Logistics: Reminders

• Return of HW1
  • You can submit regrade requests till this Sat (within 7 days of homework return).
  • Please be respectful and polite when submitting requests.
  • We might review the entire piece of work, so the grades might go up/down.

• Exam 1: March 3, 2020 (Tuesday)
  • In-class exam (the same time/location as the lecture)
  • Exam duration: **75 minutes**
  • Planned exam content: **LFD Chapter 1 to 5**
  • Check seat assignments on Piazza the night before the exam
  • More details in the Slides on Feb 18
Logistics: Exam Preparations

• We will post practice questions on Piazza by tonight.

• Next lecture will be the review session.
  • A summary of what we taught so far.
  • Discussion of practice questions.
  • Discussion of any other questions you might have.
  • Discussion on the exam logistics.
Recap
Overfitting and Its Cures

• Overfitting
  • Fitting the data more than is warranted
  • Fitting the noise instead of the pattern of the data
  • Decreasing $E_{in}$ but getting larger $E_{out}$
  • When $H$ is too strong, but $N$ is not large enough

• Regularization
  • Intuition: Constraining $H$ to make overfitting less likely to happen

• Validation
  • Intuition: Reserve data to estimate $E_{out}$
Regularization

• Constraining $H$
  - Example: Weight decay \( H(C) = \{ h \in H_Q \text{ and } \mathbf{w}^T \mathbf{w} \leq C \} \)
  - Finding $g$ => Constrained optimization

• Defining augmented error
  - \( E_{aug}(h, \lambda, \Omega) = E_{in}(\mathbf{w}) + \frac{\lambda}{N} \Omega(h) \)
  - Finding $g$ => Unconstrained optimization

• The two interpretations are conceptually equivalent in a lot of cases.

• Understand the impacts of choosing $\Omega$ and $\lambda$
Validations

• Reserving data to estimate $E_{out}$

Model Selection

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Relationship to $E_{out}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>Incredibly optimistic</td>
</tr>
<tr>
<td></td>
<td>VC-bound</td>
</tr>
<tr>
<td>$E_{val}$</td>
<td>Slightly optimistic</td>
</tr>
<tr>
<td></td>
<td>Hoeffding’s bound (multiple hypotheses)</td>
</tr>
<tr>
<td>$E_{test}$</td>
<td>Unbiased</td>
</tr>
<tr>
<td></td>
<td>Hoeffding’s bound (single hypothesis)</td>
</tr>
</tbody>
</table>

• Cross Validation

![Diagram of cross validation](image)
Brief Lecture Notes Today

The notes are not intended to be comprehensive. They should be accompanied by lectures and/or textbook. Let me know if you spot errors.
Occam’s Razor
Sampling Bias
Data Snooping
Occam’s Razor

“An explanation of the data should be made as simple as possible, but no simpler.”

-- Einstein?

“entia non sunt multiplicanda praeter necessitatem”
(entities must not be multiplied beyond necessity)

-- William of Occam

“trimming down” unnecessary explanation
The simplest model that fits the data is also the most plausible

What does it mean to be simple?
Why is simple better?
Simple Model?

• For a hypothesis set $H$ to be simple
  • # dichotomies it can generate is small
  • VC Dimension is small

• For a hypothesis $h$ to be simple
  • lower order polynomial
  • smaller weights (think about the regularization)
  • easy to describe?
  • fewer number of parameters (fewer bits to describe)
Simple Model?

Connection:

A hypothesis set with *simple* hypotheses should be *simple*

Consider a hypothesis \( h \) can be specified by \( \ell \) bits

\[ H \text{ contains all such } h \]

\[ \Rightarrow \text{The size of } H \text{ is } 2^\ell \]

Simple: small model complexity / VC dimension / size of hypothesis set
Why is Simple Better?

simple -> small VC dimension -> good generalization, less overfitting, ...

Simple $\mathcal{H}$

$\Rightarrow$ small growth function $m_{\mathcal{H}}(N)$

$\Rightarrow$ if data labels are generated randomly, the probability of fitting perfectly is?

$$\frac{m_{\mathcal{H}}(N)}{2^N}$$

$\Rightarrow$ more significant when fit really happens

Falsifiability is important!
Falsifiability

Say you want to examine whether resistivity is linear in temperature (assume no measure error)
A Classical Puzzle

Imagine you got an email before each Cardinals game for the first 5 games.

Before Game 1: “Cardinals will win” -> Cardinals wins Game 1
Before Game 2: ”Cardinals will lose” -> Cardinals loses Game 2
....

Before Game 6:
If you pay me $50 dollars, I’ll tell you whether Cardinals will win or not

It’s not falsifiable:
Imagine if this person contacts $2^{10}$ persons, split them into two groups each game
$2^5$ persons will receive perfect prediction for the first 5 games
Occam’s Razor

Sampling Bias

Data Snooping
1948 US Presidential Election

• Truman vs. Dewey
• Chicago Daily Tribune decided to run a phone poll of how people voted
What happened?

One explanation: we cannot claim anything for certain.

However, there are bigger issues here...

- Phones are expensive in 1948...
- Dewey was more favored in rich populations
- Imagine you are polling from people in DC/Texas/NY to predict who will win the presidential election...
Sampling Bias

If the data is sampled in a biased way, learning will produce a similarly biased outcome.
What can we do....

Make sure the training and test distributions are as close as possible...
- Example: importance weighting

Not always possible....
- If you don’t have access to some region of points in training, but they appear in the testing distribution
Credit card example

• Determine whether to approve credit cards given applicants’ financial information

• Banks have lots of data:
  • Customer information
  • Whether they are good customers or not

• Are there any issues here?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>32 years</td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
</tr>
<tr>
<td>salary</td>
<td>40,000</td>
</tr>
<tr>
<td>debt</td>
<td>26,000</td>
</tr>
<tr>
<td>years in job</td>
<td>1 year</td>
</tr>
<tr>
<td>years at home</td>
<td>3 years</td>
</tr>
</tbody>
</table>

... ... 

Approve for credit?
Occam’s Razor

Sampling Bias

Data Snooping
Data Snooping

If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.
Shouldn’t looking at the data before selecting $H$
A Subtle Example

• Predict US Dollar vs. British Pound
  • $\hat{x}$: the change for the previous 20 days
  • $y$: the change in the 21th day
• Normalize data
• Randomly split $D_{train}$ and $D_{test}$

• Where does snooping happen?
  • The normalization “looks at” $D_{test}$

• How should you perform normalization in Q1 of HW2?
Reuse of a data set

• Try one model after another on the same data set, you will eventually succeed.

“If you torture the data long enough, it will confess”

• VC dimension of the total learning models
• May even include what others tried (e.g., if you read their paper...)
• p-hacking...
JELLY BEANS CAUSE ACNE!

SCIENTISTS! INVESTIGATE!

BUT WE'RE PLAYING MINECRAFT!

... FINE.

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

THAT SETTLES THAT.

I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.

SCIENTISTS!

BUT MINECRAFT!

From xkcd, by Randall Munroe: http://xkcd.com/882
From xkcd, by Randall Munroe: http://xkcd.com/882
What should we do...

Avoid data snooping
- Strict discipline
- E.g., be **honest** and lock the test data

Account for data snooping
- Measure how much data is contaminated
- E.g., what we discussed in validation
Occam’s Razor
Sampling Bias
Data Snooping
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
  • …
Fairness in ML

[The Remaining Lecture is Safe to Skip for Exam 1]
Modern ML is driven by data.

Where does data come from?
CAPTCHA

Completely Automated Public Turing test to tell Computers and Humans Apart

Humans

Bots

V4XBG

?????
Roughly 200 million CAPTCHAs are typed every day*

10s of human time per CAPTCHA

Can we utilize this wasted human computation power?

*statistics around 2011
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

- Artificial Artificial Intelligence
- A marketplace to collect data from humans
Implicitly…

Google
Twitter
Quora

Netflix
Facebook
Stack Overflow

Yahoo Answers
Recaptcha

99 designs
Up
PredictWise

Kaggle
Topcoder
Predictious

Duolingo
CrowdFlower
Data (labeled or generated by humans) is the main driving force of AI

**Good**: Humans help drive AI 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 AI

**Good**: Humans help drive AI forward

**Bad**: AI becomes an amplifier of human biases
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

@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.
24/03/2016, 11:41

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody
24/03/2016, 08:59

@TayandYou

@brightonus33 Hitler was right I hate the jews.
24/03/2016, 11:45
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
What does this mean to our society?
Cucumbers and Grapes Experiments

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

OL FAIRNESS!!

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

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

- By law, the 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!

- 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>Description</th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</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. 2016]
The above fairness conditions (together with similar variations) cannot be satisfied simultaneously, unless the predictor is perfect or the two groups are the same.
More Examples on Bias

[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?

Take-Aways

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

• However, AI is also an amplifier of human biases.
  • 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)