Logistics

• Homework 5 is due Dec 2 (Friday)

• Exam 2 will be on Dec 8 (Thursday)
  • Will focus on the topics in the second half of the semester
    • Note though knowledge is cumulative, so we still assume you know the concepts earlier
  • Format / logistics will be similar with what we have in Exam 1
    • Timed exam (75 min) during lecture time in the classroom
    • Closed-book exam with 2 letter-size cheat sheets allowed (4 pages in total)
      • No format limitations (it can be typed, written, or a combination)
  • Dec 6 (Tuesday) will be a review lecture
Recap
Neural Network

• Evaluate $h(\vec{x})$ given $h$ (characterized by $\{w_{i,j}^{(\ell)}\}$)
  • Forward propagation
    $$x = x^{(0)} \xrightarrow{w^{(1)}} s^{(1)} \xrightarrow{\theta} x^{(1)} \xrightarrow{w^{(2)}} s^{(2)} \xrightarrow{\theta} x^{(2)} \cdots \xrightarrow{w^{(L)}} s^{(L)} \xrightarrow{\theta} x^{(L)} = h(x)$$

• Given $D$, learn the weights $W = \{w_{i,j}^{(\ell)}\}$
  • Backpropagation
  • Initialize $w_{i,j}^{(\ell)}$ randomly
  • For $t = 1$ to $T$
    • Randomly pick a point from $D$ (for stochastic gradient descent)
      • Forward propagation: Calculate all $x_i^{(\ell)}$ and $s_i^{(\ell)}$
      • Backward propagation: Calculate all $\delta_j^{(\ell)}$
    • Update the weights $w_{i,j}^{(\ell)} \leftarrow w_{i,j}^{(\ell)} - \eta \delta_j^{(\ell)} x_i^{(\ell-1)}$
  • Return the weights
Deep Neural Network

• “Shallow” neural network is powerful (universal approximation theorem holds with a single hidden layer). Why “deep” neural networks?

Each layer captures features of the previous layers.

We can use “raw data” (e.g., pixels of an image) as input. The hidden layer are extracting the features.

Design different network architectures to incorporate domain knowledge.
Some Techniques in Improving Deep Learning

• Regularization to mitigate overfitting
  • Weight-based, early stopping, dropout, etc

• Incorporating domain knowledges
  • Network architectures (e.g., Convolutional Neural Nets)

• Improving computation with huge amount of data
  • Hardware architecture to improve parallel computation

• Improving gradient-based optimization
  • See more in LFD 7.5 (Steepest descent, conjugate gradient, higher-order optimization)
  • Choosing better initialization points
Machine Learning Lifecycle

- Feedback
- Task Definition
- Deployment
- Dataset Construction
- Model Definition
- Training
- Testing
For ML to have “positive” impacts, we need to be careful in every stage.

What we covered (and majority of ML research)
Classification

• Standard setup of (supervised) machine learning

• Finding patterns from the given training datasets
• Use the pattern to make predictions on new testing data

Key assumption of ML

Training data points and testing data points are i.i.d. drawn from the same (unknown) distribution
Strategic Classification
Machine Learning Lifecycle

- Model Definition
- Dataset Construction
- Task Definition
- Model Definition
- Deployment
- Dataset Construction
- Testing
- Training
- Feedback
Today’s Lecture
Modern ML is driven by data.

Where does data come from?
CAPTCHA

Completely Automated Public Turing test to tell Computers and Humans Apart
Can we utilize this wasted human computation power?
The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

Word 1: an OCR task to solve
Word 2: tell apart humans and bots

“reCAPTCHA has completely digitized the archives of The New York Times and books from Google Books, as of 2011”

More than recognizing text

- Google acquired reCAPTCHA in 2009.
Data is often generated by humans
Explicitly: Human Labelers

- Amazon Mechanical Turk: Artificial Artificial Intelligence
  - A marketplace to collect data from humans
  - E.g., ImageNet has utilized this platform to collect image labels
Data (labeled or generated by humans) is the main driving force of ML

**Good**: Humans help drive ML forward

**But?**
Task: Acquire Image Labels [Otterbacher et al. 2019]

- Label distributions are different for images of different gender/race
  - Female images receive more labels related to the “attractiveness”.

Data (labeled or generated by humans) is the main driving force of ML

**Good**: Humans help drive ML forward

**Bad**: ML becomes an amplifier of human biases
Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy

Authors: Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, Olga Russakovsky

Microsoft Release a Twitter Chatbot in 2016

@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

@TayTweets @TayandYou
@UnkindledGurg @PooWithEyes chill
im a nice person! i just hate everybody

@TayTweets @TayandYou
@NYCitizen07 I fucking hate fe and they should all die and burn
24/03/2016, 11:41

@TayTweets @TayandYou
@ReynTheo HITLER DID NOT

@Sardor9515 well I learn from the best ;) if you don't understand that let me spell it out for you
I LEARN FROM YOU AND YOU ARE DUMB TOO
10:28 AM - 23 Mar 2016
Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source TayandYou (Twitter)

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

BUSINESS NEWS  OCTOBER 9, 2018 / 10:12 PM / A YEAR AGO

Voice Is the Next Big Platform, Unless You Have an Accent

It’s super funny that Alexa can’t understand my mom — until we need Alexa to use the web, drive a car, and do pretty much anything else.
There are also privacy concerns...

- Matchin: A Game for Collecting User Preferences on Images

- Building gender models using user labels

- Ask MTurk workers to compare 10 pairs of images.
  - Accuracy for guessing the gender: 78.3%
Machine learning models leak personal info if training data is compromised

Attackers can insert hidden samples to steal secrets

Katyanna Quach

AI unmask anonymous chess players, posing privacy risks

Software that identifies unique playing styles could lead to better tutorials and game play

12 JAN 2022 • 2:30 PM • BY MATTHEW HUTSON
Machine Learning Lifecycle

- Model Definition
- Training
- Testing
- Deployment
- Feedback
- Task Definition
- Dataset Construction
- Is there privacy leak?
- Are we being fair to humans?

Training → Testing → Dataset Construction → Task Definition → Feedback → Model Definition → Deployment → Are we being fair to humans? → Is there privacy leak?
Discussion on Privacy
Netflix Challenges

• Netflix challenge
  • Announced in 2006
  • Released a dataset of 100,480,507 ratings that 480,189 users gave to 17,770 movies.
  • Award $1 million to first team beating their algorithm by 10%
  • Data format: <user, movie, date of grade, grade>
    • User and movie names are replaced with integers

• Is there a second Netflix challenge?
  • Announced in August 2009
  • Cancelled in March 2010
  • Why?
    • Privacy lawsuits and FTC involvements
Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Netflix Dataset + IMDB Data
Why is Anonymization Hard?

- Even without explicit identifiable information (e.g., ID, name), other detailed information about you might still reveal who you are.

<table>
<thead>
<tr>
<th>office</th>
<th>department</th>
<th>date joined</th>
<th>salary</th>
<th>d.o.b.</th>
<th>nationality</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>IT</td>
<td>Apr 2015</td>
<td>£###</td>
<td>May 1985</td>
<td>Portuguese</td>
<td>Female</td>
</tr>
</tbody>
</table>

- What can we do?
  - Adding noises

<table>
<thead>
<tr>
<th>office</th>
<th>department</th>
<th>date joined</th>
<th>salary</th>
<th>d.o.b.</th>
<th>nationality</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK</td>
<td>IT</td>
<td>2015</td>
<td>£###</td>
<td>1980-1985</td>
<td>—</td>
<td>Female</td>
</tr>
</tbody>
</table>

Tradeoff between privacy and utility.
Unreasonable Privacy Expectations

• Can we get privacy for free?
  • No, privatizing means information loss (=> accuracy loss)

• Absolute privacy is not likely.
  • E.g., who you are friends with might reveal who you are

MIT Students’ Facebook ‘Gaydar’ Raises Privacy Issues
Differential Privacy

• A formal notion to characterize privacy.

• History
  • Proposed by Dwork et al. 2006
  • Win the Gödel Prize in 2017
  • Apple announced to adopt the notion of differential privacy in iOS 10 in 2016
Differential Privacy
Differential Privacy

- I have provided our exam distribution for Exam 1 on Piazza.

- How much of the private information (your individual grades) do I reveal?
- What if there are only 2 students in the class?
Differential Privacy

Data → Analysis (Computation) → Outcome

Data w/ your info removed → Analysis (Computation) → Outcome

$\epsilon$="similar"
Differential Privacy

• Notations
  • $A$: a randomized algorithm.
  • $D_1, D_2$: two “neighboring” database (with only one-entry difference)
  • $\epsilon$: privacy budget

• $\epsilon$-differentially private
  • $A$ is $\epsilon$-differentially private if for any neighboring databases $D_1$ and $D_2$, and for any algorithm output $Y$, we have

$$\Pr[A(D_1) \in Y] \leq e^\epsilon \Pr[A(D_2) \in Y]$$

Intuition:
The change of output is small if the change of data is small.
How to Be Differentially Private

• Let the output of $A$ be the average of students’ grades
• Consider two extreme cases
  • If the size of the database is infinity
  • If the size of the database is 1

• Add noise
  • We can tune the amount of noise to tradeoff privacy and accuracy

• The majority of the differentially private algorithms use a similar approach
Discussion on Fairness
Cucumbers and Grapes Experiments

• https://youtu.be/-KSryJXDpZo
BRIEF HISTORY OF FAIRNESS IN ML

PAPERS


LOL FAIRNESS!!

OH, CRAP.
Isn’t the point of ML to discriminate?

Want to avoid “unjustified” discrimination.
Example: Loan Applications

• By law, banks can’t discriminate people according to their race.
• First natural approach (fairness through blindness)
  • remove the race attribute from the data
• Guess what happened?
  • Redlining
What should we do?

- From computer scientists / engineers’ point of view....
  - Give me an operational definition of fairness, I’ll implement a system that satisfy it!

- One potential approach:
  - Minimize error subject to fairness constraints (Recall regularizations)
    \[ \text{minimize } Error(\overline{w}) \text{ subject to fairness constraints} \]
    \[ \text{minimize } Error(\overline{w}) + \lambda \ast [\text{fairness violations}] \]

- Several recent research and open-source libraries are done this way
  - Fairlearn: A toolkit for assessing and improving fairness in AI
  - GerryFair: Auditing and Learning for Subgroup Fairness
  - ...
How should we define fairness?
Another Example: Probation Decisions

• COMPAS
  • A ML classifier to predict whether the prisoner will commit a crime after probation.
Controversy and Debates

- ProPublica (a non-profit institution)
  - COMPAS is not fair!

<table>
<thead>
<tr>
<th>Labeled Higher Risk, But Didn’t Re-Offend</th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>
Controversy and Debates

• Northpointe (company that develops COMPAS)
  • COMPAS is fair!
Impossibility Result [Kleinberg et al. 2017]
The above fairness conditions (together with similar variations) cannot be satisfied simultaneously, unless the predictor is perfect or the two groups are the same.
<table>
<thead>
<tr>
<th></th>
<th>Labeled Low-Risk</th>
<th>Labeled High-Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Won’t Recidivate</td>
<td>TN1</td>
<td>FN1</td>
</tr>
<tr>
<td>Will Recidivate</td>
<td>FN1</td>
<td>TP1</td>
</tr>
</tbody>
</table>

- **Defendant**: the probability that I’m incorrectly classified high-risk is independent of my race.
  - **Equal False Positive Rate**:
    \[
    \frac{FP1}{TN1 + FP1} = \frac{FP2}{TN2 + FP2}
    \]

- **Defendant**: the probability that I’m incorrectly classified as low-risk is independent of my race.
  - **Equal False Negative Rate**:
    \[
    \frac{FN1}{FN1 + TP1} = \frac{FN2}{FN2 + TP2}
    \]

- **Decision-maker**: the ratio of people who recidivated among the ones labeled high-risk is independent of race.
  - **Equal Predictive Value**:
    \[
    \frac{TP1}{TP1 + FP1} = \frac{TP2}{TP2 + FP2}
    \]

**Impossibility Result [Kleignberg et al. 2017]**

The above three conditions cannot be satisfied simultaneously, unless the predictor is perfect or the two groups are the same.
The Same Impossibility Results Applies to Other Sets of Fairness Definitions

• Another setup
  • A: Sensitive attributes (e.g., race)
  • Y: True labels (e.g., commit a crime in the future)
  • C: Predictions (e.g., predictions of recidivism)

• Criteria:
  • C independent of A
  • C independent of A conditional on Y
  • Y independent of A conditional on C

Impossible to satisfy them simultaneously.
The Same Impossibility Results Applies to Other Sets of Fairness Definitions

- Another setup
  - \( A \): Sensitive attributes (e.g., race)
  - \( Y \): True labels (e.g., commit a crime in the future)
  - \( C \): Predictions (e.g., predictions of recidivism)

**Criteria:**
- \( C \) independent of \( A \)
- \( C \) independent of \( A \) conditional on \( Y \)
- \( Y \) independent of \( A \) conditional on \( C \)

Impossible to satisfy them simultaneously.
More Examples

[Kay et al., 2015]
Stereotype Mirroring and Exaggeration

• Is this result mirroring the real statistics or an exaggeration?

• Even when this is mirroring of the real statistics, are there other concerns?
  • Are we reinforcing the stereotypes?
  • Are we being “unfair” to disadvantage groups that are mistreated in the past?

Other Types of Fairness: Individual Fairness

• Similar people should be treated similarly

• Challenges
  • What do we mean by similar people
    • Need to define some kind of “distance” measure
  • What do we mean by being treated similarly
    • Decisions based on threshold won’t work
    • Need to impose some “smooth” notion
    • Randomization is often required
Other Types of Fairness: Counterfactual Fairness

A decision is fair towards an individual if it gives the same predictions in:

(a) the observed world and
(b) a world where the individual had always belonged to a different demographic group

I understood gender discrimination once I added “Mr.” to my resume and landed a job

Woman Who Switched to Man’s Name on Resume Goes From 0 to 70 Percent Response Rate
Other Types of Fairness: Procedural Fairness (Procedural Justice)

- **Neutrality**: Decisions are unbiased and guided by transparent reasoning.
- **Respect**: All are treated with respect and dignity.
- **Voice**: All are given a chance to tell their side of the story.
- **Trustworthiness**: Decision makers convey trustworthy motives about those impacted by their decisions.
Take-Aways

• ML is a powerful tool to help extract patterns from data.
  • If you have data, ML might be able to help!

• However, ML may also be an amplifier of human biases
  • Biases could creep in through many stages of the ML life cycle, such as data, task definition, model choice, parameter tuning, ...

• No silver bullet (yet)
  • Being aware of the issues is the important first step
  • ”Solving” the issues (if at all possible) requires communications among people in different disciplinaries
An Emerging Research Agenda on AI/ML + Humans/Society

• WashU Division of Computational and Data Sciences
  • A new PhD program hosted by CSE, Political Science, Social Work, Psychology and Brain Science

• MIT Institute for Data, Systems, and Society
• CMU Societal Computing
• Stanford Institute for Human-Centered Artificial Intelligence
• USC Center for AI in Society

• ACM FAT* (Fairness, Accountability, and Transparency)
• AAAI/ACM AIES (AI, Ethics, and Society)
Course Wrap-Up
Revisit Our Course Plan

• Foundations
  • What’s machine learning
  • Feasibility of learning
  • Generalization
  • Linear models
  • Non-linear transformations
  • Overfitting and how to avoid it
    • Regularization
    • Validation

• Techniques
  • Decision tree
  • Ensemble learning
    • Bagging and random forest
    • Boosting and Adaboost
  • Nearest neighbors
  • Support vector machine
  • Neural networks
  • …
There are a lot more...