Machine Learning for Large-Scale Data Analysis and Decision Making
80-629-17A

Week #1
Today

• Introduction to machine learning

• The course (syllabus)

• Math review (probability + linear algebra)

• The future of machine learning

• Why are you enrolled and what do you want from this class?
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

Substantive Expertise

Danger Zone

Machine Learning

Data Science (BI)

Traditional Research

[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
Digital cards whiz
AI beats humans at challenging poker variant p. 508
• 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
• Your domain of expertise...
Course Introduction & Goals
Course webpage

http://www.cs.toronto.edu/~lcharlin/courses/80-629/

- Google my name. There's a link from my home page.
Fit with other courses

• HEC
  • PhD level
  • Computationally oriented

• Other ML courses in Montreal (UdM, Polytechnique, McGill)
  • More applied
  • Emphasize on large-scale data using distributed computational methods
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.
  \[ a = 3 \]
  \[ a \in \mathbb{R}, a \in \mathbb{N} \]

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

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

- Indexing elements of a vector: $a_i$

$$a = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} \quad \leftarrow \quad a_1$$

- 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**

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

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

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

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

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

\[
(A + B)_{ij} = A_{ij} + B_{ij}
\]
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: \( A(BC) = (AB)C \)

\[ (AB)^T = B^T A^T \]
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:

  $$Ax = b \quad x = A^{-1}b$$

A square matrix not inversable is singular
Norms

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

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

• Standard norms in ML:
  
  • Euclidean norm (p=2)  
    $$\| a \|_2 = \sqrt{\left( \sum_i |a_i|^2 \right)}$$
  
  • Dot product w. 2-norm:  
    $$a^t b = \| a \|_2 \| b \|_2 \cos \theta_{xy}$$
  
  • Frobenius norm (matrix):  
    $$\| A \|_2 = \sqrt{\left( \sum_i \sum_j |a_{ij}|^2 \right)}$$
Special matrices & vectors

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

\[
I_3 = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

• Symmetric: $A = A^T$

• Unit vector: $\|a\|_2 = 1$

• Orthogonal vectors: $a^T b = 0$

• Orthonormal vectors: unit and orthogonal

• Orthogonal matrix: Orthonormal rows & columns

\[
A^T A = AA^T = I
\]
• 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:

$$u(x = x_i; a, b) = \frac{1}{b - a}$$

- E.g., Gaussian distribution
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:** $\mathbb{E}[X] = \sum_i P(x = x_i)x_i \quad \mathbb{E}[aX] = a\mathbb{E}[X]$

- **Variance:** $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$

- **Covariance:** $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

- **Correlation:** $\rho(x, y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$
What is the goal of ML?
• My (imperfect) view:
  Understand data through predictive models

• A bit of historical context

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

• Recent ML makes progress toward “AI tasks”

• In that context: create a machine with human-like capacities? Or a machine that can help humans?
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 Bishop
  • Ch.2-3 of Goodfellow et al.

• Slightly more advanced:
Why did you enroll?

• Why are you taking this class?
  • What do you hope to learn from it?
  • How do you plan on using it?
• Any particular topics you would like us to cover?