Lecture 1: Introduction

CIS 4190/5190 (Fall 2022)
August 31
CIS 4190/5190: Applied Machine Learning

- **Course goals**
  - Identify opportunities for applying machine learning (ML) algorithms
  - Train ML models
  - 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
Agenda

• Logistics
  • Course description
  • Course policies
  • Tentative Schedule
  • Grading

• Introduction
  • Motivation
  • Basic definitions
  • Examples
Instructors

Prof. Osbert Bastani

Prof. Zachary Ives
Teaching Assistants

• Brian Chen
• Swati Gupta
• Jio Jeong
• Jinhui Luo
• Jason Ma
• Rushab Manthripragada
• Ramya Ramalingam
• Peizhi Wu
Course Website

• Course website
  • https://www.seas.upenn.edu/~cis5190/fall2022/
  • Syllabus
  • Tentative schedule
  • Links to all assignments
  • Resources
Office Hours

• Each TA and instructor will have 1-2 hours of office hours each week
  • Times still being decided

• We are aiming to have a mix of in person and remote office hours
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
Attendance Policy

• **We expect all students to attend classes regularly**
  • Weekly quizzes designed to make sure students are following course material

• Please **do not come to class** if you are not feeling well or if you test positive for Covid-19
  • We will provide course recordings and lecture notes for students unable to make it to class
  • These will be available on Canvas (though we reserve the right to change our policy if attendance becomes an issue)
Masking Policy

• There are some students in the class who are at higher risk from COVID exposure

• Thus, we are setting a policy that we will enforce wearing a mask in class and during office hours

• We appreciate your consideration of your classmates
Waitlist Policy

• We are considering students roughly in the order (up to the cap):
  
  1. Absolutely need to take this course during this semester (Fall 2022) for a graduation requirement
  2. Absolutely need to take this course during next semester (Spring 2023) for a graduation requirement
  3. Graduating in this semester (Fall 2022)

• In addition, you must satisfy prerequisites and have correctly answered all waitlist questions (https://waitlist.cis.upenn.edu)
Waitlist Policy

• If you satisfy one of these categories, email me with which category you are in and a clear explanation for why you are in that category
  • Email: obastani@seas.upenn.edu
  • Deadline: Thursday at 8pm

• Final decisions at the discretion of the instructors
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

Also see: 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 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
## Tentative Schedule

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

• Homeworks (6×): 30%
• Project (3 milestones, teams of 3): 20%
• Final exam (during exam week): 35%
• Quizzes (12 ×, roughly weekly): 10%
  • 50% correct sufficient for full credit
• Class participation: 5%
Grading Scheme

• A+: 90+
• A: 85-90
• A-: 80-85
• B+: 75-80
• B: 70-75
• B-: 65-70
• Lower passing grades: 50-65

• May be curved up
Late Policy

• 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 late by 20 hours
  • 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
Homework Schedule

• 6 homeworks
  • Released every other Wednesday
  • Due Wednesday 2 weeks later (with an exception for HW 2)
  • Due at 8pm
Homework Structure

• **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 or code; acknowledge all your discussions at beginning of your report
Homework 1

• Designed to test mathematical background
• Opportunity to get used to the workflow
• Full points if you score 50% or more
• We will not answer questions about the HW 1 (except clarifying questions)
Quiz Schedule

• **12 quizzes**
  - Released every Wednesday (starting next week)
  - Due Thursday 8 days later

• Checks basic understanding of material covered the previous week
Agenda

• **Logistics**
  • Course description
  • Course policies
  • Tentative Schedule
  • Grading

• **Introduction**
  • Motivation
  • Basic definitions
  • Examples
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: Game Playing

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

• $T = \text{playing Checkers}$
• $P = \text{win rate against opponents}$
• $E = \text{playing games against itself}$
Example: Prediction

Photo by NASA Goddard

Image: https://www.flickr.com/photos/gsfc/5937499688/
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 $f$

New input

Predicted output
Example: Prediction

![Image of Arctic Sea Ice](https://www.flickr.com/photos/gsfc/5937599688/)

**NSIDC Index of Arctic Sea Ice in September**

- Photo by NASA Goddard
- Data from [https://nsidc.org/arcticseaicenews/sea-ice-tools/](https://nsidc.org/arcticseaicenews/sea-ice-tools/)
Our focus

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
Types of Learning

• **Supervised learning**
  • **Input:** Examples of inputs and outputs
  • **Output:** Model that predicts unknown 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 \(x\)

![Image: https://www.flickr.com/photos/gsfc/5937599688/](https://www.flickr.com/photos/gsfc/5937599688/)

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

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

• **Regression**: Labels \(y\) are real-valued

---

Image: https://www.flickr.com/photos/gsfc/5937599688/
Data from https://nsidc.org/arcticseaicenews/sea-ice-tools/
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)

- \(f(x)\)
- Tumor Size

- Malignant
- Benign

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

![Diagram of Age vs. Tumor Size with data points]

- 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
Applications of Machine Learning
Everyday Applications

COVID-19 PAYMENT

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
Radiology and Medicine

**Input:** Brain scans

**Output:** Neurological disease labels

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 Esteves, 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 Sakai1 · Kel Yamada1
Scientific Discovery

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

Creating Images & Text

https://thispersondoesnotexist.com/

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

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
# When should we use machine learning?

<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>NO</td>
</tr>
<tr>
<td>Checking large prime numbers</td>
<td>NO</td>
</tr>
<tr>
<td>Adding two numbers</td>
<td>NO</td>
</tr>
<tr>
<td>Solving differential equations</td>
<td>YES, SOMETIMES</td>
</tr>
<tr>
<td>Weather forecasting</td>
<td>MAYBE?</td>
</tr>
<tr>
<td>Recognizing animals from pictures</td>
<td>YES!</td>
</tr>
<tr>
<td>Predict fashion in 20 years</td>
<td>NO, PROBABLY</td>
</tr>
<tr>
<td>Make art and music</td>
<td>YES!</td>
</tr>
<tr>
<td>Get robots to make sandwiches</td>
<td>YES, PROBABLY</td>
</tr>
<tr>
<td>Predict fashion in 20 years</td>
<td>NO, PROBABLY</td>
</tr>
<tr>
<td>Weather forecasting</td>
<td>MAYBE?</td>
</tr>
</tbody>
</table>
Danger of Out-of-Domain Machine Learning

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

<table>
<thead>
<tr>
<th>Week</th>
<th>Content</th>
<th>Homework</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Linear Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Linear Regression, Cont.</td>
<td></td>
<td>HW 1 Due</td>
</tr>
<tr>
<td>4</td>
<td>Logistic Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Decision Trees</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Neural Networks</td>
<td>HW 2 Due</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Computer Vision</td>
<td></td>
<td>Milestone 1 Due</td>
</tr>
<tr>
<td>8</td>
<td>Bayesian Networks</td>
<td>HW 3 Due</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Unsupervised Learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Natural Language Processing</td>
<td>HW 4 Due</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Reinforcement Learning</td>
<td></td>
<td>Milestone 2 Due</td>
</tr>
<tr>
<td>12</td>
<td>Recommender Systems</td>
<td>HW 5 Due</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Ethics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Ethics</td>
<td>HW 6 Due</td>
<td>Milestone 3 Due</td>
</tr>
<tr>
<td>15</td>
<td>Review</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Review</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
First Assignments

• **HW 1**: Released today, **due 9/14**  
  • No office hours planned for HW 1  
  • 50% = full credit

• **Quiz 1**: Released 9/7, **due 9/15**