Lecture 1: Introduction

CIS 4190/5190
Spring 2023
Agenda

• Logistics
  • Course description
  • Tentative Schedule
  • Grading

• Introduction
  • Motivation
  • Basic definitions
  • Examples
Description

• **Key skills**
  • Identify opportunities for applying machine learning (ML) algorithms
  • Diagnose and debug issues in ML models

• Lectures will focus on developing mathematical understanding
• Assignments will focus on applying this understanding to implementing ML solutions
Prerequisites

• **Math:** University-level courses in probability, linear algebra, and multivariable calculus
  • Understand prior and posterior probabilities, \( \mathbb{E}[X] = \int p(x)dx \), etc.
  • Understand matrix ranks, inverses, and eigenvalues
  • Understand how to compute \( \nabla_A (Ax) \) for a matrix \( A \) and vector \( x \)
  • Tested in HW 1

• **Programming:** Previously coded up projects (preferably in Python) that were at least 100 lines of code long
  • We will provide Python help (primer + office hours) for students who know how to program but do not know Python
Course Comparisons

• **CIS 4190/5190 (this course)**
  - Basic mathematical ideas behind ML
  - Apply existing ML algorithms to new problems as an engineer or researcher

• **CIS 5200**
  - Deeper, more mathematically demanding introduction to ML
  - Perhaps do fundamental ML research in the future

Also see: [https://priml.upenn.edu/courses](https://priml.upenn.edu/courses)
Course Comparisons

• **CIS 4190/5190 (this course)**
  • Basic mathematical ideas behind ML
  • Apply existing ML algorithms to new problems as an engineer or researcher

• **CIS 5220**
  • Deep learning techniques and applications in more detail

• **CIS 5450**
  • Data science workflow including data wrangling, ML modeling, and analytics
  • Scaling ML to big datasets and clusters

Also see: [https://priml.upenn.edu/courses](https://priml.upenn.edu/courses)
CIS 4190 vs. 5190

• 5190 will have extra, **mandatory** components in the HW, which are **optional** for 4190

• **Example**
  • HW may have 45 points for 4190, and 5 extra points for 5190 (total of 50)
  • Student taking 4190 will get 100% if they get at least 45 points (typically by skipping the 5190 problem, but not necessarily)
  • You cannot score more than 100%
  • The written and coding portion are counted separately; you cannot make up written points using coding points and vice versa
## Schedule (Tentative)

<table>
<thead>
<tr>
<th>Week</th>
<th>Content</th>
<th>Homework</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Linear Regression</td>
<td>HW 1 Due</td>
</tr>
<tr>
<td>4</td>
<td>Logistic Regression</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Decision Trees</td>
<td>HW 2 Due</td>
</tr>
<tr>
<td>6</td>
<td>Neural Networks</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Computer Vision</td>
<td>HW 3 Due</td>
</tr>
<tr>
<td>8</td>
<td>Unsupervised Learning</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Natural Language Processing</td>
<td>HW 4 Due</td>
</tr>
<tr>
<td>10</td>
<td>Bayesian Networks</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Reinforcement Learning</td>
<td>HW 5 Due</td>
</tr>
<tr>
<td>12</td>
<td>Recommender Systems</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Additional Topics</td>
<td>HW 6 Due</td>
</tr>
<tr>
<td>14</td>
<td>Additional Topics</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Additional Topics</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Review</td>
<td></td>
</tr>
</tbody>
</table>
Grading Scheme (Tentative)

- **Homeworks (6×)**: 30%
- **Project**: 20%
- **Final exam (during exam week)**: 35%
- **Quizzes (12×, roughly weekly)**: 10%
  - 50% correct sufficient for full credit
- **Class participation**: 5%
Grading Scheme (Tentative)

• A+: 95+
• A: 90-95
• A-: 85-90
• B+: 80-85
• B: 75-80
• B-: 70-75
• Lower passing grades: 50-70

• May be curved up
Late Policy (Tentative)

• For each hour late, lose 0.5% on the points for that assignment
  • Homeworks, quizzes, project milestones
  • Max 48 late hours per assignment

• Example
  • Submit HW 1 20 hours late
  • Lose $20 \times 0.5 = 10\%$ on HW 1 (0.5% of overall grade)

• If you have a medical reason, email both professors a copy of your medical visit report, and we will grant an extension (typically 2 days)
  • We will consider other reasons on a case-by-case basis
Office Hours

• Each instructor & TA will have 1 hour of office hours each week
  • Times still being decided
Communication

• We will use Ed Discussion for questions and course discussions
  • Send a message to “instructors” to contact professors and TAs
  • You can contact the professors directly on Ed Discussion or by email (posted on course website); always contact both of us
Homework Schedule

• **6 homeworks**
  • Released every other Wednesday
  • Due Wednesday 2 weeks later (with an exception for HW 6)

• **HW 1**
  • Designed to test mathematical background
  • **Expected time:** 3 hours
  • Full points if you score 50% or more
  • Opportunity to get used to the workflow
Homework

- **Written problems**: GradeScope submission
  - LaTeX encouraged; handwritten + scanned at your own risk
  - Won’t be graded if you don’t annotate your answers correctly!

- **Coding problems**: AutoGrader + GradeScope submission of notebook
  - Colab/iPython notebook skeletons; AutoGrader as unit tests within skeleton
  - Only difference between AutoGrader and unit tests is different data
  - If code passes the unit tests and you didn’t “game” it, it should pass AutoGrader

• Discussion permitted for clarifications, but never share solutions/code; acknowledge all your discussions at the beginning of your report
Quiz Schedule

• 12 quizzes
  • Released every Wednesday
  • Due Thursday 8 days later

• Checks basic understanding of material covered the previous week
Agenda

• Logistics
  • Course description
  • Tentative Schedule
  • Grading

• Introduction
  • Motivation
  • Basic definitions
  • Examples
First Assignments

• **HW 1**: Released today, **due 1/25**
  • No office hours planned for HW 1
  • You can ask questions via Ed Discussion
  • 50% = full credit

• **Quiz 1**: Released 1/19, **due 1/26**
What is Machine Learning?

“Learning is any process by which a system improves performance from experience.”

Herbert Simon
What is Machine Learning?

“Machine learning ... gives computers the ability to learn without being explicitly programmed.”

Arthur Samuel
What is Machine Learning?

- **Tom Mitchell**: Algorithms that
  - improve their **performance** $P$
  - at **task** $T$
  - with **experience** $E$

- A well-defined machine learning task is given by $(P, T, E)$
Example: Prediction

Photo by NASA Goddard

Image: https://www.flickr.com/photos/gsfc/59375999688/
Data from https://nsidc.org/arcticseainews/sea-ice-tools/
Example: Prediction

• **Tom Mitchell:** Algorithms that
  • improve their **performance** $P$
  • at some **task** $T$
  • with **experience** $E$

  • $T = \text{predict Arctic sea ice extent}$
  • $P = \text{prediction error (e.g., absolute difference)}$
  • $E = \text{historical data}$
Machine Learning for Prediction

Data $Z$ → Machine learning algorithm → Model $f$
Machine Learning for Prediction

New input

Model $f$

Predicted output
**Example: Prediction**

Arctic Sea Ice Extent (millions of sq km) in September:

<table>
<thead>
<tr>
<th>Year</th>
<th>NSIDC Index of Arctic Sea Ice in September</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
</tr>
<tr>
<td>2025</td>
<td></td>
</tr>
</tbody>
</table>

Image: [https://www.flickr.com/photos/gsfc/5937599688/](https://www.flickr.com/photos/gsfc/5937599688/)
Data from [https://nsidc.org/arcticseaicenews/sea-ice-tools/](https://nsidc.org/arcticseaicenews/sea-ice-tools/)
Example: Game Playing

• **Tom Mitchell:** Algorithms that
  • improve their **performance** $P$
  • at some **task** $T$
  • with **experience** $E$

• $T =$ playing Chess
• $P =$ win rate against opponents
• $E =$ playing games against itself
Machine Learning Workflow

- Framing an ML problem (Mitchell’s P, T, E)
- Data curation (sourcing, scraping, collection, labeling)
- Data analysis / visualization
- ML Design (hypothesis class, loss function, optimizer, hyperparameters, features)
- Train model
- Validate / Evaluate
- Deploy (and generate new data)
- Monitor performance on new data

Our focus
Types of Learning

• **Supervised learning**
  • **Input:** Examples of inputs and desired outputs
  • **Output:** Model that predicts output given a new input

• **Unsupervised learning**
  • **Input:** Examples of some data (no “outputs”)
  • **Output:** Representation of structure in the data

• **Reinforcement learning**
  • **Input:** Sequence of interactions with an environment
  • **Output:** Policy that performs a desired task
Supervised Learning

• Given \((x_1, y_1), \ldots, (x_n, y_n)\), learn a function that predicts \(y\) given \(x\)

![Graph showing the NSIDC Index of Arctic Sea Ice in September](https://www.flickr.com/photos/gsfc/5937599688/)

Data from [https://nsidc.org/arcticseaicenews/sea-ice-tools/](https://nsidc.org/arcticseaicenews/sea-ice-tools/)
Supervised Learning

- Given $(x_1, y_1), ..., (x_n, y_n)$, learn a function that predicts $y$ given $x$
- **Regression**: Labels $y$ are real-valued

![Arctic Sea Ice Extent](https://www.flickr.com/photos/gsfc/5937399688/)

*Image by NASA Goddard*

*Source: [https://www.flickr.com/photos/gsfc/5937399688/](https://www.flickr.com/photos/gsfc/5937399688/)
Data from [https://nsidc.org/arcticseaindex/](https://nsidc.org/arcticseaindex/)*
Supervised Learning

• Given \((x_1, y_1), \ldots, (x_n, y_n)\), learn a function that predicts \(y\) given \(x\)
• **Classification:** Labels \(y\) are categories

Ocular Tumor (Malignant / Benign)

\(f(x)\)

- **Malignant**
- **Benign**

Predict Benign  Predict Malignant

Tumor Size

Image: https://eyecancer.com/uncategorized/choroidal-metastasis-test/
Supervised Learning

- Given \((x_1, y_1), \ldots, (x_n, y_n)\), learn a function that predicts \(y\) given \(x\)
- Inputs \(x\) can be multi-dimensional

![Graph showing relationship between Age and Tumor Size]

- Patient age
- Clump thickness
- Tumor Color
- Cell type
- ...

Image: https://eyecancer.com/uncategorized/choroidal-metastasis-test/
Unsupervised Learning

• Given $x_1, ..., x_n$ (no labels), output hidden structure in $x$’s
  • E.g., clustering
Unsupervised Learning

Find Subgroups in Social Networks

Identify Types of Exoplanets

Visualize Data

Image Credits:
https://en.wikipedia.org/wiki/Exoplanet
Reinforcement Learning

• Learn how to perform a task from interactions with the environment

• Examples:
  • Playing chess (interact with the game)
  • Robot grasping an object (interact with the object/real world)
  • Optimize inventory allocations (interact with the inventory system)
Reinforcement Learning

https://www.youtube.com/watch?v=iaF43Ze1oel
When should we use machine learning ...? ...
over traditional programming?

<table>
<thead>
<tr>
<th>Analytical Modeling/Understanding</th>
<th>Data Quantity and Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flying rockets to other planets</td>
<td>Adding two numbers</td>
</tr>
<tr>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Checking large prime numbers</td>
<td>Solving differential equations</td>
</tr>
<tr>
<td>NO</td>
<td>YES, SOMETIMES</td>
</tr>
<tr>
<td>Weather forecasting</td>
<td></td>
</tr>
<tr>
<td>MAYBE?</td>
<td></td>
</tr>
<tr>
<td>Predict fashion in 20 years</td>
<td>Recognizing animals from pictures</td>
</tr>
<tr>
<td>NO, PROBABLY</td>
<td>YES!</td>
</tr>
<tr>
<td>Make art and music</td>
<td>Get robots to make sandwiches</td>
</tr>
<tr>
<td>YES!</td>
<td>YES, PROBABLY</td>
</tr>
</tbody>
</table>
Applications of Machine Learning
Everyday Applications

COVID-19 PAYMENT   Spam x

Miller, Jane
to me

This message seems dangerous
It contains a suspicious link that was used to steal people's personal information.

Good morning,
You are advised to download the attached invoice for your review. Please get back to us as soon as possible.

Thanks,
Jane

---

Netfli

Because you watched New Girl

Schitts Creek

Crazy Ex Girlfriend

Amazon

Uber
Scientific Discovery

https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery


http://www.mousemotorlab.org/deeplabcut
Radiology and Medicine

**Input:** Brain scans

![Brain scans](https://www.nature.com/articles/s41746-020-00376-2)

**Output:** Neurological disease labels

![Neurological disease labels](https://www.nature.com/articles/s41573-019-0024-5)

Applications of machine learning in drug discovery and development

https://www.nature.com/articles/s41573-019-0024-5

Deep learning-enabled medical computer vision

Andre Esteve, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean & Richard Socher

https://www.nature.com/articles/s41746-020-00376-2

Machine learning studies on major brain diseases: 5-year trends of 2014–2018

Koji Sakai, Kōei Yamada

---

*Image credit: [Nature](https://www.nature.com)*
Semiconductor Manufacturing

**Input:** light spectra after bouncing off silicon wafer

- $10^{-3}$ nm
- 1 nm
- $10^3$ nm

- X-rays
- UV
- Visible
- Infrared

**Output:** defective / perfect

- EM simulations to recursively match collected spectra

- huge gains in manufacturing throughput

  $\rightarrow$ cheaper phones and computers!

Nanofabrication

Removing individual molecules autonomously with a scanning probe microscope to manufacture new material

RL agent does better than humans!

Autonomous robotic nanofabrication with reinforcement learning

https://advances.sciencemag.org/content/6/36/eabb6987
Creating Music, Art, and Photos

https://thispersondoesnotexist.com/

https://openai.com/blog/jukebox/
Text Generation

SYSTEM PROMPT (HUMAN-WRITTEN)

Recycling is good for the world.

NO! YOU COULD NOT BE MORE WRONG!!

MODEL COMPLETION (MACHINE-WRITTEN, 25 TRIES)

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I’m not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a

https://transformer.huggingface.co/doc/gpt2-large
Ethical Considerations

“The Pennsylvania Board of Probation and Parole has begun using machine learning forecasts to help inform parole release decisions. In this paper, we evaluate the impact of the forecasts on those decisions and subsequent recidivism.”

“In 2013, the University of Texas at Austin’s computer science department began using a machine-learning system called GRADE to help make decisions about who gets into its Ph.D. program”

“Videos about vegetarianism led to videos about veganism. Videos about jogging led to videos about running ultramarathons. It seems as if you are never ‘hard core’ enough for YouTube’s recommendation algorithm. It promotes, recommends and disseminates videos in a manner that appears to constantly up the stakes. Given its billion or so users, YouTube may be one of the most powerful radicalizing instruments of the 21st century.”
Danger of Out-of-Domain Machine Learning

Any time you are evaluating on data “far” from your training data, beware!