CSE 123 Summer 2024

LEC 14: Machine Learning

Questions during Class?
Raise hand or send here

sli.do    #cse123

BEFORE WE START

Talk to your neighbors:
How prepared do you feel for the final? What do you need to study?

Music: 123 24su Lecture Tunes ☀️

Instructor: Joe Spaniac
TAs: Andras Daniel
      Eric Nicole
      Sahej Trien
      Zach
Lecture Outline

• Announcements/Reminders

• Machine Learning (ML)
  - Definition / MLE
  - Applications

• ML Pipeline

• Spam Classifier
  - Decision Trees

• Questions to Consider
Announcements

• C3 / R5 feedback released sometime after lecture today
• P3 due tonight (8/7) at 11:59pm
  - Submit something so we can provide some feedback!
• Programming Assignment 4 releases tomorrow (8/8)
  - No resubmission opportunities :(  
  - New assignment, expecting some hiccups – please ask questions if you’re confused!
• Section tomorrow: TAs choice
  - TAs talking about CS topics that interest them
  - Attend any / all that interest you! We have a schedule on the Ed board
• Check-in 4 in section this upcoming Tuesday (8/13)
  - Final exam review, good practice
  - No longer guaranteeing that it will be a problem you’ll see on a quiz
• Resubmission period 7 & 8 release
  - Extra resubmission opportunity open to all previous assignments!
Lecture Outline

• Announcements/Reminders

• Machine Learning (ML)
  - Definition / MLE
  - Applications

• ML Pipeline

• Spam Classifier
  - Decision Trees

• Questions to Consider
What is Machine Learning (ML)?

• Subset of Computer Science concerned with “learning” data trends
What is Machine Learning (ML)?

- Subset of Computer Science concerned with “learning” data trends
- Simple example: maximum likelihood estimation (MLE)

\[ D = (HHTHT) \]

What’s the likelihood of the next coin flip being heads?
\[ \theta = \text{likelihood}, n = \text{flips seen}, k = \text{heads seen} \]

\[ P(D|\theta) = \theta^k (1 - \theta)^{n-k} \]

Goal: find \( \hat{\theta}_{\text{MLE}} \), value that maximizes probability of what we saw
Maximum Likelihood Estimation

\[ P(D|\theta) = \theta^k (1 - \theta)^{n-k} \]

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

\[ \log P(D|\theta) = k \log \theta + (n - k) \log(1 - \theta) \]
\[ \hat{\theta}_{MLE} = \arg\max_\theta k \log \theta + (n - k) \log(1 - \theta) \]

\[ \frac{\partial}{\partial \theta} (k \log \theta + (n - k) \log(1 - \theta)) = 0 \]
\[ k/\theta - n - k/(1 - \theta) = 0 \]
Maximum Likelihood Estimation

\[
k / \theta - n - k / (1 - \theta) = 0
\]

\[
(1 - \theta)k - \theta(n - k) = 0
\]

\[
k - k\theta - \theta n + \theta k = 0
\]

\[
k - \theta n = 0
\]

\[
\hat{\theta}_{MLE} = k / n
\]

Takeaway: There are formal, mathematical ways to verify intuition!
+ We can perform this process with more complicated distributions!
**What is Machine Learning (ML)?**

- Subset of Computer Science concerned with “learning” data trends
- Simple example: maximum likelihood estimation (MLE)
  - As \( n \to \infty \), we know that \( \hat{\theta}_{MLE} \to \theta^* \) (true distribution)
  - With enough data points, we can estimate any statistical distribution!
  - Central limit theorem: all probability distributions are effectively Gaussian...

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
What is Machine Learning (ML)?

• Subset of Computer Science concerned with “learning” data trends

• Simple example: maximum likelihood estimation (MLE)
  - As \( n \to \infty \), we know that \( \hat{\theta}_{MLE} \to \theta^* \) (true distribution)
  - With enough data points, we can estimate any statistical distribution!
  - Central limit theorem: all probability distributions are effectively Gaussian...

Given enough previous examples, we can estimate the underlying distribution and make predictions about... anything!
Applications of ML

- **Opinion Polls**
  - How does a population feel about an issue?

- **Content Recommendation**
  - Can we predict how much someone will like a movie based on past ratings?

- **Object Recognition**
  - Identify {Car, Road, Plane, Bird, Person} within an image?

- **Text Generation**
  - Can computers generate text written like a human?

- **Image Generation**
  - Can computers generate images from a prompt

For each of the following, what would $D$ and $\theta^*$ represent?
Lecture Outline

• Announcements/Reminders

• Machine Learning (ML)
  - Definition / MLE
  - Applications

• ML Pipeline

• Spam Classifier
  - Decision Trees

• Questions to Consider
ML Pipeline

• Generally, building an ML model involves the following steps:
  
  - Data Collection / Sanitization
  - Featurization
  - Model Training / Tuning
  - Evaluation
  - Deployment

• Notice that you can step backwards!
  - ML in particular is an applied science, it’s all an experiment!
1. Data Collection

- **We need** example data to understand a distribution
  - Lots and lots of it too \((n \to \infty)\)

- **Where does this data come from?**
  - Language: Reddit, Twitter, Facebook, Wikipedia, Blogs, etc.
  - Images: Google, Twitter, Websites
  - Code: Github
  - Really, anywhere publicly (or not) accessible on the Internet

- **Who determines what data is used?**  
  - Often companies buy preprocessed data from others
  - Let’s say that you accidentally post your phone number on your twitter
    - A model could scrape that info, memorize it, and regurgitate it when prompted

- **Data carries PII / bias that we need to account for**
Data Bias

- Image results for searching the term “CEO” on Google (2015)
  - Notice anything about the results?

Data Bias

• Fix: Image results for searching “CEO” and “CEO United States” (2022)

https://www.washington.edu/news/2022/02/16/googles-ceo-image-search-gender-bias-hasnt-really-been-fixed
Data Sanitization

- Data carries PII / bias that we need to account for
- We don’t want our model to memorize a phone number
  - Let’s just remove all phone numbers from our inputs!
  - Is this an effective solution?

- Sanitization can be ethically gray – does it disproportionately affect subpopulations?
  - E.g. Swear words & AAVE
  - Correlated features

Our models are only as strong as the data they’re built upon. Garbage in, garbage out.
2. Featurization

• Now that we have all our data, we need to convert it into something a computer can understand (numbers)
  - How can we convert text / images into numbers?

• Determine what aspects of the data interest you (features)

• Words can be “vectorized”
  - Converted into $n$-dimensional vectors $n \in \{50, 200, 500, \ldots \}$
  - Determined from the word2vec algorithm

• Images are already numbers... (2d array of RGB values)
Word Embeddings

• We call these word vectors “embeddings” and they’re pretty interesting to mess around with

• Can perform mathematical operations on them
  - Find the nearest vectors to any given word (synonyms)
  - Compute comparisons (dog is to puppy as cat is to ___)
    - Take the difference between puppy and dog (age vector) and add it to cat
    - Find the nearest vectors to the result and you’ll likely see “kitten”

• These operations can further reveal bias
  - man is to doctor as woman is to ______
  - Any model trained from biased data points will estimate a biased distribution
3. Model Training

• Pick some way of using data to estimate
• Lots of different flavors of this
  - Regression (linear, logistic)
  - Neural Networks (CNNs, RNNs, Transformer, etc.)
  - Nearest neighbors
  - Decision trees
• Provide additional computation (memory / GPUs / time) until desired result is achieved

*It’s all one big experiment – try options until something sticks.*

*This should feel somewhat concerning...*
4. Evaluation

• Does your model actually work?

• Typically we split our initial dataset into 3 different subsets:
  - Train (provided to the model during training)
  - Test (used after a model has trained to compare to previous iterations)
  - Validation (used once a model has been chosen to see how it performs)

• Determine whether or not your model is over / underfitting

• Most ML applications go no further than this step
  - No attempt to determine *why* a particular model is working well
5. Deployment

• Put your model out into the real world and see what happens
  - Does it perform the job as expected? Should further work be put into development?

• At this point, often the next iteration of refinement takes place
  - GPT 2.0 -> 3.0 -> 3.5 -> 4.0
  - Options include:
    - Collect more data, use more compute, discover better tuning, discover better model

• Often, not much effort is put into understanding negative impacts
  - Case in point: ChatGPT and the education system
ML Pipeline

- That’s it – in essence, that’s how every ML model is created

Does this knowledge change your perspective on ML / AI?
Lecture Outline

• Announcements/Reminders

• Machine Learning (ML)
  - Definition / MLE
  - Applications

• ML Pipeline

• Spam Classifier
  - Decision Trees

• Questions to Consider
SpamClassifier

• Your last programming assignment will involve part 3 of this pipeline
  - You’ll implement a *decision tree* capable of detecting spam emails

• Extra credit involves steps 1/2
  - Finding and featurizing another dataset
Decision Trees

- Tree structure where each intermediary node contains a feature/threshold pair (split) and leaf nodes are labels.
Decision Trees

- Let’s say we wanted to classify the following
  - “hello, i am here at your office but the door is locked. are you there?”
Decision Trees

- Let’s say we wanted to classify the following
  - “hello, i am here at your office but the door is locked. are you there?"
Decision Trees

• Let’s say we wanted to classify the following
  - “hello, i am here at your office but the door is locked. are you there? 
Decision Trees

• Let’s say we wanted to classify the following
  - “hello, i am here at your office but the door is locked. are you there?

```
Key
Split
Label

Input
here: 0.0667
dolphin: 0.00
you: 0.0667

Overall Root
wordPercent~here
0.1

wordPercent~dolphin
0.0375

wordPercent~you
0.05814

Ham

Spam

Ham

Ham
```
Lecture Outline

• Announcements/Reminders

• Machine Learning (ML)
  - Definition / MLE
  - Applications

• ML Pipeline

• Spam Classifier
  - Decision Trees

• Questions to Consider
Questions to Consider

• Are ML Models actually capable of “learning” anything?
  - I.e. is it possible to “learn” just by observing / memorizing?
  - Does ChatGPT actually “understand” language?

• If all output from ML models is based on previous examples, who gets credit / takes responsibility for generation?
  - Think AI art and your C3 / P3 reflection responses

• What harm could come from deploying ML models we don’t fully understand?

• If society itself is biased, how much should we worry about the bias present in data / ML models?
  - To what extent should concern about bias hinder further advancements?