

• Thanks to Ian Goodfellow for providing slides.
Why probabilities?

• To capture uncertainty

  E.g., What time will I get home tonight?

• Probabilities provide a formalism for making statements about “data generating processes” (L. Wasserman)

  E.g., what happens when I flip a fair coin?
The example

- Generate data by throwing a fair die.

- What do we know about a single throw?
  - 6 possible outcomes. (sample space)
  - Each outcome (e.g., 1). (element, state)
  - A subset of outcomes (e.g., <3). (event)
  - Outcomes are equiprobable. (uniform distribution)
Random variables and probabilities

• A random variable (r.v.) is a probabilistic outcome.

• For example,

  • Die throw ($X$)

  • The actual outcome is $\in \{1, 2, 3, 4, 5, 6\}$. ($x$)

• A probability function ($P$) assigns a real number to each possible event: $P(x) \geq 0$, $\forall x \in X$

  \[ P(\bigcup x) = 1 \]
Discrete RVs

- An RV is discrete if it takes a finite number of values\(^1\)

\[
P(x = x_i) \geq 0, \forall i
\]

\[
\sum_{i} P(x = x_i) = 1
\]

- E.g., uniform distribution:

\[
P(x = x_i) = \frac{1}{k}, \forall i
\]

- E.g., Poisson distribution:

\[
P(x = x_i; \lambda) = \frac{\lambda^{x_i} \exp^{-\lambda}}{x_i!}
\]

\(^1\) technically: it must be countable

both images are from: wikipedia.org
Continuous RVs

- An RV is continuous if \( f(x) \geq 0, \forall x \in X \)

\[
\int_{-\infty}^{\infty} f(x) \, dx = 1
\]

\[
P(a < x < b) = \int_a^b f(x) \, dx
\]

- \( f(x) \) is a probability density function (PDF)

- E.g., (continuous) uniform distribution:

\[
u(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}
\]

- E.g., Gaussian distribution

from: wikipedia.org
A few useful properties
(shown for discrete variables for simplicity)

• Sum rule: $P(x) = \sum_y P(x, y)$

• Product rule: $P(x, y) = P(x|y)P(y)$

• Chain rule: $P(x_1, \ldots, x_n) = P(x_1)P(x_2 | x_1)P(x_3 | x_2, x_1)$
  \[= P(x_1) \prod_i P(x_i | x_1, \ldots, x_{i-1})\]

• If $x$ and $y$ are independent: $P(x, y) = P(x)P(y)$

• Bayes’ Rule: $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
Moments

- **Expectation:** \( E[X] = \sum_{i} P(x = x_i)x_i \quad E[aX] = aE[X] \)
- **Variance:** \( \sigma^2 = E[(X - E[X])^2] \)
- **Covariance:** \( \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \)
- **correlation:** \( \rho(x, y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \)
What is the goal of ML?
• A bit of historical context

• When I started my PhD very few in ML talked about AI

• Recent ML makes progress toward “AI tasks”
  
  • Examples of AI tasks: translation, object recognition

• In that context: create a machine with human-like capacities? Or a machine that can help humans?
• My (imperfect) view:

  Understand data through predictive models

  Understand the world through predictive models
Further Reading

- Prologue to “The Master Algorithm”
  http://homes.cs.washington.edu/~pedrod/Prologue.pdf

- Ch. 1 of Hastie et al.

- Math Preparation
  - Ch.2 of Pattern Recognition and Machine Learning [PRML]
  - Ch.2-3 of Deep Learning [DL]

- Slightly more advanced: