18-661: Introduction to ML for Engineers

Course Overview

Spring 2025

ECE – Carnegie Mellon University

18-661: Introductory ML for Engineers

Welcome!

New students: Welcome to CMU and to this intro to ML class

18-661: Introductory ML for Engineers

Welcome!

New students: Welcome to CMU and to this intro to ML class

About this class

- Introductory Machine Learning for Engineers, assumes no prior background in machine learning
- Offered every semester since Fall 2018
- Offered across the Pittsburgh, Silicon Valley and Rwanda campuses
- The undergraduate version 18-461 has the same lectures and assignments as 18-661, but a separate (easier) grading curve

Registration

- If you're not registered, we encourage you to stay patient
- You are welcome to keep attending the lectures until the waitlists are sorted out

Direct all waitlist-related questions to:
ece-waitlists@andrew.cmu.edu

Course Prerequisites

- Probability theory
- Linear algebra
- Calculus: Differentiation, Integration, Convexity
- Python programming, in particular, `numpy`

If you don't satisfy these pre-requisites, we strongly encourage you to take the class after reviewing introductory material (see readings for Lecture 2).

Instructors & TAs

- Carlee Joe-Wong, Instructor
- Gauri Joshi, Instructor

Instructors & TAs

- Carlee Joe-Wong, Instructor
- Gauri Joshi, Instructor

- Neharika Jali, TA (Pitt)
- Jong-Ik Park, TA (Pitt)
- Arian Raje, TA (Pitt)
- Siddharth Shah, TA (Pitt)
- Steven Zeng, TA (Pitt)
- Landelin Gihozo, TA (Kigali)
- John Waithaka, TA (Kigali)

Credit & thanks to:

- Virginia Smith, CMU
- Yuejie Chi, CMU
- Pulkit Grover, CMU
- Anit Sahu and Joao Saude, CMU
- Guannan Qu, CMU
- Ameet Talwalkar, CMU
- Fei Sha, USC
- Emily Fox, Stanford

Outline

1. What is Machine Learning?
2. Course Goals
3. Course Logistics
4. Probability Review
5. A Simple Learning Problem: MLE/MAP Estimation

What is Machine Learning?

Let's ask!

JO

What is machine learning?

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. There are various techniques and algorithms used in machine learning, including decision trees, neural networks, and Bayesian methods.

Let's ask!

JO What is machine learning?

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. There are various techniques and algorithms used in machine learning, including decision trees, neural networks, and Bayesian methods.

- We will cover decision trees, neural networks, and Bayesian methods in this course.

Let's ask!

JO What is machine learning?

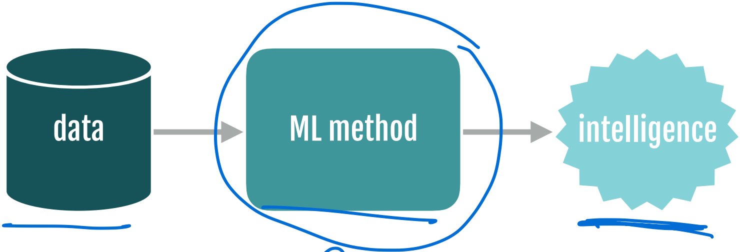
Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. There are various techniques and algorithms used in machine learning, including decision trees, neural networks, and Bayesian methods.

- We will cover decision trees, neural networks, and Bayesian methods in this course.
- This is generated from ChatGPT: `https://openai.com/blog/chatgpt/`
- ChatGPT has learned to answer questions (make decisions), based on observing text from the Internet (data).

A More Concrete Definition

- **Machine learning is**: the study of methods that *improve their performance on some task with experience*
- Can you concretely define performance and experience for the tasks shown in the pictures?

Machine Learning Pipeline



We are going to study this block in great detail

Examples

Task 1: Regression

How much should you sell your house for?

Task 1: Regression

How much should you sell your house for?

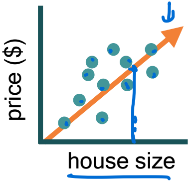
Task 1: Regression

How much should you sell your house for?

input: houses & features

Task 1: Regression

How much should you sell your house for?



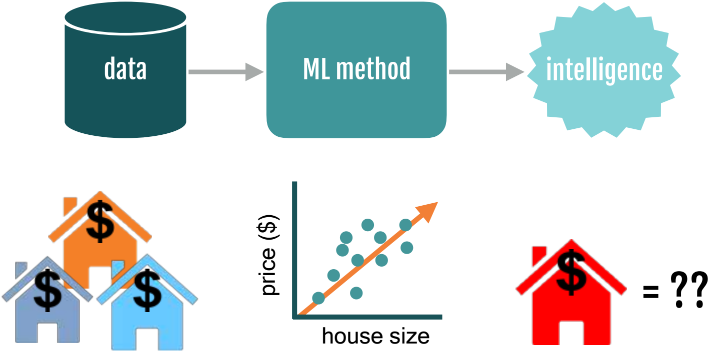
input: houses & features **learn**: $x \rightarrow y$ relationship

house features

price

Task 1: Regression

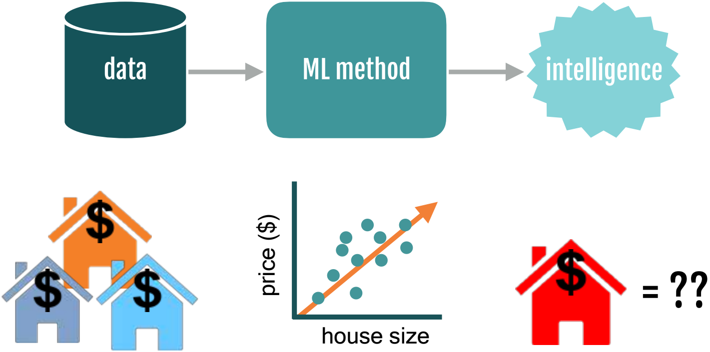
How much should you sell your house for?



input: houses & features **learn**: $x \rightarrow y$ relationship **predict**: y (*continuous*)

Task 1: Regression

How much should you sell your house for?



input: houses & features **learn**: $x \to y$ relationship **predict**: y (*continuous*)

Course Covers: Feature Scaling, Linear/Ridge Regression, Loss Function, (Stochastic) Gradient Descent, Regularization, Cross Validation

Task 2: Classification

Cat or dog?

Task 2: Classification

Cat or dog?

input: cats and dogs

Task 2: Classification

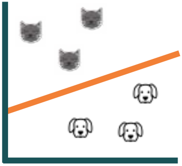
Cat or dog?

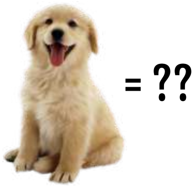
input: cats and dogs

learn: $x \to y$ relationship

Task 2: Classification

Cat or dog?





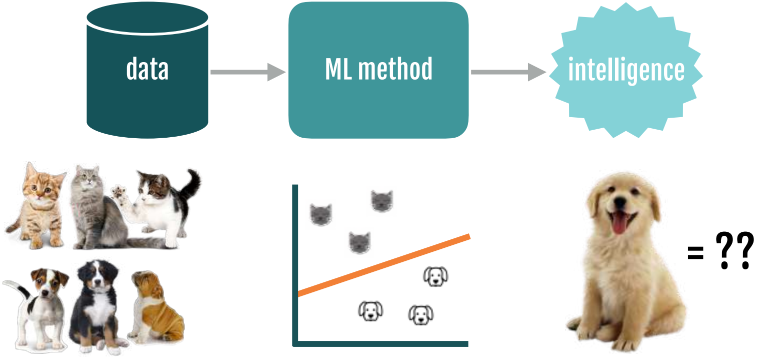
input: cats and dogs

learn: $x \to y$ relationship

predict: y (*categorical*)

Task 2: Classification

Cat or dog?



input: cats and dogs **learn**: $x \to y$ relationship **predict**: y (*categorical*)

Course Covers: Naïve Bayes, Logistic Regression, SVMs, Neural Nets, Decision Trees, Boosting, Nearest Neighbors

Task 3: Clustering

How to segment an image?

Task 3: Clustering

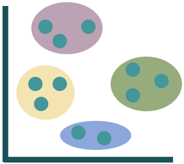
How to segment an image?

input: raw pixels $\{x\}$

Task 3: Clustering

How to segment an image?

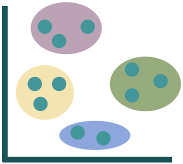
input: raw pixels $\{x\}$



separate: $\{x\}$ into sets

Task 3: Clustering

How to segment an image?

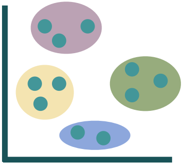


input: raw pixels $\{x\}$ **separate**: $\{x\}$ into sets **output**: cluster labels $\{z\}$

Task 3: Clustering

How to segment an image?

input: raw pixels $\{x\}$



separate: $\{x\}$ into sets

output: cluster labels $\{z\}$

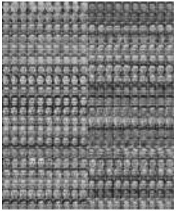
Course Covers: K-means, K-means++, GMM clustering

Task 4: Embedding

How to efficiently represent data?

Task 4: Embedding

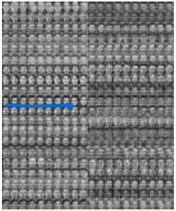
How to efficiently represent data?

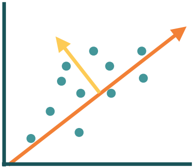


input: large dataset $\{x\}$

Task 4: Embedding

How to efficiently represent data?



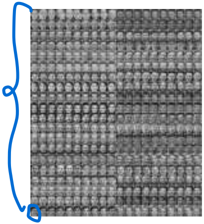


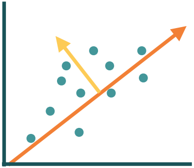
input: large dataset $\{x\}$

find: sources of variation

Task 4: Embedding

How to efficiently represent data?



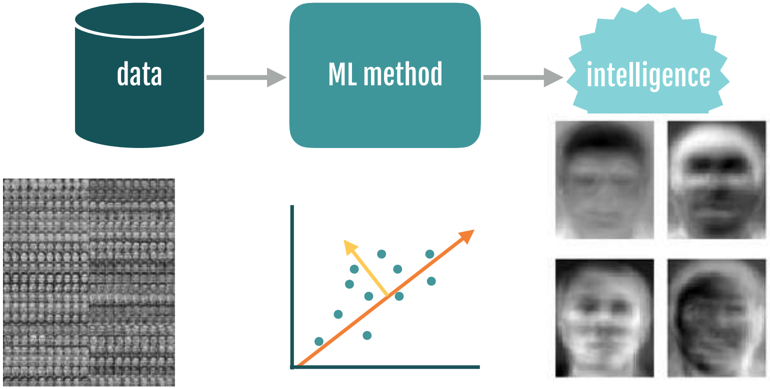


input: large dataset $\{x\}$ **find**: sources of variation **return**: representation $\{z\}$

$$\text{Image } 1 = 0.5 \times B_1 + 0.2\, B_2 + 0.3\, B_3 + 0.1\, B_4$$

Task 4: Embedding

How to efficiently represent data?



input: large dataset $\{x\}$ **find**: sources of variation **return**: representation $\{z\}$

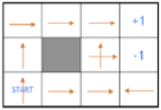
Course Covers: Dimensionality Reduction, PCA

Task 5: Reinforcement Learning

How to take the actions that maximize reward?

Task 5: Reinforcement Learning

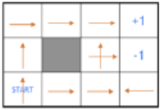
How to take the actions that maximize reward?

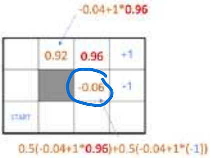


input:
state-action-reward
trajectories

Task 5: Reinforcement Learning

How to take the actions that maximize reward?



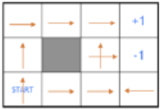


input:
state-action-reward trajectories

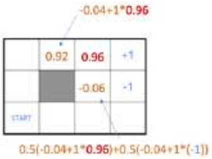
find: Value function or Q function

Task 5: Reinforcement Learning

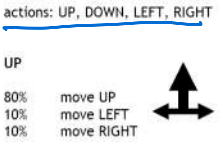
How to take the actions that maximize reward?



input:
state-action-reward trajectories



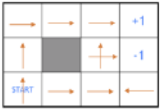
find: Value function or Q function



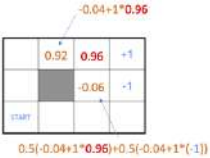
return: Optimal Policy π

Task 5: Reinforcement Learning

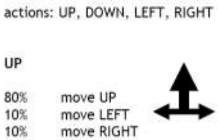
How to take the actions that maximize reward?



input:
state-action-reward trajectories



find: Value function or Q function

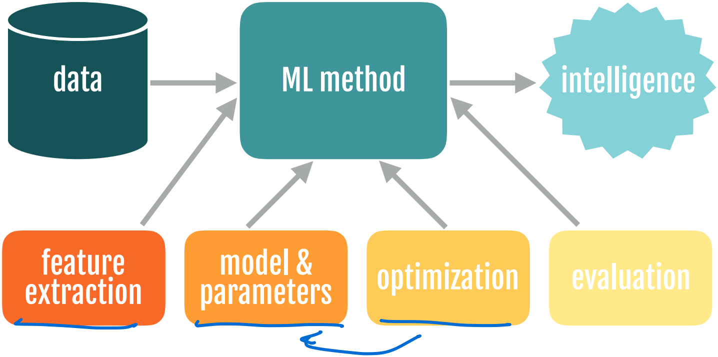


return: Optimal Policy π

Course Covers: Online learning and bandits, Bellman equation, Policy Evaluation, Q-learning

Course Goals

Goal of the Course



Equip you with the *tools* to *develop* and *deploy* machine learning for engineering applications

- Fundamental Understanding: Algorithms, Theoretical Analysis
- Applications: Implementation in Python, PyTorch

Key Topics

Models

Mapping
$x \to y$

- Linear and Ridge Regression
- Linear classification: logistic regression, SVM
- Nonlinear models: kernels, neural networks & deep learning, decision trees
- Nearest neighbors, clustering
- Graphical Models

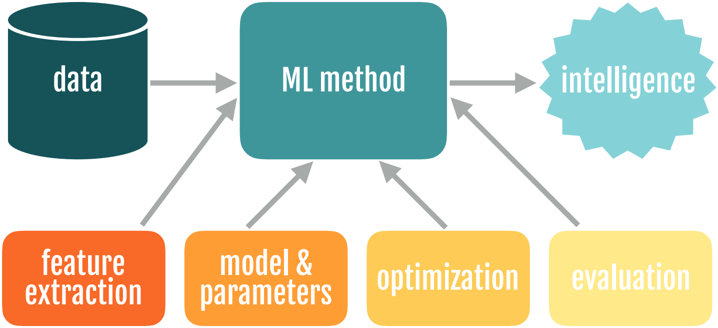
Methods

- Gradient descent
- Boosting
- k-means
- PCA
- EM

Concepts

- Point estimation, MLE, MAP
- Loss functions, bias-variance tradeoff, cross-validation
- Sparsity, overfitting, model selection
- Types of ML (supervised, unsupervised, reinforcement)

Similar Courses



- Most similar CMU Courses are 10-601 and 10-701
- This class is geared towards engineers and will include Python & PyTorch implementation of ML methods on real datasets

Course Logistics

Logistics

- Course Website: `https://www.andrew.cmu.edu/course/18-661/`: Slides and other reading materials
- Gradescope: Homework submission and grading (Entry Code: 4J36BK)
- Piazza: Course discussions and announcements, homework assignments (Sign-up Link: `https://piazza.com/cmu/spring2025/1846118661/home`)

Please make sure that you have access to Gradescope and Piazza ASAP.

Lectures, Recitations and Office Hours

Lectures

- Mon/Wed - 12:00-1:50 pm ET, TEP3500/CMR F205/B23 118
- You are expected to attend lectures in person on each campus. Zoom links are only provided for exceptional circumstances
- Recorded lectures will be uploaded to Canvas under the Panopto recordings tab, usually within 24 hours

Lectures, Recitations and Office Hours

Lectures

- Mon/Wed - 12:00-1:50 pm ET, TEP3500/CMR F205/B23 118
- You are expected to attend lectures in person on each campus. Zoom links are only provided for exceptional circumstances
- Recorded lectures will be uploaded to Canvas under the Panopto recordings tab, usually within 24 hours

Recitations

- Friday 11:00 am -12:20 pm ET (Pittsburgh/Kigali), 3:30pm - 4:50pm ET (Pittsburgh/SV)
- Pittsburgh students are welcome to attend either recitation
- Attendance is strongly encouraged but not mandatory
- Practice homework questions, supplementary material, exam review

Lectures, Recitations and Office Hours

Lectures

- Mon/Wed - 12:00-1:50 pm ET, TEP3500/CMR F205/B23 118
- You are expected to attend lectures in person on each campus. Zoom links are only provided for exceptional circumstances
- Recorded lectures will be uploaded to Canvas under the Panopto recordings tab, usually within 24 hours

Recitations

- Friday 11:00 am -12:20 pm ET (Pittsburgh/Kigali), 3:30pm - 4:50pm ET (Pittsburgh/SV)
- Pittsburgh students are welcome to attend either recitation
- Attendance is strongly encouraged but not mandatory
- Practice homework questions, supplementary material, exam review

Office Hours

- Dates/times will be posted on the course website & Piazza
- Zoom links/locations and weekly updates will be posted on Piazza

Homeworks and Exams

- Homeworks (40%): Both math and programming problems
- Miniexams (15%): Three during the semester
- Midterm Exam (15%): Linear and Logistic Regression, Naïve Bayes, SVMs (subject to change)
- Final Exam (25%): Everything else (Nearest Neighbors, Neural Networks, Decision Trees, Boosting, Clustering, PCA, etc.), plus pre-midterm topics (subject to change)
- Gradescope Quizzes (5%): Short multiple-choice question quizzes conducted in random (possibly all) lectures to encourage class attendance and attention. We will take the best 10 quiz scores.

Homeworks and Exams

- Homeworks (40%): Both math and programming problems
- Miniexams (15%): Three during the semester
- Midterm Exam (15%): Linear and Logistic Regression, Naïve Bayes, SVMs (subject to change)
- Final Exam (25%): Everything else (Nearest Neighbors, Neural Networks, Decision Trees, Boosting, Clustering, PCA, etc.), plus pre-midterm topics (subject to change)
- Gradescope Quizzes (5%): Short multiple-choice question quizzes conducted in random (possibly all) lectures to encourage class attendance and attention. We will take the best 10 quiz scores.

Grading

- Default grades are 90%+ A range, 80 - 89% B range, 70 - 79% C range, 60 - 69% D range, 59% and below F
- We may curve up grades at the end of the course based on the overall distribution, attendance, and class/Piazza participation

Homework and Exam Logistics

Homeworks

- Five homeworks (see website for tentative release and due dates)
- Released on Piazza/~~Canvas~~; scanned solutions to be submitted on Gradescope
- You have a total of 5 late days (max 2 late days per homework)
- Collaboration is encouraged, but you need to write your own answers **and** list the names of your collaborators on each homework
    - Generative AI tools like ChatGPT may be used. However, you must write out your own answers and include query printouts. You cannot directly ask the AI tool to solve any homework problem.
- Show your work to receive full (or partial) credit

Homework and Exam Logistics

Homeworks

- Five homeworks (see website for tentative release and due dates)
- Released on Piazza/Canvas; scanned solutions to be submitted on Gradescope
- You have a total of 5 late days (max 2 late days per homework)
- Collaboration is encouraged, but you need to write your own answers **and** list the names of your collaborators on each homework
  - Generative AI tools like ChatGPT may be used. However, you must write out your own answers and include query printouts. You cannot directly ask the AI tool to solve any homework problem.
- Show your work to receive full (or partial) credit

Exams

- Conducted in-person on each campus.
- No collaboration allowed.
- The use of generative AI tools is NOT allowed and will be reported to the university as an academic violation.

Student Wellness

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

Student Wellness

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress. If you feel overwhelmed or want to talk to someone, please feel free to contact the following resources:

- Counseling and Psychological Services (CaPS) in Pittsburgh at 412-268-2922 or `http://www.cmu.edu/counseling/`.
- Director of Student Affairs in SV at 650-335-2846, Building 19, Room 1041 or student-services@sv.cmu.edu.

Quick Polls

- MS students who are starting in Spring 2025?

Quick Polls

- MS students who are starting in Spring 2025?
- MS students who started in Fall 2024 or earlier?

Quick Polls

- MS students who are starting in Spring 2025?
- MS students who started in Fall 2024 or earlier?
- PhD students?

Quick Polls

- MS students who are starting in Spring 2025?
- MS students who started in Fall 2024 or earlier?
- PhD students?
- Waitlist / Hoping to Register ?

Course Prerequisites

By next week, you should be familiar with the following topics:

- Probability theory: Bayes' theorem, Gaussian distribution, expectation, variance
- Linear algebra: Matrix Inverse, Matrix Rank, Eigen values, SVD
- Calculus: Partial Differentiation, Integration, Convexity
- Python programming, in particular, `numpy`

Course Prerequisites

By next week, you should be familiar with the following topics:

- Probability theory: Bayes' theorem, Gaussian distribution, expectation, variance
- Linear algebra: Matrix Inverse, Matrix Rank, Eigen values, SVD
- Calculus: Partial Differentiation, Integration, Convexity
- Python programming, in particular, `numpy`

The first two lectures and the first recitation will go over these concepts.

These are meant to be representative samples of the math and programming you will need to succeed in this course. If you don't satisfy these pre-requisites, we strongly encourage you to take the class after reviewing introductory material.

Questions?

Probability Review

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events

$P: \mathcal{F} \to [0,1]$

$[0,1]$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$ $2^6 = 64$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - P(rolling an odd number)

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - P(rolling an odd number)

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - $P(\text{rolling an odd number}) = P(\{1, 3, 5\}) = \frac{1}{2}$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - $P(\text{rolling an odd number}) = P(\{1, 3, 5\}) = \frac{1}{2}$
- Tossing a fair coin twice
  - Ω: $\{HH, HT, TH, TT\}$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - $P(\text{rolling an odd number}) = P(\{1, 3, 5\}) = \frac{1}{2}$
- Tossing a fair coin twice
  - Ω: $\{HH, HT, TH, TT\}$
  - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - $P(\text{rolling an odd number}) = P(\{1, 3, 5\}) = \frac{1}{2}$
- Tossing a fair coin twice
  - Ω: $\{HH, HT, TH, TT\}$
  - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{\underline{HH, HT}\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$
  - $P(\text{first flip is heads})$

Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
    - Ω: $\{1, 2, 3, 4, 5, 6\}$
    - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
    - $P(\text{rolling an odd number}) = P(\{1, 3, 5\}) = \frac{1}{2}$
- Tossing a fair coin twice
    - Ω: $\{HH, HT, TH, TT\}$
    - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$
    - $P(\text{first flip is heads})$

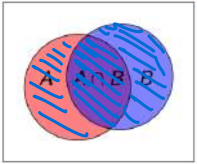
Probability Terminology

Name	Type	Symbol	Meaning
Sample Space	set	Ω, S	possible outcomes
Event Space	a set of subsets (of Ω)	$\mathcal{F}$, E	the events that have probabilities
Probability Measure	measure	P, π	assigns probabilities to events
Probability Space	a triple	$(\Omega, \mathcal{F}, P)$	

- Rolling a fair die
  - Ω: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
  - $P(\text{rolling an odd number}) = P(\{1, 3, 5\}) = \frac{1}{2}$
- Tossing a fair coin twice
  - Ω: $\{HH, HT, TH, TT\}$
  - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$
  - $P(\text{first flip is heads}) = P(\{HH, HT\}) = \frac{1}{2}$

Axioms of Probability

The probability measure satisfies the following properties:

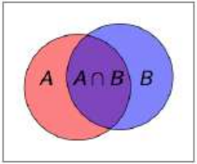


Ω

- $0 \leq P(A) \leq 1, \forall A \in \mathcal{F}$
- $P(\Omega) = 1, \; P(\varnothing) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Axioms of Probability

The probability measure satisfies the following properties:



- $0 \leq P(A) \leq 1, \ \forall A \in \mathcal{F}$
- $P(\Omega) = 1, \ P(\varnothing) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Question: For two tosses of a fair coin, suppose A is the event that at least one is H, and B is the event that there is exactly one T. Then what is $P(A \cup B)$?

$A = \{HH, HT, TH\}$

$B = \{HT, TH\}$

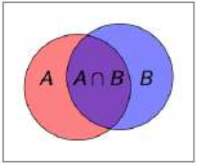
$P(A \cap B) = 1/2$

$P(A \cup B) = 3/4 + 1/2 - 1/2$

$= 3/4$

Axioms of Probability

The probability measure satisfies the following properties:

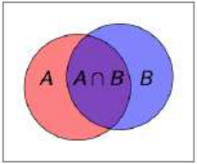


- $0 \leq P(A) \leq 1, \ \forall A \in \mathcal{F}$
- $P(\Omega) = 1, \ P(\varnothing) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Question: For two tosses of a fair coin, suppose A is the event that at least one is H, and B is the event that there is exactly one T. Then what is $P(A \cup B)$?

$$\begin{aligned} P(A \cup B) &= P(\{HH, HT, TH\}) + P(\{HT, TH\}) - P(\{HT, TH\}) \\ &= 0.75 + 0.5 - 0.5 = 0.75 \end{aligned}$$

Axioms of Probability

The probability measure satisfies the following properties:

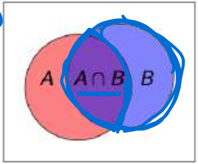


- $0 \leq P(A) \leq 1, \ \forall A \in \mathcal{F}$
- $P(\Omega) = 1, \ P(\varnothing) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Question: For two tosses of a fair coin, suppose A is the event that at least one is H, and B is the event that there is exactly one T. Then what is $P(A \cup B)$?

$$\begin{aligned} P(A \cup B) &= P(\{HH, HT, TH\}) + P(\{HT, TH\}) - P(\{HT, TH\}) \\ &= 0.75 + 0.5 - 0.5 = 0.75 \end{aligned}$$

Conditional Probability (Bayes Rule)

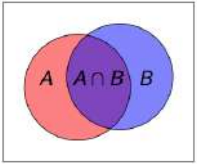
$\Omega \rightarrow$



- For events $A, B \in \mathcal{F}$, the conditional probability of A given B is given by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A,B)}{P(B)}$$

Conditional Probability (Bayes Rule)



- For events $A, B \in \mathcal{F}$, the conditional probability of A given B is given by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

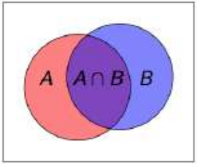
- Question: For two tosses of a fair coin, what is the probability of at least one T, given that the event TT did not occur? 1/3? 2/3? or 1/2?

$A = \{HT, TH, TT\}$

$B = \{HT, TH, HH\}$

$$\frac{P(A \cap B) = 1/2}{P(B) = 3/4} = 2/3$$

Conditional Probability (Bayes Rule)

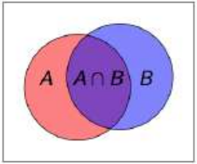


- For events $A, B \in \mathcal{F}$, the conditional probability of A given B is given by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

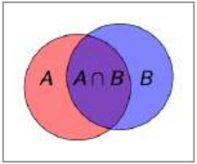
- Question: For two tosses of a fair coin, what is the probability of at least one T, given that the event TT did not occur?

Conditional Probability (Bayes Rule)



- For events $A, B \in \mathcal{F}$, the conditional probability of A given B is given by:
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$
- Question: For two tosses of a fair coin, what is the probability of at least one T, given that the event TT did not occur? ANS: 2/3

Conditional Probability (Bayes Rule)



Pf $B = \phi$ $\quad P(B) = 0$

$P(A \cap B) = 0$

$P(A \mid B) = \frac{0}{0} = 0$

- For events $A, B \in \mathcal{F}$, the conditional probability of A given B is given by:

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

- Question: For two tosses of a fair coin, what is the probability of at least one T, given that the event TT did not occur? ANS: 2/3
- Bayes rule:

$$\boxed{P(B \mid A)P(A)} = P(A \cap B) = \boxed{P(A \mid B)P(B)}$$

$$\implies P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Random Variables

Formally, a random variable is a function $X : \Omega \to \mathbb{R}$ that assigns a numerical value to each outcome s within a probability space $(\Omega, \mathcal{F}, P)$.

Example: Rolling a fair die

- Ω: $\{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
- $X(S) = S$ for each $S \in \Omega$

Random Variables

Formally, a random variable is a function $X : \Omega \to \mathbb{R}$ that assigns a numerical value to each outcome s within a probability space $(\Omega, \mathcal{F}, P)$.

Example: Rolling a fair die

- Ω: $\{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
- $X(S) = S$ for each $S \in \Omega$

The expectation of a random variable, $\mathbb{E}[X]$, is defined as $\sum_{s \in \Omega} X(s)P(s)$, i.e., the average value of X.

- $\mathbb{E}[X] = \sum_{s=1}^{6} sP(s) = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}$

Random Variables → eg Gaussian, Exponential, Bernoulli, Binomial, ...

Formally, a random variable is a function $X : \Omega \to \mathbb{R}$ that assigns a numerical value to each outcome s within a probability space $(\Omega, \mathcal{F}, P)$.

Example: Rolling a fair die

- Ω: $\{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
- $X(S) = S$ for each $S \in \Omega$

The expectation of a random variable, $\mathbb{E}[X]$, is defined as $\sum_{s \in \Omega} X(s)P(s)$, i.e., the average value of X.

- $\mathbb{E}[X] = \sum_{s=1}^{6} sP(s) = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}$

The variance of a random variable is $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$, which measures how much X can vary from its expected value $\mathbb{E}[X]$.

- $\mathrm{Var}[X] = \sum_{s=1}^{6} \frac{1}{6}(s - \frac{7}{2})^2 = \frac{35}{12}$

Some Other Concepts that You Should Know

- Discrete and continuous random variables
- PMF (probability mass function), PDF (probability density function), CDF (cumulative distribution function) of random variables
- Entropy of a random variable

Some of these will be covered in HW1

A Simple Learning Problem: MLE/MAP Estimation

Dogecoin

- Scenario: You find a coin on the ground.

- *You ask yourself: Is this a fair or biased coin? What is the probability that I will flip a heads?*

- You flip the coin 10 times . . .

- You flip the coin 10 times . . .

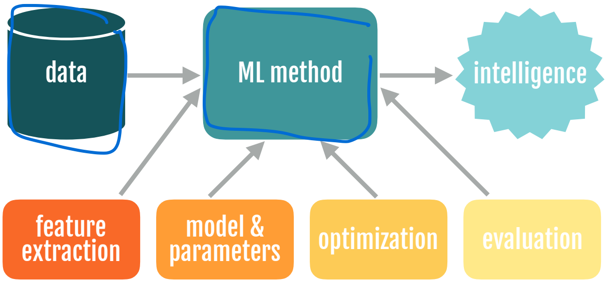
Data

- It comes up as 'H' 8 times and 'T' 2 times

- You flip the coin 10 times . . .
- It comes up as 'H' 8 times and 'T' 2 times
- Can we learn the bias of the coin from this data?

$p = 0.8$? ←

Recall: Machine Learning Pipeline



Two methods that we will discuss today:

- Maximum likelihood Estimation (MLE)
- Maximum a posteriori Estimation (MAP)

Maximum Likelihood Estimation (MLE)

- **Data**: Observed sequence D of n_H heads and n_T tails

Maximum Likelihood Estimation (MLE)

- **Data**: Observed sequence D of n_H heads and n_T tails
- **Model**: Each flip follows a Bernoulli distribution

$$P(H) = \theta,\ P(T) = 1 - \theta,\ \theta \in [0, 1]$$

Maximum Likelihood Estimation (MLE)

- **Data**: Observed sequence D of n_H heads and n_T tails
- **Model**: Each flip follows a Bernoulli distribution

$$P(H) = \theta,\ P(T) = 1 - \theta,\ \theta \in [0, 1]$$

- Thus, the likelihood of observing sequence D is

$$P(D \mid \theta) = \theta^{n_H}(1 - \theta)^{n_T}$$

Maximum Likelihood Estimation (MLE)

- **Data**: Observed sequence D of n_H heads and n_T tails
- **Model**: Each flip follows a Bernoulli distribution

$$P(H) = \theta,\ P(T) = 1 - \theta,\ \theta \in [0, 1]$$

- Thus, the likelihood of observing sequence D is

$$P(D \mid \theta) = \theta^{n_H}(1-\theta)^{n_T}$$

- Question: Given this model and the data we've observed, can we calculate an estimate of θ?

Maximum Likelihood Estimation (MLE)

- **Data**: Observed sequence D of n_H heads and n_T tails
- **Model**: Each flip follows a Bernoulli distribution

$$P(H) = \theta,\ P(T) = 1 - \theta,\ \theta \in [0, 1]$$

- Thus, the likelihood of observing sequence D is

$$P(D \mid \theta) = \theta^{n_H}(1 - \theta)^{n_T}$$

- Question: Given this model and the data we've observed, can we calculate an estimate of θ?
- **MLE**: Choose θ that maximizes the *likelihood* of the observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

How to Solve?

- $\log(x)$ is a monotone increasing function; will not affect the arg max

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} P(D \mid \theta) \\
&= \arg\max_{\theta} \log P(D \mid \theta) \\
&= \arg\max_{\theta} \log\left(\theta^{n_H}(1-\theta)^{n_T}\right) \\
&= \arg\max_{\theta} \underbrace{n_H \log(\theta) + n_T log(1-\theta)}_{\text{concave}}
\end{aligned}$$

How to Solve?

- $\log(x)$ is a monotone increasing function; will not affect the arg max

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} P(D \mid \theta) \\
&= \arg\max_{\theta} \log P(D \mid \theta) \\
&= \arg\max_{\theta} \log\left(\theta^{n_H}(1-\theta)^{n_T}\right) \\
&= \arg\max_{\theta} \underbrace{n_H \log(\theta) + n_T log(1-\theta)}_{\text{concave}}
\end{aligned}$$

How to Solve?

- $\log(x)$ is a monotone increasing function; will not affect the arg max

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} P(D \mid \theta) \\
&= \arg\max_{\theta} \log P(D \mid \theta) \\
&= \arg\max_{\theta} \log\left(\theta^{n_H}(1-\theta)^{n_T}\right) \\
&= \arg\max_{\theta} \underbrace{n_H \log(\theta) + n_T log(1-\theta)}_{\text{concave}}
\end{aligned}$$

- Take derivative $\frac{\partial}{\partial\theta} \log P(D \mid \theta)$ and set equal to zero

$$\frac{n_H}{\theta} + \frac{n_T}{1-\theta} \times -1 = 0$$

$$\theta = \frac{n_H}{n_H + n_T}$$

How to Solve?

- $\log(x)$ is a monotone increasing function; will not affect the arg max

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} P(D \mid \theta) \\
&= \arg\max_{\theta} \log P(D \mid \theta) \\
&= \arg\max_{\theta} \log\left(\theta^{n_H}(1-\theta)^{n_T}\right) \\
&= \arg\max_{\theta} \underbrace{n_H \log(\theta) + n_T log(1-\theta)}_{\text{concave}}
\end{aligned}
$$

- Take derivative $\frac{\partial}{\partial\theta} \log P(D \mid \theta)$ and set equal to zero

How to Solve?

- $\log(x)$ is a monotone increasing function; will not affect the arg max

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta} P(D \mid \theta) \\ &= \arg\max_{\theta} \log P(D \mid \theta) \\ &= \arg\max_{\theta} \log\left(\theta^{n_H}(1-\theta)^{n_T}\right) \\ &= \arg\max_{\theta} \underbrace{n_H \log(\theta) + n_T log(1-\theta)}_{\text{concave}}\end{aligned}$$

- Take derivative $\frac{\partial}{\partial\theta} \log P(D \mid \theta)$ and set equal to zero

$$\begin{aligned}0 &= \frac{\partial}{\partial\theta}\left(n_H \log(\theta) + n_T log(1-\theta)\right) \\ &= \frac{n_H}{\theta} - \frac{n_T}{1-\theta} \\ \Longrightarrow \hat{\theta}_{MLE} &= \frac{n_H}{n_H + n_T}\end{aligned}$$

Going Back to Our Scenario

- You flip the coin 10 times ...
- It comes up as 'H' 8 times and 'T' 2 times
- Can we learn the bias θ of the coin from this data?

Going Back to Our Scenario

- You flip the coin 10 times ...
- It comes up as 'H' 8 times and 'T' 2 times

- Can we learn the bias θ of the coin from this data?

$$\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T} = 0.8$$

Going Back to Our Scenario

- You flip the coin 10 times ...
- It comes up as 'H' 8 times and 'T' 2 times

- Can we learn the bias θ of the coin from this data?

$$\boxed{\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T} = 0.8}$$

Here, we are trusting the data completely. But there could be too little data or noisy data...

Going Back to Our Scenario

- You flip the coin 10 times . . .
- It comes up as 'H' 8 times and 'T' 2 times

- Can we learn the bias θ of the coin from this data?

$$\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T} = 0.8$$

Here, we are trusting the data completely. But there could be too little data or noisy data... We will cover this in the next lecture.

You Should Know

- Structure of the machine learning pipeline
- Basics of probability theory: Bayes' Theorem and conditional probabilities, expectation, variance
- Maximum likelihood estimation formulation and procedure

Course Prerequisites

By next week, you should be familiar with the following topics:

- Probability theory: Bayes' theorem, Gaussian distribution, expectation, variance
- Linear algebra: Matrix Inverse, Matrix Rank, Eigen values, SVD
- Calculus: Partial Differentiation, Integration, Convexity
- Python programming, in particular, `numpy`

Course Prerequisites

By next week, you should be familiar with the following topics:

- Probability theory: Bayes' theorem, Gaussian distribution, expectation, variance
- Linear algebra: Matrix Inverse, Matrix Rank, Eigen values, SVD
- Calculus: Partial Differentiation, Integration, Convexity
- Python programming, in particular, `numpy`

The first two lectures and the first recitation will go over these concepts.

These are meant to be representative samples of the math and programming you will need to succeed in this course. If you don't satisfy these pre-requisites, we strongly encourage you to take the class after reviewing introductory material.

Math quiz

- Today's math quiz will hopefully mitigate attrition later
  - Representative of mathematical concepts you are excepted to know
  - Graded to assess your background (but not part of final grade)
  - We may contact students who perform poorly

Math quiz

- Today's math quiz will hopefully mitigate attrition later
  - Representative of mathematical concepts you are excepted to know
  - Graded to assess your background (but not part of final grade)
  - We may contact students who perform poorly
- Be honest / realistic with yourself about your background
- If not today, you should be able to solve the quiz completely by the end of this week, after Thursday's lecture on probability/linear algebra review
  - It's better for you and your classmates to drop the course now rather than a month from now, so that others can be admitted off the waitlist

Math quiz

- Today's math quiz will hopefully mitigate attrition later
  - Representative of mathematical concepts you are excepted to know
  - Graded to assess your background (but not part of final grade)
  - We may contact students who perform poorly
- Be honest / realistic with yourself about your background
- If not today, you should be able to solve the quiz completely by the end of this week, after ~~Thursday~~ Wednesday's lecture on probability/linear algebra review
  - It's better for you and your classmates to drop the course now rather than a month from now, so that others can be admitted off the waitlist

TAs will discuss the math quiz during this Friday's recitations

Questions?

Math Quiz
On Gradescope (Entry Code: 4J36BK)
Score won't affect grade
But is an indication of your preparedness for the course