12 January 2023

CIS 5200: MACHINE LEARNING

INTRODUCTION AND OVERVIEW

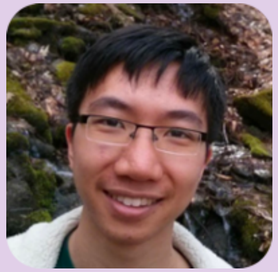
Surbhi Goel and Eric Wong

Spring 2023

STAFF

INSTRUCTORS


Surbhi Goel


Eric Wong

TEACHING ASSISTANTS


Abhinav Atrishi

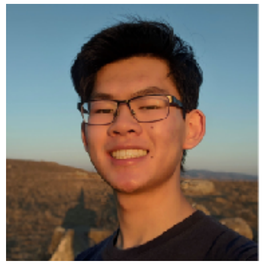

Jordan Hochman


Bowen Jiang


Pavlos Kallinikidis


William Liang


Heyi Liu


David LuoZhang


Aryan Nagariya


Jeffrey Pan


Aditya Singh


Tianyi Wei

Head TAs


Keshav Ramji


Wendi Zhang

PREREQUISITES - WHAT SHOULD YOU KNOW?

The course will have a significant mathematical component along with programming

- Undergraduate level training or coursework on linear algebra, (multivariate) calculus, and basic probability and statistics
- Basic programming in Python
- Undergraduate level training or coursework in the analysis of algorithms

It is okay if you do not have this

Homework 0 is out!

Great way to check if you have the required prerequisites

OTHER COURSES & WAITLIST - IS THIS COURSE FOR YOU?

CIS 5200	Machine Learning	+Math, +Theory
CIS 4190/5190	Applied Machine Learning	+Programming
CIS 5450	Big Data Analytics	+Data
ESE 5450	Data Mining	+Data, +Math

- Remaining waitlist priority for those that complete HW 0
- Struggling with HW0? Consider taking some prerequisites or an alternative course that better fits your background
- Course is offered every semester

COURSE GOALS - WHAT WILL YOU LEARN?

By the end of the semester, you should

- be familiar with a variety of ML methods, both classical and modern
- understand the math behind them
- know their strengths and drawbacks
- be able to use them in practice
- take more advanced ML courses

Course will be fast-paced!

Lectures will help you learn the material

Homeworks will help you test your understanding

LOGISTICS - WHAT WILL THE COURSE LOOK LIKE?

Lectures:

- live not recorded, slides and notes will be made available after

Recitations:

- Some TA office hours will operate as supplementary recitations
- Sign up for a designated recitation slot (coming soon)
- First half will be recitation, then convert to normal office hours
- Optional attendance, announced weekly

Office Hours:

- ~20 over the week, some weeks will be recitation style
- time/location on the course website

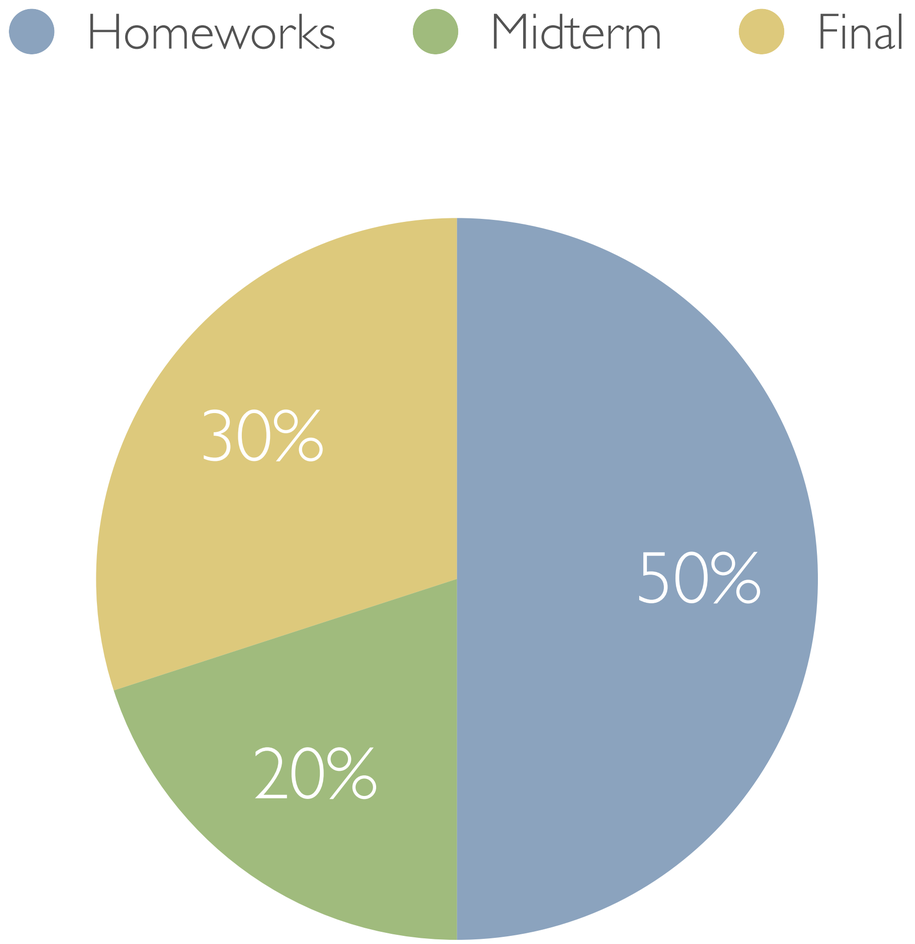
LOGISTICS - WHAT WILL YOU DO?

Homeworks:

- 6 homeworks - HW0 (2% for completing), HW1-HW4 (10% each), HW5 (8%)
- conceptual (in latex) and coding (in ipython notebook), submitted via Gradescope

Exams:

- Midterm (20%) - during class on 2/23
- Final (30%) - TBA during finals week 5/1-5/9



LOGISTICS - WHERE TO FIND ANSWERS?

Course website: https://machine-learning-upenn.github.io/

- refer for logistics, schedule, policies, slides, homework etc.
- office hour time/location

Gradescope/Canvas:

- submit homework, check grades

Ed:

- main channel for communication
- ask/answer questions, reach out to instructors or TAs (when necessary)

POLICIES - COLLABORATION

Collaboration:

- encouraged and permitted, though each student must write their **own** solutions
- give credit, acknowledge who all you collaborated with

Honor Code:

- encouraged to use online resources for learning
- don't try to find solutions to homeworks, for e.g. previous year's solutions, chatGPT, StackOverflow, your friend's answers, etc.
- cheating will be punished according to university regulations, **zero tolerance**

Collaborate, don't cheat!

Homeworks and exams are for you to get better with the material

POLICIES - LATE DAYS AND REGRADING

Homeworks:

- 144 hours (6 days), max 48 hours (2 days) per homework, zero credit after
- Gradescope tracks late time but does not enforce a max (will be calculated at end of semester)

Regrading:

- submit your request within 1 week from the date when grades are released
- all requests must be made through Ed with a written explanation of the concern
- will not be entertained during the OHs or lectures.

POLICIES - SAFETY AND WELLBEING

Masking:

- for the health and safety of everyone, we are requiring that all students wear a mask during class and office hours
- if you are feeling sick, we encourage you to stay at home and recover.
- we will reevaluate as things change

Wellness and Inclusion:

- we value and actively seek to include all of you and your unique identities
- your mental health and wellbeing are incredibly important to us
- if you experience any challenges, please reach out to us, **we are ready to help**
- see course website for a list of resources

FEEDBACK - WHAT IF THINGS AREN'T OKAY?

We are new to this, we are learning!

Give us **lots of feedback**, we want to make the course better for you

You can use Ed to send us anonymous feedback, or come chat with us

We care!

Help us improve

NOW TO THE FUN PART!

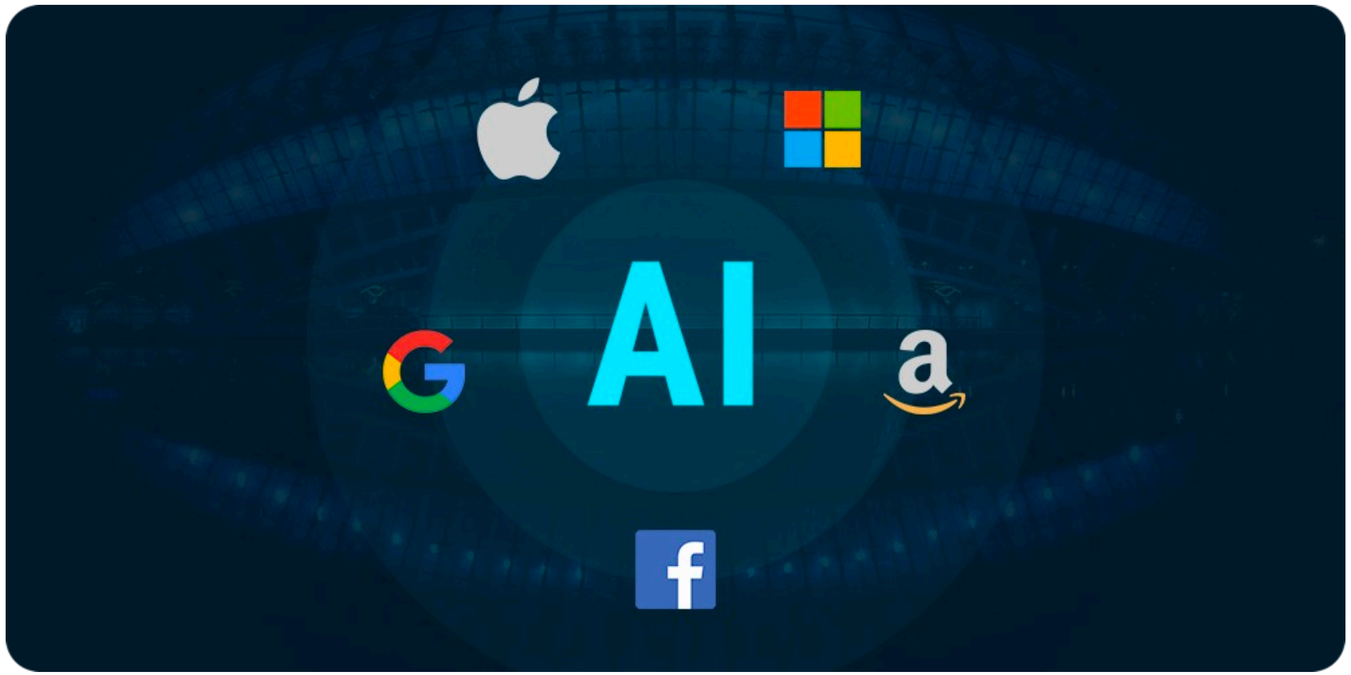
Let's begin!

The following draws from slides by Vatsal Sharan (USC), Aarti Singh (CMU), Christopher De Sa (Cornell)

WHO USES MACHINE LEARNING? - TECH INDUSTRY

The usual suspects:

- Google
- Meta
- Microsoft
- OpenAI
- Amazon
- Many more…



RECENT DEVELOPMENTS - TEXT & IMAGE GENERATION

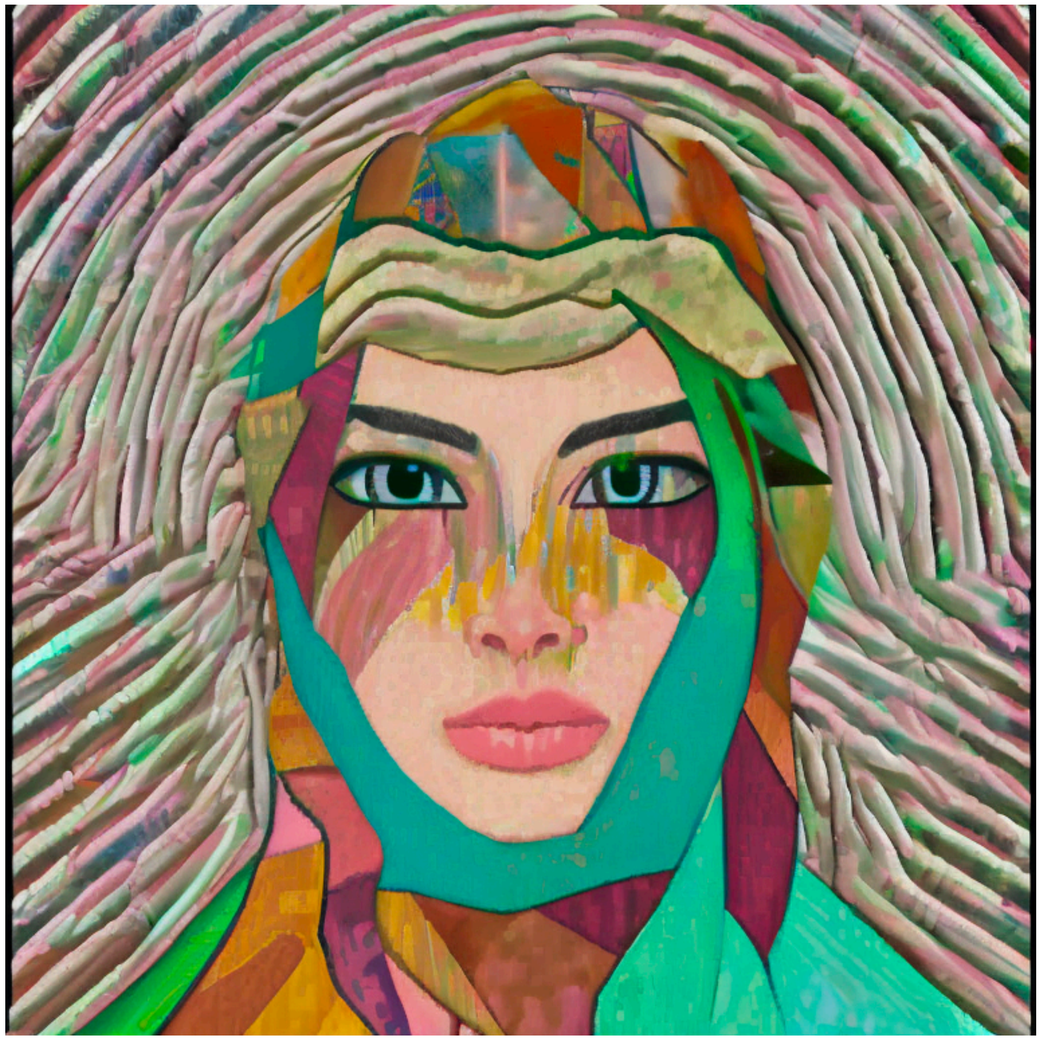
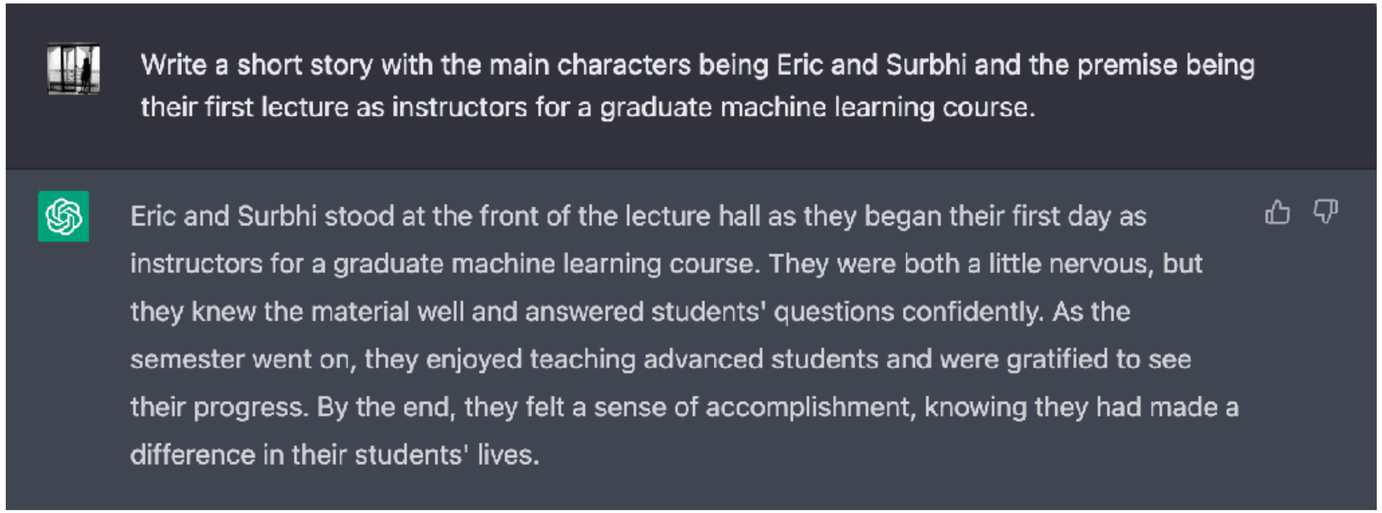
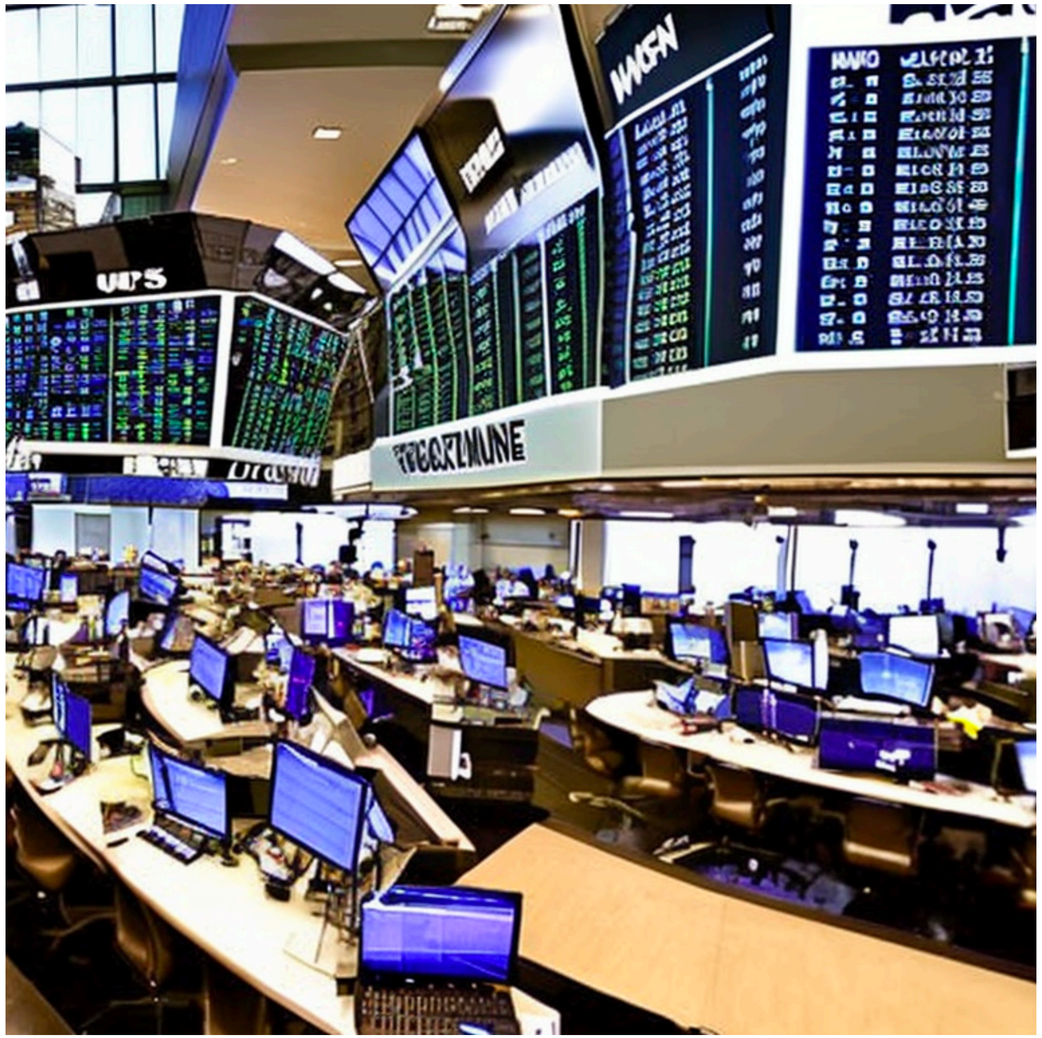


Image generation (Dall-E)



Text generation (ChatGPT)

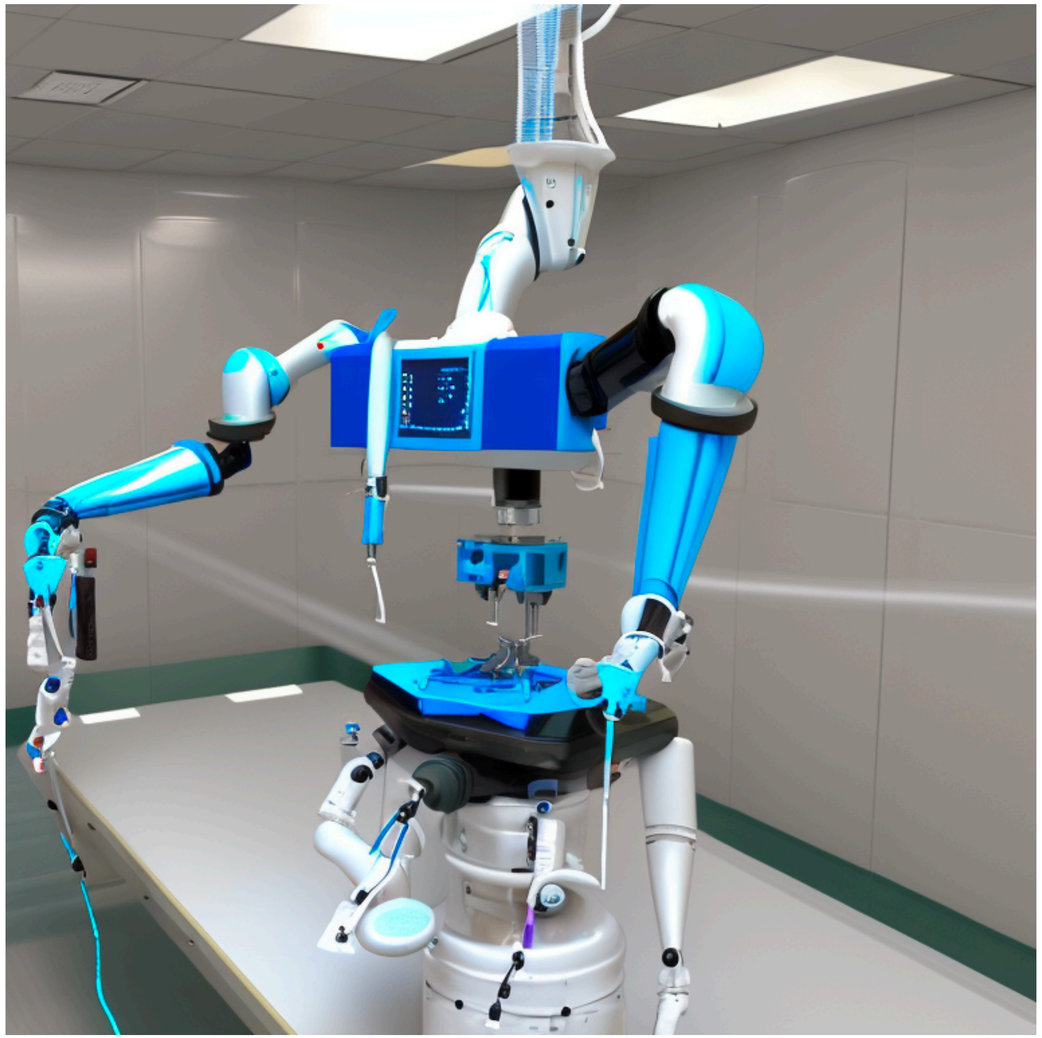
WHO USES MACHINE LEARNING? - FINANCE



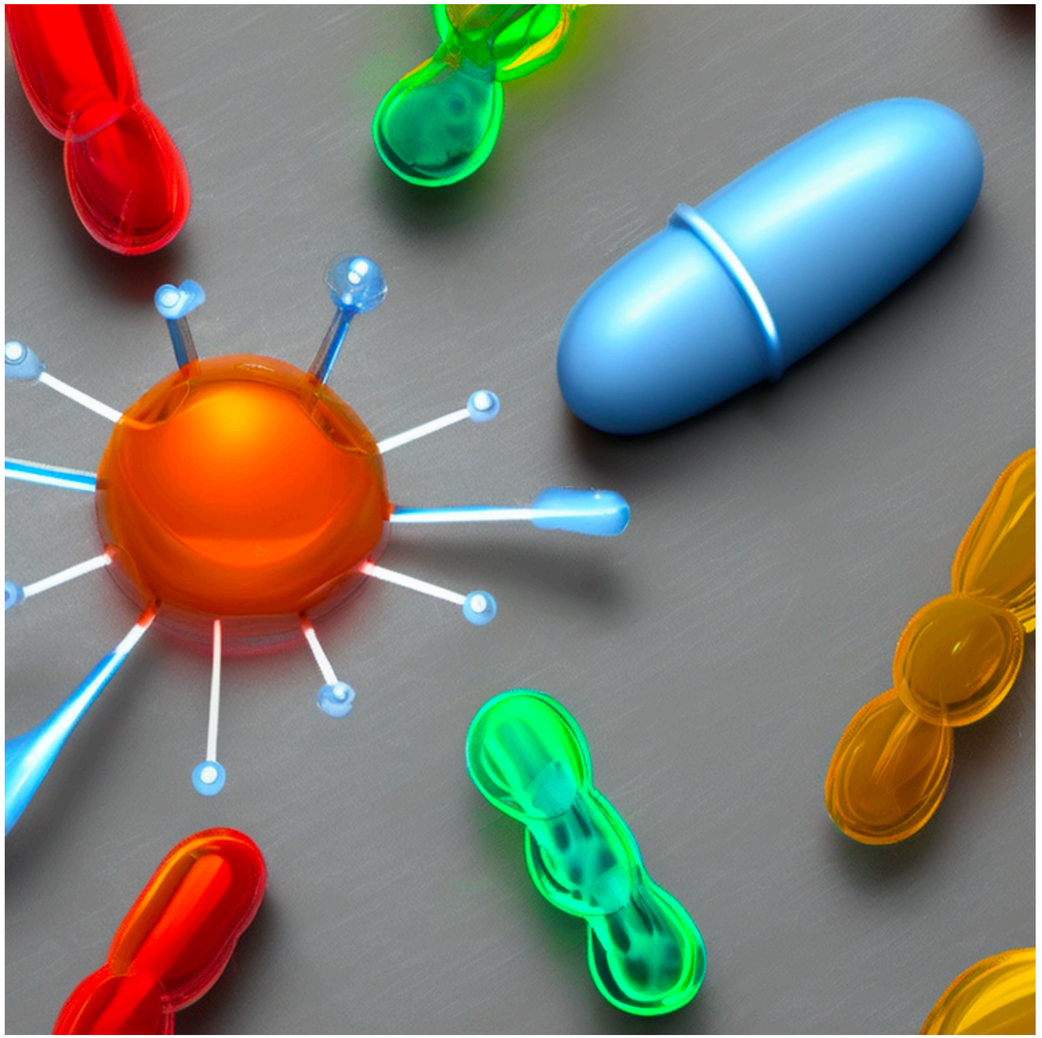
Algorithmic trading

Credit lending

WHO USES MACHINE LEARNING? - HEALTHCARE



Surgical robots



Drug discovery

WHO USES MACHINE LEARNING? - MANUFACTURING

Production lines

Logistics

WHO USES MACHINE LEARNING? - AUTONOMOUS VEHICLES

Unmanned Aerial Vehicles

Submarines

CHALLENGES OF ML - UNDERSTANDING FUNDAMENTALS

NHTSA deepens its probe into Tesla collisions with stationary emergency vehicles

The agency added six more incidents since the investigation started.

Spencer Platt via Getty Images

AI May Be More Prone to Errors in Image-Based Diagnoses Than Clinicians

New research indicates that AI may be more prone to making mistakes than humans in image-based medical diagnoses because of the features they use for analysis.

Source: Getty Images

Forbes

Armed With ChatGPT, Cybercriminals Build Malware And Plot Fake Girl Bots

Thomas Brewster Forbes Staff

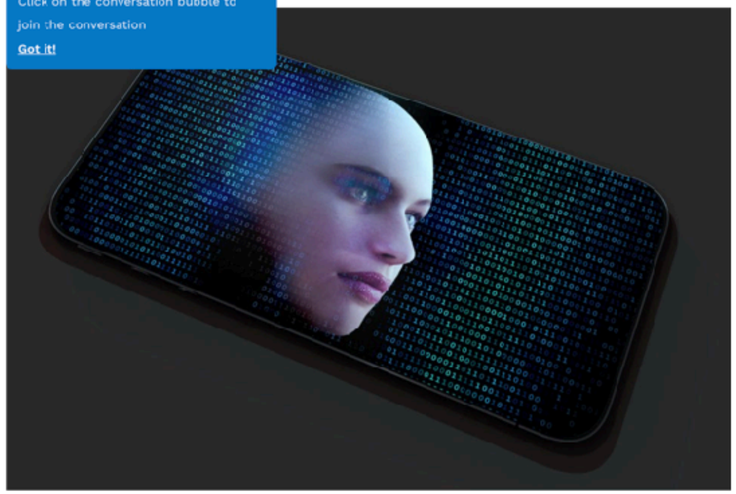
Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

Follow

1

Jan 6, 2023, 12:06pm EST

New!
Click on the conversation bubble to join the conversation
Got it!



Hackers are testing ChatGPT's ability to create female chatbots as part of their efforts to scam men attracted to the digital persona. GETTY

To fix ML systems, we need to understand them first!

CHALLENGES OF ML - HARD TO UNDERSTAND

ML systems are trending towards **opaque** and **black box**

The process of building ML systems have been criticized as resembling **alchemy**

We can use **mathematics**, **theory**, and/or **compute** to understand and research ML

WHAT IS MACHINE LEARNING? - OVERVIEW

Design and analysis of algorithms that

- Improve **performance**
- at some **task**
- given **experience**

In many cases, the learning algorithm is optimization

MACHINE LEARNING - TASKS

Supervised learning:

- Regression, classification

Unsupervised learning:

- Density estimation, clustering, dimensionality reduction

Semi-supervised learning

Active learning

Online learning

Reinforcement learning

And so on...

MACHINE LEARNING - SUPERVISED LEARNING

Predict future outcomes based on past outcomes

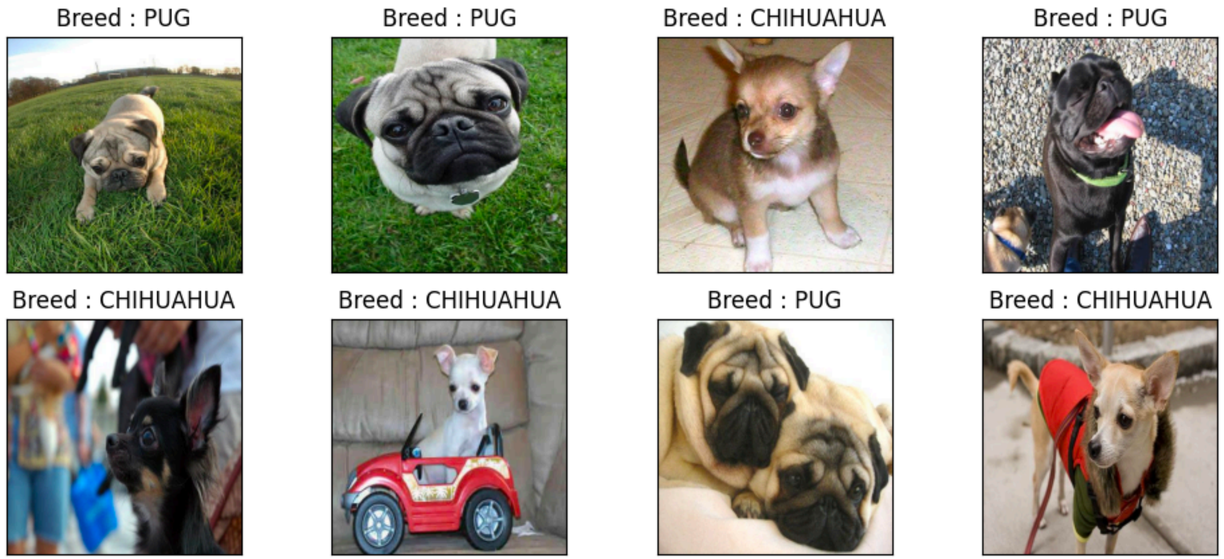


Image classification

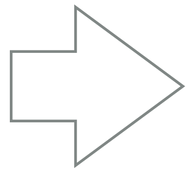
Stock prediction

MACHINE LEARNING - SUPERVISED LEARNING

Inputs $x \in \mathcal{X}$

Labels $y \in \mathcal{Y}$

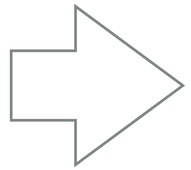
($\mathcal{Y}$ = Breeds)
"Pug"
"Chihuahua"



Classification

Discrete labels

($\mathcal{Y}$ = Stock prices)
"$130.02"



Regression

Continuous labels

Task: Learn predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

SUPERVISED LEARNING - SETUP

Input space: $\mathcal{X} \subset \mathbb{R}^d$

- Features in d dimensions
- e.g. images → pixel values, stock price → open value, close value, etc.

Output space: $\mathcal{Y}$

- $\mathcal{Y} = \{0,1\}$ for dog breed classification where 1 is "pug" and 0 is "chihuahua"
- $\mathcal{Y} = \mathbb{R}_+$ for stock price prediction

Predictor function: $f : \mathcal{X} \to \mathcal{Y}$

How do we select the predictor?

SUPERVISED LEARNING - LOSS FUNCTION

Loss function

* $\ell(f(x), y)$ as a measure of how good/bad the prediction is

Examples

0-1 loss for classification: $$\ell(f(x), y) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise.} \end{cases}$$

square loss for regression: $$\ell(f(x), y) = (f(x) - y)^2$$

SUPERVISED LEARNING - TRAINING

Training Dataset

* Set of labelled examples $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Empirical Risk Minimization:

With the given loss function, we find a good predictor by minimizing loss on the training dataset

$$\hat{f} = \arg\min_{f} \underbrace{\frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i)}_{\hat{R}(f)}$$

Empirical Risk

SUPERVISED LEARNING - QUESTION

If I give you the following dataset $\{(0,1),(1,2),(2,3)\}$ and ask you, what is the value at $x = 3$, would you be able to answer this?

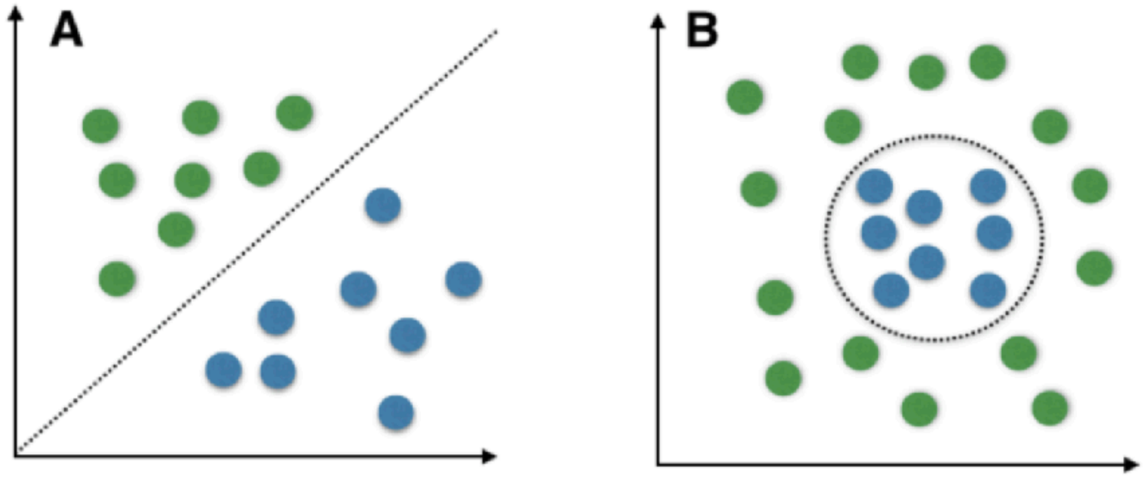
It is impossible to know what this would be unless we have some assumptions!

The No Free Lunch Theorem states that every successful ML algorithm must make assumptions, no one algorithm works for all instances

SUPERVISED LEARNING - HYPOTHESIS CLASS

Hypothesis class / Model family / Function class

- Encodes inductive bias (assumptions) about the prediction function
- We use prior knowledge about the problem to choose hypothesis class $\mathscr{F}$



A: Linear separator

$$f(x) = \textbf{sign}(w^\top x + b)$$

$\mathscr{F}$ contains all linear separators

B: Non-linear separator

$$f(x) = \textbf{sign}(\|x - \theta\| - r)$$

$\mathscr{F}$ contains all ball separators

Image source: https://www.codingninjas.com/codestudio/library/linear-vs-non-linear-classification

SUPERVISED LEARNING - TRAINING

Training Dataset

- Set of labelled examples $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Empirical Risk Minimization:

With the given loss function and **hypothesis class**, we find a good predictor by minimizing loss on the training dataset

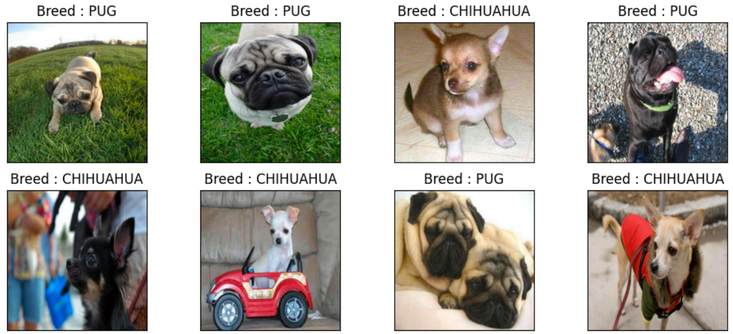
$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i)$$

Search is restricted to a hypothesis class

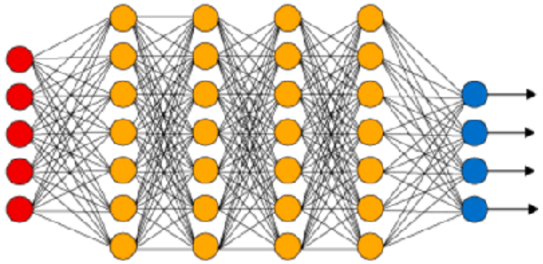
SUPERVISED LEARNING - PUTTING IT TOGETHER

Training dataset

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$



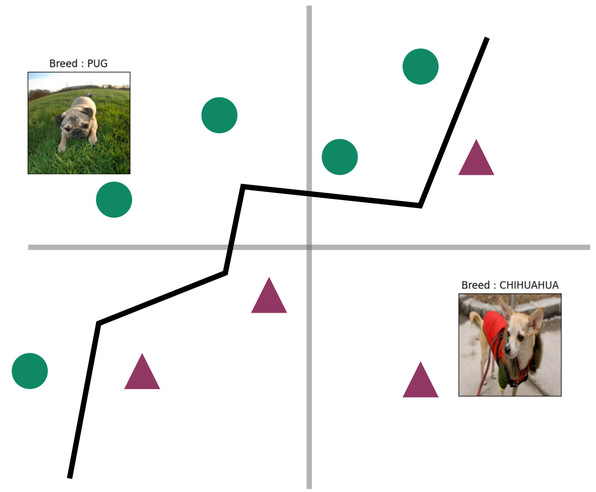
Hypothesis class $\mathcal{F}$



Minimize loss on training data

$$\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i)$$

Prediction function $\hat{f}$



How do we evaluate performance of predictor $\hat{f}$?

SUPERVISED LEARNING - BAD EXAMPLE

Memorizer predictor $f_{\text{mem}}(\cdot)$

$$f_{\text{mem}}(x) = \begin{cases} y_i & \text{if } \exists (x_i, y_i) \in \mathcal{S}, x = x_i, \\ 0 & \text{otherwise.} \end{cases}$$

This gets 0 training loss $\hat{R}(f_{\text{mem}}) = 0$, but do you think it is a good predictor?

SUPERVISED LEARNING - GENERALIZATION

We want the predictor to perform well not just on the training data but on examples it will see in the future.

How do we formalize this?

(i.i.d. assumption) Training dataset is drawn independently and identically from some unknown but fixed distribution $\mathscr{D}$

loss on future examples = loss over the distribution

$$R(\hat{f}) = \mathbb{E}_{(x,y)\sim\mathscr{D}}\left[\ell(\hat{f}(x), y)\right]$$

Recall, our predictor minimized loss on training set, $\hat{R}(\hat{f}) = \frac{1}{m}\sum_{i=1}^{m}\ell(\hat{f}(x_i), y_i)$

SUPERVISED LEARNING - GENERALIZATION

loss on future examples = loss over the distribution

$$R(\hat{f}) = \mathbb{E}_{(x,y)\sim\mathscr{D}}\left[\ell(\hat{f}(x), y)\right]$$

How do we estimate this?

We split the data into

- Training data - the subset on which we train the model
- Test data - the subset on which we evaluate the model

Test data is unseen during learning,
therefore good estimate of generalization performance

SUPERVISED LEARNING - FAILURES

What can go wrong? When do we fail to learn?

Underfitting

Error on training set $\hat{R}(\hat{f})$ is high; leads to high error on new samples $R(\hat{f})$

Overfitting

Error on training set $\hat{R}(\hat{f})$ is low; but error on new samples $R(\hat{f})$ is high

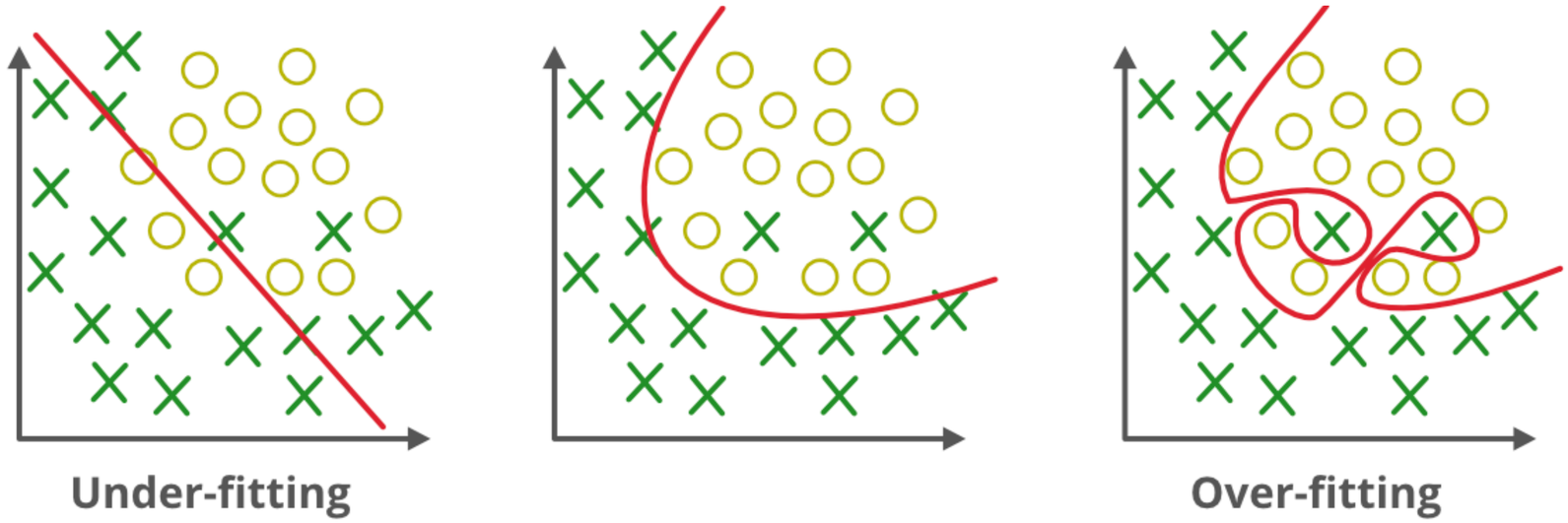


Image source: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

SUPERVISED LEARNING - SUMMARIZED

Loss function: What is the right loss function for the task?

Representation: What class of functions should we use for the task?

Optimization: How can we efficiently solve the empirical risk minimization?

Generalization: Will the predictor perform well on unseen data?

All are related, and data is the fuel!