18-461/661: Introduction to ML for Engineers

Probability and Linear Algebra Review

Spring 2020

ECE – Carnegie Mellon University
Registration

- Course is currently full and we can’t increase class size
- Expect several students will drop the course
- Course will be offered again in Fall 2020

Direct all waitlist-related questions to Megan Oliver (Pittsburgh) or Brittany Walker (SV):
mvoliver@andrew.cmu.edu and bmw2@andrew.cmu.edu

- 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
Announcements

- Graded quizzes will be posted on Gradescope – Entry Code 9NEDVR
- HW1 will be posted on Fri Jan 17, due on Mon Jan 27
- Recitation this Friday will go through the math quiz
- No class on Monday due to the MLK Day holiday
1. Recap: What is Machine Learning?

2. Probability Review

3. A Simple Learning Problem: MLE/MAP Estimation

4. Linear Algebra Review
Recap: What is Machine Learning?
• **Machine learning is:** the study of methods that *improve their performance on some task with experience*
Goal: Choose the Right ML Method for a Given Task

- **data**
- **feature extraction**
- **model & parameters**
- **optimization**
- **evaluation**
- **ML method**
- **intelligence**
Task 1: Regression

How much should you sell your house for?

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

Course Covers: Feature Scaling, Linear/Ridge Regression, Loss Function, SGD, Regularization, Cross Validation
Task 2: Classification

Cat or dog?

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

Course Covers: Naive Bayes, Logistic Regression, SVMs, Neural Nets, Decision Trees, Boosting
Task 3: Clustering

How to segment an image?

input: raw pixels \( \{x\} \)

separate: \( \{x\} \) into sets

output: cluster labels \( \{z\} \)

Course Covers: Nearest Neighbors, K-means clustering
Task 4: Embedding

How to reduce size of dataset?

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

Course Covers: Dimensionality Reduction, PCA
1. Recap: What is Machine Learning?

2. Probability Review

3. A Simple Learning Problem: MLE/MAP Estimation

4. Linear Algebra Review
Probability Review
Probability Terminology

### Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>Ω, S</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets</td>
<td>F, E</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>P, π</td>
<td>assigns probabilities to events</td>
</tr>
</tbody>
</table>

#### Probability Space

A triple $(\Omega, F, P)$

#### Examples

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

- **Tossing a fair coin twice**
  - $\Omega$: \{HH, HT, TH, TT\}
  - $F$ = \{\{HH\}, \{HT\}, ..., \{HH, HT\}, ..., \{HH, HT, TH, TT\}, \{\}\}
  - $P(\text{first flip is heads}) = P(\{HH, HT\}, \{HT\}) = \frac{1}{2}$
# Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega$, $S$</td>
<td>possible outcomes</td>
</tr>
</tbody>
</table>
### Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega, S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}, E$</td>
<td>the events that have probabilities</td>
</tr>
</tbody>
</table>
### Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega, S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}, E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P, \pi$</td>
<td>assigns probabilities to events</td>
</tr>
</tbody>
</table>
# Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega, S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}, E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P, \pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

**Examples:**

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

- **Tossing a fair coin twice**
  - $\Omega$: $\{HH, HT, TH, TT\}$
  - $F$: $\{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{}\}$
  - $P$(first flip is heads) = $\frac{1}{2}$
### Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega$, $S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}$, $E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P$, $\pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

**Examples:**

- Rolling a fair die
  - $\Omega$: $\{1, 2, 3, 4, 5, 6\}$
## Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>Ω, S</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of Ω)</td>
<td>𝒇, 𝐸</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>𝑃, 𝜋</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>(Ω, 𝒇, 𝑃)</td>
<td></td>
</tr>
</tbody>
</table>

**Examples:**

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega$, $S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}$, $E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P$, $\pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

### Examples:
- Rolling a fair die
  - $\Omega$: $\{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}, \{\}\}$
## Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>Ω, S</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of Ω)</td>
<td>F, E</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>P, π</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>(Ω, F, P)</td>
<td></td>
</tr>
</tbody>
</table>

**Examples:**

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega$, $S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}$, $\mathcal{E}$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P$, $\pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

### Examples:

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega, S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}, E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P, \pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

### Examples:

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>Ω, S</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of Ω)</td>
<td>𝐹, 𝐸</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>𝑃, 𝜖</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>(Ω, 𝐹, 𝑃)</td>
<td></td>
</tr>
</tbody>
</table>

Examples:

- 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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega, S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}, E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P, \pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

### Examples:

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

- **Tossing a fair coin twice**
  
  - $\Omega$: $\{HH, HT, TH, TT\}$
  
  - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$
## Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega$, $S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}$, $E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P$, $\pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

### Examples:

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

- Tossing a fair coin twice
  - $\Omega$: $\{HH, HT, TH, TT\}$
  - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$
  - $P(\text{first flip is heads})$
Probability Terminology

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>( \Omega, S )</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of ( \Omega ))</td>
<td>( \mathcal{F}, E )</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>( P, \pi )</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>((\Omega, \mathcal{F}, P))</td>
<td></td>
</tr>
</tbody>
</table>

Examples:

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

- Tossing a fair coin twice
  - \( \Omega: \{HH, HT, TH, TT\} \)
  - \( \mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\} \)
  - \( P(\text{first flip is heads}) \)
**Probability Terminology**

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Space</td>
<td>set</td>
<td>$\Omega$, $S$</td>
<td>possible outcomes</td>
</tr>
<tr>
<td>Event Space</td>
<td>a set of subsets (of $\Omega$)</td>
<td>$\mathcal{F}$, $E$</td>
<td>the events that have probabilities</td>
</tr>
<tr>
<td>Probability Measure</td>
<td>measure</td>
<td>$P$, $\pi$</td>
<td>assigns probabilities to events</td>
</tr>
<tr>
<td>Probability Space</td>
<td>a triple</td>
<td>$(\Omega, \mathcal{F}, P)$</td>
<td></td>
</tr>
</tbody>
</table>

**Examples:**

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

- **Tossing a fair coin twice**
  - $\Omega$: $\{HH, HT, TH, TT\}$
  - $\mathcal{F} = \{\{HH\}, \{HT\}, \ldots, \{HH, HT\}, \ldots, \{HH, HT, TH, TT\}, \{\}\}$
  - $P$(first flip is heads) = $P(\{HH\}, \{HT\}) = \frac{1}{2}$
Axioms of Probability

- \( 0 \leq P(A) \leq 1 \)
- \( P(\Omega) = 1, \; P(\emptyset) = 0 \)
- \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)

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 \( P(A \cup B) \)?

\[
P(A \cup B) = 0.75 + 0.5 - 0.5 = 0.75
\]
Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1, \ P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

**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 $P(A \cup B)$?
Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$, $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **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 $P(A \cup B)$?

$$P(A \cup B) = 0.75 + 0.5 - 0.5 = 0.75$$
For events $A, B \in \mathcal{F}$, the conditional probability of $A$ given $B$ is given by:

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

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

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

**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?
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: $2/3$

**Bayes rule:**

$$P(B \mid A)P(A) = P(A \cap B) = P(A \mid B)P(B)$$

$$\Rightarrow P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$
Some Other Concepts that You Should Know

- Discrete and Continuous Random Variables
- PMF, PDF, CDF
- Expectation and Variance
- Entropy
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 . . .
• 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?
Two approaches that we will discuss today:

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

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

Thus, the likelihood of observing sequence $D$ is $P(D|\theta) = \theta^{n_H} (1-\theta)^{n_T}$.
Maximum Likelihood Estimation (MLE)

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

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

Thus, the likelihood of observing sequence $D$ is
• **Data**: Observed set $D$ 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 set $D$ 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 $\theta$?
Maximum Likelihood Estimation (MLE)

- **Data**: Observed set $D$ 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 $\theta$?
- **MLE**: Choose $\theta$ that maximizes the likelihood of the observed data
  
  $$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D \mid \theta)$$
• \( \log(x) \) is a monotone increasing function; will not affect the \( \arg \max \)

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

\[
= \arg \max_{\theta} \log P(D \mid \theta)
\]

\[
= \arg \max_{\theta} \log (\theta^nH(1 - \theta)^nT)
\]
• log(x) is a monotone increasing function; will not affect the arg max

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

\[ = \arg \max_{\theta} \log P(D \mid \theta) \]

\[ = \arg \max_{\theta} \log (\theta^{n_H} (1 - \theta)^{n_T}) \]

\[ = \arg \max_{\theta} \underbrace{n_H \log(\theta) + n_T \log(1 - \theta)}_{\text{concave}} \]
How to solve?

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

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

\[
= \arg \max_{\theta} \log P(D \mid \theta)
\]

\[
= \arg \max_{\theta} \log (\theta^{n_H}(1 - \theta)^{n_T})
\]

\[
= \arg \max_{\theta} \underbrace{n_H \log(\theta) + n_T \log(1 - \theta)}_{\text{concave}}
\]
How to solve?

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

\[
\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} P(D \mid \theta)
\]

\[
= \arg \max_{\theta} \log P(D \mid \theta)
\]

\[
= \arg \max_{\theta} \log (\theta^{n_H}(1 - \theta)^{n_T})
\]

\[
= \arg \max_{\theta} \left( n_H \log(\theta) + n_T \log(1 - \theta) \right)
\]

- 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

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

\[
= \arg\max_{\theta} \log P(D \mid \theta)
\]

\[
= \arg\max_{\theta} \log (\theta^{n_H} (1 - \theta)^{n_T})
\]

\[
= \arg\max_{\theta} n_H \log(\theta) + n_T \log(1 - \theta)
\]

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

\[
0 = \frac{\partial}{\partial \theta} n_H \log(\theta) + n_T \log(1 - \theta)
\]
How to solve?

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

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

\[
= \arg \max_{\theta} \log P(D \mid \theta)
\]

\[
= \arg \max_{\theta} \log (\theta^{n_H} (1 - \theta)^{n_T})
\]

\[
= \arg \max_{\theta} \frac{n_H \log(\theta) + n_T \log(1 - \theta)}{\text{concave}}
\]

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

\[
0 = \frac{\partial}{\partial \theta} n_H \log(\theta) + n_T \log(1 - \theta)
\]

\[
= \frac{n_H}{\theta} - \frac{n_T}{1 - \theta}
\]
How to solve?

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

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

\[
= \arg \max_{\theta} \log P(D \mid \theta)
\]

\[
= \arg \max_{\theta} \log (\theta^n (1 - \theta)^t)
\]

\[
= \arg \max_{\theta} \underbrace{n_H \log(\theta) + n_T \log(1 - \theta)}_{\text{concave}}
\]

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

\[
0 = \frac{\partial}{\partial \theta} n_H \log(\theta) + n_T \log(1 - \theta)
\]

\[
= \frac{n_H}{\theta} - \frac{n_T}{1 - \theta}
\]

\[
\implies \hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T}
\]
Going back to our scenario

- You flip the coin 10 times . . .
Going back to our scenario

- You flip the coin 10 times . . .
- It comes up as 'H' 8 times and 'T' 2 times
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 \( \theta \) 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 $\theta$ of the coin from this data?
• You flip the coin 10 times . . .
• It comes up as 'H' 8 times and 'T' 2 times

**Can we learn the bias $\theta$ of the coin from this data?**

$$\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T} = 0.8$$
• You flip the coin 10 times . . .
• It comes up as 'H' 8 times and 'T' 2 times
• Can we learn the bias $\theta$ 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.
What about prior knowledge?

- We believe the coin is *supposed* to be close to 50-50
What about prior knowledge?

- We believe the coin is *supposed* to be close to 50-50
- Rather than completely “trusting” the data as-is, we want to use the data to update our prior beliefs
What about prior knowledge?

- We believe the coin is *supposed* to be close to 50-50
- Rather than completely “trusting” the data as-is, we want to use the data to update our prior beliefs
What about prior knowledge?

- We believe the coin is *supposed* to be close to 50-50
- Rather than completely “trusting” the data as-is, we want to use the data to update our prior beliefs
Bayesian Learning

- How to incorporate prior knowledge?

\[ P(\theta | D) = \frac{P(D | \theta) P(\theta)}{P(D)} \]

Or, equivalently:

\[ P(\theta | D) \propto P(D | \theta) P(\theta) \]
Bayesian Learning

• How to incorporate prior knowledge?
• Use Bayes’ Rule:

\[ P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \]
Bayesian Learning

- How to incorporate prior knowledge?
- Use Bayes’ Rule:

\[ P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \]

- Or, equivalently:

\[ P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \]

posterior \hspace{1cm} likelihood \hspace{1cm} prior
Bayesian Learning

• How to incorporate prior knowledge?
• Use Bayes’ Rule:

\[ P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \]

• Or, equivalently:

\[ P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \]

posterior \hspace{1cm} likelihood \hspace{1cm} prior
Bayesian Learning

• How to incorporate prior knowledge?
• Use Bayes’ Rule:

\[
P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}
\]

• Or, equivalently:

\[
P(\theta \mid D) \propto P(D \mid \theta)P(\theta)
\]

posterior \hspace{1cm} \text{likelihood} \hspace{1cm} \text{prior}
\[ \hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(D \mid \theta) P(\theta) \]

- Recall that \( P(D \mid \theta) = \theta^{n_H} (1 - \theta)^{n_T} \)
\[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(D \mid \theta) P(\theta) \]

- Recall that \( P(D \mid \theta) = \theta^n (1 - \theta)^T \)
- How should we set the prior, \( P(\theta) \)?
\[
\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(D \mid \theta) P(\theta)
\]

- Recall that \(P(D \mid \theta) = \theta^nH(1 - \theta)^nT\)
- How should we set the prior, \(P(\theta)\)?
- Common choice for a binomial likelihood is to use the Beta distribution, \(\theta \sim \text{Beta}(\alpha, \beta)\):

\[
P(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}, \text{where } B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta
\]
\[ \hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(D \mid \theta) P(\theta) \]

- Recall that \( P(D \mid \theta) = \theta^{n_H} (1 - \theta)^{n_T} \)
- How should we set the prior, \( P(\theta) \)?
- Common choice for a binomial likelihood is to use the Beta distribution, \( \theta \sim \text{Beta}(\alpha, \beta) \):

\[
P(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \text{ where } B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta
\]

- Interpretation: \( \alpha = \) number of expected heads, \( \beta = \) number of expected tails. Larger value of \( \alpha + \beta \) denotes more confidence (and smaller variance).
Beta Distribution

\( \frac{\alpha}{\beta} \) controls left/right bias, \( \alpha + \beta \) controls height of peak
A benefit of using the *Beta* distribution as a prior is that the posterior will also be a *Beta* distribution:

\[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(D | \theta) P(\theta) \]
A benefit of using the *Beta* distribution as a prior is that the posterior will also be a *Beta* distribution:

\[ \hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(D \mid \theta)P(\theta) \]

\[ = \arg \max_{\theta} \theta^{\alpha + n_H - 1}(1 - \theta)^{\beta + n_T - 1} \]
A benefit of using the *Beta* distribution as a prior is that the posterior will also be a *Beta* distribution:

\[
\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(D \mid \theta) P(\theta) = \arg \max_{\theta} \theta^{\alpha + n_H - 1} (1 - \theta)^{\beta + n_T - 1} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2}
\]
A benefit of using the *Beta* distribution as a prior is that the posterior will also be a *Beta* distribution:

\[
\hat{\theta}_{MAP} = \arg \max_{\theta} P(D | \theta)P(\theta)
\]

\[
= \arg \max_{\theta} \theta^{\alpha+n_H-1}(1-\theta)^{\beta+n_T-1}
\]

\[
= \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2}
\]

Note that as \( n_H + n_T \to \infty \), the effect of the prior disappears and we recover the MLE estimate.
Putting it all together

\[ \hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T} \]

\[ \hat{\theta}_{MAP} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2} \]

Suppose \( \theta^* = 0.5 \) and we observe: \( D = \{H, H, T, T, T, T\} \).

Scenario 1: We assume \( \theta \sim \text{Beta}(4, 4) \). Which is more accurate – \( \hat{\theta}_{MLE} \) or \( \hat{\theta}_{MAP} \)?

\( \hat{\theta}_{MAP} = \frac{5}{12}, \hat{\theta}_{MLE} = \frac{1}{3} \)

Scenario 2: We assume \( \theta \sim \text{Beta}(1, 7) \). Which is more accurate – \( \hat{\theta}_{MLE} \) or \( \hat{\theta}_{MAP} \)?

\( \hat{\theta}_{MAP} = \frac{1}{6}, \hat{\theta}_{MLE} = \frac{1}{3} \)
Putting it all together

\[
\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T}
\]
\[
\hat{\theta}_{MAP} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2}
\]

- Suppose \(\theta^* := 0.5\) and we observe: \(D = \{H, H, T, T, T, T\}\)
Putting it all together

\[ \hat{\theta}_{\text{MLE}} = \frac{n_H}{n_H + n_T} \]
\[ \hat{\theta}_{\text{MAP}} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2} \]

- Suppose \( \theta^* := 0.5 \) and we observe: \( D = \{H, H, T, T, T, T\} \)
- Scenario 1: We assume \( \theta \sim \text{Beta}(4, 4) \). Which is more accurate – \( \theta_{\text{MLE}} \) or \( \theta_{\text{MAP}} \)?
**Putting it all together**

\[
\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T}
\]

\[
\hat{\theta}_{MAP} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2}
\]

- Suppose \( \theta^* := 0.5 \) and we observe: \( D = \{H, H, T, T, T, T\} \)
- Scenario 1: We assume \( \theta \sim Beta(4, 4) \). Which is more accurate – \( \theta_{MLE} \) or \( \theta_{MAP} \)?
  - \( \theta_{MAP} = 5/12, \theta_{MLE} = 1/3 \)
- Scenario 2: We assume \( \theta \sim Beta(1, 7) \). Which is more accurate – \( \theta_{MLE} \) or \( \theta_{MAP} \)?
  - \( \theta_{MAP} = 1/6, \theta_{MLE} = 1/3 \)
Putting it all together

\[ \hat{\theta}_{\text{MLE}} = \frac{n_H}{n_H + n_T} \]
\[ \hat{\theta}_{\text{MAP}} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2} \]

- Suppose \( \theta^* := 0.5 \) and we observe: \( D = \{H, H, T, T, T, T\} \)
- Scenario 1: We assume \( \theta \sim \text{Beta}(4, 4) \). Which is more accurate – \( \theta_{\text{MLE}} \) or \( \theta_{\text{MAP}} \)?
  - \( \theta_{\text{MAP}} = 5/12, \theta_{\text{MLE}} = 1/3 \)
- Scenario 2: We assume \( \theta \sim \text{Beta}(1, 7) \). Which is more accurate – \( \theta_{\text{MLE}} \) or \( \theta_{\text{MAP}} \)?
Putting it all together

\[
\hat{\theta}_{\text{MLE}} = \frac{n_H}{n_H + n_T}
\]

\[
\hat{\theta}_{\text{MAP}} = \frac{\alpha + n_H - 1}{\alpha + \beta + n_H + n_T - 2}
\]

- Suppose \( \theta^* := 0.5 \) and we observe: \( D = \{H, H, T, T, T, T\} \)
- Scenario 1: We assume \( \theta \sim \text{Beta}(4, 4) \). Which is more accurate – \( \theta_{\text{MLE}} \) or \( \theta_{\text{MAP}} \)?
  - \( \theta_{\text{MAP}} = 5/12, \theta_{\text{MLE}} = 1/3 \)
- Scenario 2: We assume \( \theta \sim \text{Beta}(1, 7) \). Which is more accurate – \( \theta_{\text{MLE}} \) or \( \theta_{\text{MAP}} \)?
  - \( \theta_{\text{MAP}} = 1/6, \theta_{\text{MLE}} = 1/3 \)
Bayesians vs. Frequentists

You are no good when sample is small

You give a different answer for different priors
Why was this a ML problem?

Machine learning is: the study of methods that

*improve their performance* (the accuracy of the predicted probability)
Why was this a ML problem?

**Machine learning is:** the study of methods that

*improve their performance* (the accuracy of the predicted probability)

*on some task* (predicting the probability of 'heads')
Why was this a ML problem?

**Machine learning is**: the study of methods that

*improve their performance* (the accuracy of the predicted probability)

*on some task* (predicting the probability of 'heads')

*with experience* (the more coin flips we see, the better our guess)
Learning involves ...

- Collect some data
Learning involves ...

- Collect some data
  - e.g., coin flips
Learning involves ...

- Collect some data
  - e.g., coin flips
- Set up the problem: Choose a model / loss function

Key idea: these are choices. It's important to understand the implications of these choices and evaluate their trade-offs for the problem at hand.
Learning involves ...

- Collect some data
  - e.g., coin flips
- Set up the problem: Choose a model / loss function
  - e.g., bernoulli model, data likelihood/ a posteriori prob.

Key idea: these are choices. It's important to understand the implications of these choices and evaluate their trade-offs for the problem at hand.
Learning involves ...

- Collect some data
  - e.g., coin flips
- Set up the problem: Choose a model / loss function
  - e.g., bernoulli model, data likelihood/ a posteriori prob.
- Solve the problem: Choose an optimization procedure
Learning involves ...

- Collect some data
  - e.g., coin flips
- Set up the problem: Choose a model / loss function
  - e.g., bernoulli model, data likelihood/ a posteriori prob.
- Solve the problem: Choose an optimization procedure
  - e.g., set derivative of log to zero and solve to find MLE/MAP
Learning involves ...

- Collect some data
  - e.g., coin flips
- Set up the problem: Choose a model / loss function
  - e.g., bernoulli model, data likelihood/a posteriori prob.
- Solve the problem: Choose an optimization procedure
  - e.g., set derivative of log to zero and solve to find MLE/MAP
Learning involves ...

- Collect some data
  - e.g., coin flips
- Set up the problem: Choose a model / loss function
  - e.g., bernoulli model, data likelihood / a posteriori prob.
- Solve the problem: Choose an optimization procedure
  - e.g., set derivative of log to zero and solve to find MLE/MAP

Key idea: these are choices. It's important to understand the implications of these choices and evaluate their trade-offs for the problem at hand.
Linear Algebra Review
Recall: Task 1: Regression

How much should you sell your house for?

input: houses & features  learn: $x \rightarrow y$ relationship  predict: $y$ (continuous)
Data Can be Compactly Represented by Matrices

- Learn parameters \((w_1, w_0)\) of the orange line \(y = w_1 x + w_0\)
  - Sq.ft
  - House 1: \(1000 \times w_1 + w_0 = 200,000\)
  - House 2: \(2000 \times w_1 + w_0 = 350,000\)
Data Can be Compactly Represented by Matrices

- Learn parameters \((w_1, w_0)\) of the orange line \(y = w_1 x + w_0\)
  
  Sq.ft
  
  House 1: \(1000 \times w_1 + w_0 = 200,000\)
  House 2: \(2000 \times w_1 + w_0 = 350,000\)

- Can represent compactly in matrix notation
  
  \[
  \begin{bmatrix}
  1000 & 1 \\
  2000 & 1
  \end{bmatrix}
  \begin{bmatrix}
  w_1 \\
  w_0
  \end{bmatrix}
  =
  \begin{bmatrix}
  200,000 \\
  350,000
  \end{bmatrix}
  \]
Some Concepts That You Should Know

- Invertibility of Matrices and Computing Inverses
- Vector Norms – L2, Frobenius etc., Inner Products
- Eigenvalues and Eigen-vectors
- Singular Value Decomposition
- Covariance Matrices and Positive Semi-definite-ness

Excellent Resources:

- Essence of Linear Algebra YouTube Series by 3Blue1Brown
- Prof. Gilbert Strang’s course at MIT
Matrix multiplication

• For two matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, their product is:

$$AB = C \in \mathbb{R}^{m \times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$$
Matrix multiplication

• For two matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, their product is:

$$AB = C \in \mathbb{R}^{m \times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

• Multiplication is undefined with the number of columns in $A \neq$ the number of rows in $B$ (except in case: $cA$ where $c \in \mathbb{R}$ is a scalar)
Matrix multiplication

- For two matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, their product is:

  $$AB = C \in \mathbb{R}^{m \times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$$

- Multiplication is undefined with the number of columns in $A \neq$ the number of rows in $B$ (except in case: $cA$ where $c \in \mathbb{R}$ is a scalar)

- Special cases:
Matrix multiplication

- For two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}$, their product is:

$$AB = C \in R^{m\times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$$

- Multiplication is undefined with the number of columns in $A \neq$ the number of rows in $B$ (except in case: $cA$ where $c \in \mathbb{R}$ is a scalar)

- Special cases:
  - Inner product: $x, y \in \mathbb{R}^n, \quad x^\top y \in \mathbb{R} = \sum_{i=1}^{n} x_iy_i$
Matrix multiplication

- For two matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, their product is:

$$AB = C \in \mathbb{R}^{m \times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$$

- Multiplication is undefined with the number of columns in $A \neq$ the number of rows in $B$ (except in case: $cA$ where $c \in \mathbb{R}$ is a scalar)

- Special cases:
  - Inner product: $x, y \in \mathbb{R}^n$, $x^\top y \in \mathbb{R} = \sum_{i=1}^{n} x_iy_i$
  - Matrix-vector product: $A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n \iff Ax \in \mathbb{R}^m$

$$A = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix}, \quad Ax \in \mathbb{R}^m = \sum_{i=1}^{n} a_ix_i$$
Important properties

- Associative: $A(BC) = (AB)C$

*Not* Commutative: $AB \neq BA$

Transpose: $(AB)^\top = B^\top A^\top$
Important properties

- Associative: $A(BC) = (AB)C$
- Distributive: $A(B + C) = AB + AC$
Important properties

- Associative: \( A(BC) = (AB)C \)
- Distributive: \( A(B + C) = AB + AC \)
- *Not* Commutative: \( AB \neq BA \)
Important properties

• Associative: \( A(BC) = (AB)C \)
• Distributive: \( A(B + C) = AB + AC \)
• *Not* Commutative: \( AB \neq BA \)
• Transpose: \( (AB)^T = B^T A^T \)
• The *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that:

$$AA^{-1} = A^{-1}A = I_n$$
• The *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that:

$$AA^{-1} = A^{-1}A = I_n$$

• If $A^{-1}$ exists, then $A$ is called invertible or non-singular
Matrix Inverse

- The *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that:
  $$AA^{-1} = A^{-1}A = I_n$$

- If $A^{-1}$ exists, then $A$ is called invertible or non-singular.
- Matrix $A$ is invertible iff $\det(A) \neq 0$.

Let us solve the house-price prediction problem:

$$\begin{bmatrix} 1000 & 1 \\ 2000 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} 200,000 \\ 350,000 \end{bmatrix}$$
• The *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that:

$$AA^{-1} = A^{-1}A = I_n$$

• If $A^{-1}$ exists, then $A$ is called invertible or non-singular
• Matrix $A$ is invertible iff $\det(A) \neq 0$
• Let us solve the house-price prediction problem
• The *inverse* of a matrix $A \in \mathbb{R}^{n\times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n\times n}$ such that:

$$AA^{-1} = A^{-1}A = I_n$$

• If $A^{-1}$ exists, then $A$ is called invertible or non-singular

• Matrix $A$ is invertible iff $\det(A) \neq 0$

• Let us solve the house-price prediction problem
Matrix Inverse

• The *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that:

$$AA^{-1} = A^{-1}A = I_n$$

• If $A^{-1}$ exists, then $A$ is called invertible or non-singular

• Matrix $A$ is invertible iff $\det(A) \neq 0$

• Let us solve the house-price prediction problem

$$
\begin{bmatrix}
1000 & 1 \\
2000 & 1 \\
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0 \\
\end{bmatrix} =
\begin{bmatrix}
200,000 \\
350,000 \\
\end{bmatrix}
$$
Let us solve the house-price prediction problem

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]

(1)
Let us solve the house-price prediction problem

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  \hspace{1cm} (1)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\left( \begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix} \right)^{-1}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  \hspace{1cm} (2)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
150 & 5 \\
250 & 0
\end{bmatrix}
\]  \hspace{1cm} (3)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
150 \\
50
\end{bmatrix}
\]  \hspace{1cm} (4)
• Let us solve the house-price prediction problem

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix} =
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  

(1)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix} = \left(\begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix}\right)^{-1}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  

(2)

\[
= \frac{1}{-1000}
\begin{bmatrix}
1 & -1 \\
-2000 & 1000
\end{bmatrix}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  

(3)
Let us solve the house-price prediction problem

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= \begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  
(1)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= \left( \begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix} \right)^{-1}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  
(2)

\[
= \frac{1}{-1000} \begin{bmatrix}
1 & -1 \\
-2000 & 1000
\end{bmatrix}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]  
(3)

\[
= \frac{1}{-1000} \begin{bmatrix}
150,000 \\
-5 \times 10^7
\end{bmatrix}
\]  
(4)
Matrix Inverse

- Let us solve the house-price prediction problem

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]

(1)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= \left( \begin{bmatrix}
1000 & 1 \\
2000 & 1
\end{bmatrix} \right)^{-1}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]

(2)

\[
= \frac{1}{-1000}
\begin{bmatrix}
1 & -1 \\
-2000 & 1000
\end{bmatrix}
\begin{bmatrix}
200,000 \\
350,000
\end{bmatrix}
\]

(3)

\[
= \frac{1}{-1000}
\begin{bmatrix}
150,000 \\
-5 \times 10^7
\end{bmatrix}
\]

(4)

\[
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
150 \\
50,000
\end{bmatrix}
\]

(5)
• You could have data from many houses

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1 \\
1500 & 1 \\
\vdots & \vdots \\
2500 & 1 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
w_1 \\
w_0 \\
\end{bmatrix}
= \begin{bmatrix}
200,000 \\
350,000 \\
300,000 \\
\vdots \\
450,000 \\
\end{bmatrix}
\]

\[A \times w = y\]
Norms and Loss Functions

- You could have data from many houses

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1 \\
1500 & 1 \\
\vdots & \vdots \\
2500 & 1 \\
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0 \\
\end{bmatrix}
= 
\begin{bmatrix}
200,000 \\
350,000 \\
300,000 \\
\vdots \\
450,000 \\
\end{bmatrix}
\]

- There isn’t a \( w = [w_1, w_0]^T \) that will satisfy all equations
Norms and Loss Functions

- You could have data from many houses

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1 \\
1500 & 1 \\
\vdots & \vdots \\
2500 & 1
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_0
\end{bmatrix}
= 
\begin{bmatrix}
200,000 \\
350,000 \\
300,000 \\
\vdots \\
450,000
\end{bmatrix}
\]

- There isn't a \( w = [w_1, w_0]^T \) that will satisfy all equations
- Want to find \( w \) that minimizes the difference between \( Aw, y \)
Norms and Loss Functions

- You could have data from many houses

\[
\begin{bmatrix}
1000 & 1 \\
2000 & 1 \\
1500 & 1 \\
\vdots & \vdots \\
2500 & 1 \\
\end{bmatrix}
\]

\[
A \times [w_1, w_0] =
\begin{bmatrix}
200,000 \\
350,000 \\
300,000 \\
\vdots \\
450,000 \\
\end{bmatrix}
\]

- There isn't a \( w = [w_1, w_0]^T \) that will satisfy all equations
- Want to find \( w \) that minimizes the difference between \( Aw, y \)
- But since this a vector, we need an operator that can map the vector \( y - Aw \) to a scalar
Norms and Loss Functions

- A vector norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with

\[ f(x) \geq 0 \quad \text{and} \quad f(x) = 0 \iff x = 0 \]

- $f(ax) = |a|f(x)$ for $a \in \mathbb{R}$

- $f(x + y) \leq f(x) + f(y)$

- e.g., $\ell_2$ norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^{n} x_i^2}$

- e.g., $\ell_1$ norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
A vector norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ with

- $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$

For example, the $\ell_2$ norm:

$$\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$$

And the $\ell_1$ norm:

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

Question: What is the $\ell_1$ norm of $y - Aw$ for the following problem?

$$\begin{bmatrix}
1 & 1 \\
2 & 1 \\
1.5 & 1 \\
2 & 1.5 & 1
\end{bmatrix} = \begin{bmatrix}
2 \\
3 \\
3.5 \\
4 \\
2.5 \\
0
\end{bmatrix}$$

Answer:

$$\|y - Aw\|_1 = 0.5$$
Norms and Loss Functions

- A vector norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with
  - $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  - $f(ax) = |a|f(x)$ for $a \in \mathbb{R}$

- Example:
  - $\ell_2$ norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^{n} x_i^2}$
  - $\ell_1$ norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
Norms and Loss Functions

- A vector norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with
  - $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  - $f(ax) = |a|f(x)$ for $a \in \mathbb{R}$
  - $f(x + y) \leq f(x) + f(y)$

\[ \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2.5 & 1 \end{pmatrix} \begin{pmatrix} 1.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 3.5 \\ 4.5 \\ 3.5 \end{pmatrix} \]

- Example: $\ell_2$ norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$
- Example: $\ell_1$ norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$
Norms and Loss Functions

• A vector norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with
  • $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  • $f(ax) = |a| f(x)$ for $a \in \mathbb{R}$
  • $f(x + y) \leq f(x) + f(y)$

• e.g., $\ell_2$ norm: $\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}$
Norms and Loss Functions

• A vector norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ with
  • $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  • $f(ax) = |a|f(x)$ for $a \in \mathbb{R}$
  • $f(x + y) \leq f(x) + f(y)$

• e.g., $\ell_2$ norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^{n} x_i^2}$
• e.g., $\ell_1$ norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
Norms and Loss Functions

- A vector norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ with
  - $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  - $f(ax) = |a|f(x)$ for $a \in \mathbb{R}$
  - $f(x + y) \leq f(x) + f(y)$
- e.g., $\ell_2$ norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^{n} x_i^2}$
- e.g., $\ell_1$ norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

**Question:** What is the $\ell_1$ norm of $y - Aw$ for the following problem?

$$
\[
\begin{pmatrix}
1 & 1 \\
2 & 1 \\
1.5 & 1 \\
2.5 & 1 \\
\end{pmatrix}
\begin{pmatrix}
2 \\
3.5 \\
3 \\
4.5 \\
\end{pmatrix}
= 
\begin{pmatrix}
1.5 \\
0.5 \\
\end{pmatrix}
\] = y
$$
Norms and Loss Functions

- A vector norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with
  - $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  - $f(ax) = |a|f(x)$ for $a \in \mathbb{R}$
  - $f(x + y) \leq f(x) + f(y)$
- e.g., $\ell_2$ norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$
- e.g., $\ell_1$ norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$  

**Question:** What is the $\ell_1$ norm of $y - Aw$ for the following problem?

\[
\begin{bmatrix}
1 & 1 \\
2 & 1 \\
1.5 & 1 \\
2.5 & 1
\end{bmatrix}
\begin{bmatrix}
1.5 \\
0.5
\end{bmatrix}
= 
\begin{bmatrix}
2 \\
3.5 \\
3 \\
4.5
\end{bmatrix}
\]

- **Answer:** $\|y - Aw\|_1 = 0.5$
Matrix as Linear Transformation (see 3Blue1Brown)

- How exactly does square matrix multiplication transform vectors?

\[
\begin{pmatrix}
1 & 2 \\
4 & 3
\end{pmatrix}
\begin{pmatrix}
1 \\
0
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
4
\end{pmatrix}
\]

\[
\begin{pmatrix}
1 & 2 \\
4 & 3
\end{pmatrix}
\begin{pmatrix}
0 \\
1
\end{pmatrix}
= 
\begin{pmatrix}
2 \\
3
\end{pmatrix}
\]

Now we can express any vector as a linear combination of the above matrix-unit-vector products:

\[
\begin{pmatrix}
1 & 2 \\
4 & 3
\end{pmatrix}
\begin{pmatrix}
2 \\
1
\end{pmatrix}
= 2
\begin{pmatrix}
1 \\
4
\end{pmatrix}
+ 1
\begin{pmatrix}
2 \\
3
\end{pmatrix}
= 
\begin{pmatrix}
4 \\
11
\end{pmatrix}
\]
Matrix as Linear Transformation (see 3Blue1Brown)

- How exactly does square matrix multiplication transform vectors?
- Its columns correspond to re-scaled unit vectors

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
4
\end{bmatrix}
\]
Matrix as Linear Transformation (see 3Blue1Brown)

• How exactly does square matrix multiplication transform vectors?
• It’s columns correspond to re-scaled unit vectors

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
4
\end{bmatrix}
\]
Matrix as Linear Transformation (see 3Blue1Brown)

- How exactly does square matrix multiplication transform vectors?
- It’s columns correspond to re-scaled unit vectors

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
= \begin{bmatrix}
1 \\
4
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
0 \\
1
\end{bmatrix}
= \begin{bmatrix}
2 \\
3
\end{bmatrix}
\]

Now we can express any vector as a linear combination of the above matrix-unit-vector products.
Matrix as Linear Transformation (see 3Blue1Brown)

- How exactly does square matrix multiplication transform vectors?
- It’s columns correspond to re-scaled unit vectors

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix} =
\begin{bmatrix}
1 \\
4
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
0 \\
1
\end{bmatrix} =
\begin{bmatrix}
2 \\
3
\end{bmatrix}
\]

- Now we can express any vector as a linear combination of the above matrix-unit-vector products

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
2 \\
1
\end{bmatrix} = 2 \begin{bmatrix}
1 \\
4
\end{bmatrix} + 1 \begin{bmatrix}
2 \\
3
\end{bmatrix}
\]
• How exactly does square matrix multiplication transform vectors?
• It’s columns correspond to re-scaled unit vectors

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\begin{bmatrix}
1 \\
0 \\
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
4 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\begin{bmatrix}
0 \\
1 \\
\end{bmatrix}
= 
\begin{bmatrix}
2 \\
3 \\
\end{bmatrix}
\]

• Now we can express any vector as a linear combination of the above matrix-unit-vector products

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\begin{bmatrix}
2 \\
1 \\
\end{bmatrix}
= 2 \begin{bmatrix}
1 \\
4 \\
\end{bmatrix}
+ 1 \begin{bmatrix}
2 \\
3 \\
\end{bmatrix}
\]
Matrix as Linear Transformation (see 3Blue1Brown)

- How exactly does square matrix multiplication transform vectors?
- It’s columns correspond to re-scaled unit vectors

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
4
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
0 \\
1
\end{bmatrix}
= 
\begin{bmatrix}
2 \\
3
\end{bmatrix}
\]

- Now we can express any vector as a linear combination of the above matrix-unit-vector products

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
2 \\
1
\end{bmatrix}
= 2
\begin{bmatrix}
1 \\
4
\end{bmatrix}
+ 1
\begin{bmatrix}
2 \\
3
\end{bmatrix}
\]

\[
= 
\begin{bmatrix}
4 \\
11
\end{bmatrix}
\]
• For $A \in \mathbb{R}^{n \times n}$, $\lambda$ is an eigenvalue and $x \neq 0$ is an eigenvector if $Ax = \lambda x$. 
• For $A \in \mathbb{R}^{n \times n}$, $\lambda$ is an eigenvalue and $x \neq 0$ is an eigenvector if $Ax = \lambda x$.

• Eigenvalues are the roots of $\det(A - \lambda I_n) = 0$
Eigenvalues and Eigenvectors

- For \( A \in \mathbb{R}^{n \times n} \), \( \lambda \) is an eigenvalue and \( x \neq 0 \) is an eigenvector if \( Ax = \lambda x \).
- Eigenvalues are the roots of \( \det(A - \lambda I_n) = 0 \).
- Eigenvalues are non-zero solutions of \( Ax = \lambda x \).
Eigenvalues and Eigenvectors

- For $A \in \mathbb{R}^{n \times n}$, $\lambda$ is an eigenvalue and $x \neq 0$ is an eigenvector if $Ax = \lambda x$.
- Eigenvalues are the roots of $\det(A - \lambda I_n) = 0$.
- Eigenvalues are non-zero solutions of $Ax = \lambda x$.
- Viewing $A$ as a linear transformation, the vectors remain unchanged and only get re-scaled are the eigen-vectors. Their scaling factors are the eigen-values!

Question: Find the eigen-values and eigen-vectors of \[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]
Eigenvalues and Eigenvectors

• For $A \in \mathbb{R}^{n \times n}$, $\lambda$ is an eigenvalue and $x \neq 0$ is an eigenvector if $Ax = \lambda x$.

• Eigenvalues are the roots of $\det(A - \lambda I_n) = 0$

• Eigenvalues are non-zero solutions of $Ax = \lambda x$

• Viewing $A$ as a linear transformation
  • The vectors remain unchanged and only get re-scaled are the eigen-vectors.
For \( A \in \mathbb{R}^{n \times n} \), \( \lambda \) is an eigenvalue and \( x \neq 0 \) is an eigenvector if \( Ax = \lambda x \).

Eigenvalues are the roots of \( \det(A - \lambda I_n) = 0 \).

Eigenvalues are non-zero solutions of \( Ax = \lambda x \).

Viewing \( A \) as a linear transformation:

- The vectors remain unchanged and only get re-scaled are the eigenvectors.
- Their scaling factors are the eigenvalues!
• For $A \in \mathbb{R}^{n \times n}$, $\lambda$ is an eigenvalue and $x \neq 0$ is an eigenvector if $Ax = \lambda x$.

• Eigenvalues are the roots of $\det(A - \lambda I_n) = 0$

• Eigenvalues are non-zero solutions of $Ax = \lambda x$

• Viewing $A$ as a linear transformation
  • The vectors remain unchanged and only get re-scaled are the eigen-vectors.
  • Their scaling factors are the eigen-values!

• **Question:** Find the eigen-values and eigen-vectors of

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$
• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\]
• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• **Eigen-values:**
• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• Eigen-values:
• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• **Eigen-values:**

\[
\det \left( \begin{bmatrix}
1 - \lambda & 2 \\
4 & 3 - \lambda
\end{bmatrix} \right) = 0
\]
• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\]

• **Eigen-values:**

\[
\det \left( \begin{bmatrix}
1 - \lambda & 2 \\
4 & 3 - \lambda \\
\end{bmatrix} \right) = 0
\]

\[
(1 - \lambda)(3 - \lambda) - 8 = 0
\]
Question: Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

Eigen-values:

\[
\text{det}
\begin{bmatrix}
1 - \lambda & 2 \\
4 & 3 - \lambda
\end{bmatrix}
= 0
\]

\[
(1 - \lambda)(3 - \lambda) - 8 = 0
\]

\[
\lambda^2 - 4\lambda - 5 = 0
\]
• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• **Eigen-values:**

\[
\text{det}
\begin{pmatrix}
1 - \lambda & 2 \\
4 & 3 - \lambda
\end{pmatrix}
= 0
\]

\[
(1 - \lambda)(3 - \lambda) - 8 = 0
\]

\[
\lambda^2 - 4\lambda - 5 = 0
\]

\[
(\lambda - 5)(\lambda + 1) = 0
\]
Question: Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

Eigen-values:

\[
\det \left( \begin{bmatrix} 1 - \lambda & 2 \\ 4 & 3 - \lambda \end{bmatrix} \right) = 0
\]

\[
(1 - \lambda)(3 - \lambda) - 8 = 0
\]

\[
\lambda^2 - 4\lambda - 5 = 0
\]

\[
(\lambda - 5)(\lambda + 1) = 0
\]

\[
\lambda = 5, \lambda = -1
\]
Eigenvalues and Eigenvectors

• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

Eigen-values: \( \lambda = 5, \lambda = -1 \)

Eigen-vectors:

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = 5
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = \begin{bmatrix}1 \\ 2\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = -1
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = \begin{bmatrix}-1 \\ 1\end{bmatrix}
\]
• **Question**: Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• Eigen-values: \( \lambda = 5, \lambda = -1 \)
Eigenvalues and Eigenvectors

- **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

- Eigen-values: \( \lambda = 5, \lambda = -1 \)
- Eigen-vectors:

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 5
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\]
Eigenvalues and Eigenvectors

• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• **Eigen-values:** \( \lambda = 5, \lambda = -1 \)
• **Eigen-vectors:**

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 5
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= -1
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\]
• Question: Find the eigen-values and eigen-vectors of

\[
\begin{pmatrix}
1 & 2 \\
4 & 3
\end{pmatrix}
\]

• Eigen-values: \( \lambda = 5, \lambda = -1 \)
• Eigen-vectors:

\[
\begin{pmatrix}
1 & 2 \\
4 & 3
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= 5
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= \begin{pmatrix}
1 \\
2
\end{pmatrix}
\]

\[
\begin{pmatrix}
1 & 2 \\
4 & 3
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= -1
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= \begin{pmatrix}
-1 \\
1
\end{pmatrix}
\]

• **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\]

• **Eigen-values:** \( \lambda = 5, \lambda = -1 \)

• **Eigen-vectors:**

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 5
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\implies
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= \begin{bmatrix}
1 \\
2
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= -1
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\]
**Eigenvalues and Eigenvectors**

- **Question:** Find the eigen-values and eigen-vectors of

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\]

- **Eigen-values:** \( \lambda = 5, \lambda = -1 \)

- **Eigen-vectors:**

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
= 5
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
\implies
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
= \begin{bmatrix}
1 \\
2 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & 2 \\
4 & 3 \\
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
= -1
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
\implies
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
= \begin{bmatrix}
-1 \\
1 \\
\end{bmatrix}
\]
• Group the eigen-vectors and eigen values into the following matrices.

\[ P = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \]

\[ \Lambda = \begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix} \]
• Group the eigen-vectors and eigen values into the following matrices.

\[ P = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \]

\[ \Lambda = \begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix} \]
• Group the eigen-vectors and eigen values into the following matrices.

\[ P = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \]

\[ \Lambda = \begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix} \]

• If the eigen-vectors are linearly independent, we can express \( A \) as

\[ A = P\Lambda P^{-1} \]

\[ = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}^{-1} \]
• Why is this useful?
• Why is this useful?
• Suppose we want to find powers of $A$, eg. $A^4$
Eigen-Value Decomposition

• Why is this useful?
• Suppose we want to find powers of $A$, eg. $A^4$
• One option, that is quite tedious is:

$$A^4 = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$
Why is this useful?

Suppose we want to find powers of $A$, eg. $A^4$

One option, that is quite tedious is:

$$A^4 = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$
• Why is this useful?
• Suppose we want to find powers of $A$, eg. $A^4$
• One option, that is quite tedious is:

$$A^4 = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$

• Instead we could use the eigen-value decomposition

$$A^4 = P \Lambda P^{-1} P \Lambda P^{-1} P \Lambda P^{-1} P \Lambda P^{-1}$$
Eigen-Value Decomposition

• Why is this useful?
• Suppose we want to find powers of $A$, eg. $A^4$
• One option, that is quite tedious is:

$$A^4 = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$

• Instead we could use the eigen-value decomposition

$$A^4 = P \Lambda P^{-1} P \Lambda P^{-1} P \Lambda P^{-1} P \Lambda P^{-1}$$
• Why is this useful?
• Suppose we want to find powers of $A$, eg. $A^4$
• One option, that is quite tedious is:

$$A^4 = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$

• Instead we could use the eigen-value decomposition

$$A^4 = P\Lambda P^{-1}P\Lambda P^{-1}P\Lambda P^{-1}P\Lambda P^{-1} = P\Lambda^4 P^{-1} = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 5^4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}^{-1}$$
Singular value decomposition (SVD)

- EVD only works for square, diagonalizable matrices

\[ A = U \Sigma V^\top, \]

- \( U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n} \) are orthogonal matrices (i.e. \( U^\top U = U U^\top = I \))

- \( \Sigma \in \mathbb{R}^{m \times n} \) is a diagonal matrix with singular values of \( A \) denoted by \( \sigma_i \) appearing by non-increasing order:
  \[ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0. \]

- The square singular values of \( A \) are the eigenvalues of the matrix \( AA^\top \) or \( A^\top A \), i.e.,
  \[ \sigma_i(A) = \sqrt{\lambda_i(AA^\top)} = \sqrt{\lambda_i(A^\top A)}. \]

- \( V \) is the matrix of eigen-vectors of \( A^\top A \)

- \( U \) is the matrix of eigen-vectors of \( AA^\top \)
Singular value decomposition (SVD)

- EVD only works for square, diagonalizable matrices
- SVD works for matrices of any size! It decomposes $A \in \mathbb{R}^{m \times n}$ as follows.

$$A = U \Sigma V^\top,$$

- $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $U^\top = U^{-1}$)
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with singular values of $A$ denoted by $\sigma_i$ appearing by non-increasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$.
- The square singular values of $A$ are the eigenvalues of the matrix $AA^\top$ or $A^\top A$, i.e., $\sigma_i(A) = \sqrt{\lambda_i(AA^\top)} = \sqrt{\lambda_i(A^\top A)}$.
- $V$ is the matrix of eigen-vectors of $A^\top A$.
- $U$ is the matrix of eigen-vectors of $AA^\top$. 
Singular value decomposition (SVD)

- EVD only works for square, diagonalizable matrices
- SVD works for matrices of any size! It decomposes $A \in \mathbb{R}^{m \times n}$ as follows.

$$A = U \Sigma V^\top,$$

- $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $U^\top = U^{-1}$)
Singular value decomposition (SVD)

- EVD only works for square, diagonalizable matrices.
- SVD works for matrices of any size! It decomposes $A \in \mathbb{R}^{m \times n}$ as follows.

$$A = U \Sigma V^\top,$$

- $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $U^\top = U^{-1}$).
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with singular values of $A$ denoted by $\sigma_i$ appearing by non-increasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$. 

$\sigma_i$ are the square singular values of $A$. They are the eigenvalues of the matrices $AA^\top$ or $A^\top A$. 

$V$ is the matrix of eigen-vectors of $A^\top A$. 

$U$ is the matrix of eigen-vectors of $AA^\top$. 
Singular value decomposition (SVD)

- EVD only works for square, diagonalizable matrices
- SVD works for matrices of any size! It decomposes $A \in \mathbb{R}^{m \times n}$ as follows.

$$A = U \Sigma V^\top,$$

- $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $U^\top = U^{-1}$)
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with singular values of $A$ denoted by $\sigma_i$ appearing by non-increasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$.
- The square singular values of $A$ are the eigenvalues of the matrix $AA^\top$ or $A^\top A$, i.e., $\sigma_i(A) = \sqrt{\lambda_i(AA^\top)} = \sqrt{\lambda_i(A^\top A)}$
Singular value decomposition (SVD)

• EVD only works for square, diagonalizable matrices
• SVD works for matrices of any size! It decomposes $A \in \mathbb{R}^{m \times n}$ as follows.

$$A = U \Sigma V^T,$$

• $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $U^T = U^{-1}$)
• $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with singular values of $A$ denoted by $\sigma_i$ appearing by non-increasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$.
• The square singular values of $A$ are the eigenvalues of the matrix $AA^T$ or $A^T A$, i.e., $\sigma_i(A) = \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^T A)}$
• $V$ is the matrix of eigen-vectors of $A^T A$
Singular value decomposition (SVD)

- EVD only works for square, diagonalizable matrices
- SVD works for matrices of any size! It decomposes $A \in \mathbb{R}^{m \times n}$ as follows.

$$A = U \Sigma V^\top,$$

- $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $U^\top = U^{-1}$)
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with singular values of $A$ denoted by $\sigma_i$ appearing by non-increasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$.
- The square singular values of $A$ are the eigenvalues of the matrix $AA^\top$ or $A^\top A$, i.e., $\sigma_i(A) = \sqrt{\lambda_i(AA^\top)} = \sqrt{\lambda_i(A^\top A)}$
- $V$ is the matrix of eigen-vectors of $A^\top A$
- $U$ is the matrix of eigen-vectors of $AA^T$
1. Recap: What is Machine Learning?

2. Probability Review

3. A Simple Learning Problem: MLE/MAP Estimation

4. Linear Algebra Review