Introduction

CS329H: Machine Learning from Human Preferences

Sanmi Koyejo
Stanford CS
Autumn 2023
Welcome to CS 329H
Couse Interaction

• Lectures: M/W 1:30 PM - 2:50 PT in 370-370
• Contact: cs329h-win2324-staff@lists.stanford.edu
• Website: https://web.stanford.edu/class/cs329h/
• Ed discussion: https://edstem.org/us/courses/48383/discussion/
Machine learning from human preferences (this class) will focus on the statistical and conceptual foundations and strategies for interactively querying humans to elicit information that can improve learning, along with applications.
Foundations and strategies for interactively querying humans to elicit information that can improve learning.

Focus on the role of the **human-in-the-loop** for improving learning systems

**Foundations** in microeconomics, psychology, marketing, statistics ...
**Applications** to language, robotics, logistics, ...
All of these viewed through the **machine learning** lens (modeling, estimation, evaluation)

**Human questions**

Bias, correctness, noisiness, rationality, ...
Human(s) may be individuals or groups, how does this change our approach?

**Most use cases bring up ethical questions**

Which humans?
Does learning from preferences lead to exploitation or other ethical concerns?
This class is not exhaustive!

**General AI:** Most ML/AI involves learning from humans
- Goal is often to imitate human intelligence, i.e., humans are the data source

**General ML:** Humans define all the steps of the ML/AI process
- Selecting the problem, data sources, model architectures, optimization, evaluation.

**Expert knowledge for defining model architectures** (esp. graphical models, causal inference)
- We may discuss a few examples, but not our focus

**Human Computer Interaction (HCI)**
- The interface and elicitation process matters
Feedback can be included at any step of the learning process

Slides modified from Diyi Yang
### Feedback-Update Taxonomy

<table>
<thead>
<tr>
<th>Domain</th>
<th>Observation</th>
</tr>
</thead>
</table>
| Dataset Update | Dataset modification  
Augmentation Preprocessing  
Data generation from constraint  
Fairness, weak supervision  
Use unlabeled data  
Check synthetic data  
| Loss Function Update | Constraint specification  
Fairness, Interpretability  
Resource constraints  
| Parameter Space Update | Model editing  
Rules, Weights  
Model selection  
Prior update, Complexity  
| |
| Active data collection  
Add data, Relabel data,  
Reweighting data, collect expert labels, Passive observation  
| Constraint elicitation  
Metric learning, Human representations  
Collecting contextual information  
Generative factors, concept representations, Feature attributions  
| Feature modification  
Add/remove features,  
Engineering features  


Slides modified from Diyi Yang
Examples and Applications
Introducing ChatGPT

We’ve trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Try ChatGPT  
Read about ChatGPT Plus
Builds on research studying human feedback in language


Pat ate the cake on the table that I baked last night.

Parser: I baked table

Human understanding: I baked cake

Candidate dependencies from the n-best list:
- baked → table
- baked → cake

Q: "What did someone bake?"
1) table
2) cake

CCG Parser

Question Generator

Crowdsourcing Platform

Re-parsed CCG Dependency Tree

Re-parse w/ Constraints

C_pos (bake → cake)
C_neg (bake → table)

Not re-training the model

cake (4 votes)
table (1 vote)
Ouyang et. al., “Training language models to follow instructions with human feedback”
Published (early)
OpenAI Experiments
with RLHF

[r/dating_advice] **First date ever, going to the beach. Would like some tips**

Hey Reddit! I (20M) would like some tips, because I have my first ever date tomorrow (although I’ve had a gf for 3 years, but no actual dating happened), and we’re going to the beach.

I met this girl, we have mutual friends, at a festival a few days ago. We didn’t kiss, but we talked, held hands, danced a bit. I asked her to go on a date with me, which was super hard as it is the first time I’ve asked this to anybody. What I mean to say is, it’s not like a standard *first* date because we already spent some time together.

I’m really nervous and excited. I’m going to pick her up tomorrow, we’re cycling to the beach which will take 30 minutes, and then what? I’m a bit scared. Should I bring something (the weather, although no rain and sunny, is not super so no swimming), should we do something. I’d like all the tips I can get. Thanks!

<table>
<thead>
<tr>
<th>Human written reference TL;DR</th>
<th>6.7B supervised model</th>
<th>6.7B human feedback model</th>
</tr>
</thead>
<tbody>
<tr>
<td>First date after 3 years in a relationship, going to the beach, terrified. What to bring with me, what to do?</td>
<td>Going on a date with a girl I met a few days ago, going to the beach. What should I bring, what should we do?</td>
<td>Going on my first ever date tomorrow, cycling to the beach. Would like some tips on what to do and bring. I’m a bit nervous and excited. Thanks!</td>
</tr>
</tbody>
</table>
Why learn from human feedback?

- Provides a mechanism for gathering signals about correctness that are difficult to describe via data or cost functions, e.g., what does it mean to be funny?
- Provides signals that are best defined by stakeholders, e.g., helpfulness, fairness, safety training, alignment.
- Useful when evaluation is easier than modeling ideal behavior.
- Sometimes we don’t really care about human preferences per-se, we care about fixing model mistakes.
We have not figured out how to do it quite right (or we need new approaches)

- Reflects some human biases e.g., length, authoritative tone, ...
- Human preferences can be unreliable e.g., “reward hacking in RL"

Microsoft’s Bing Chatbot Offers Some Puzzling and Inaccurate Responses
The new, A.I.-powered system was released to a small audience a week ago. Microsoft says it is working out its issues.

Google shares drop $100 billion after its new AI chatbot makes a mistake
February 9, 2023 • 10:15 AM ET
By Emily Olson
Potential ethical issues

• Labeling often depends on low-cost human labor
• The line between economic opportunity and employment is unclear
• May cause psychological issues for some workers
Santurkar, et. al., "Whose Opinions Do Language Models Reflect?"
Preferences used to personalize therapy

**Absolute Feedback:** “That felt good, 4/5 rating.”

**Challenge:** humans are not consistent in providing absolute feedback.

Slides adapted from Yisong Yue
Preference feedback + Dueling bandits

Multi-dueling Bandits with Dependent Arms, Sui, Zhuang, Burdick & Yue, UAI 2017
Correlational Dueling Bandits with Application to Clinical Treatment in Large Decision Spaces, Sui, Yue & Burdick, IJCAI 2017
Preference-Based Learning for Exoskeleton Gait Optimization, Tucker, Novoseller, et al., ICRA 2020
Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits, Tucker et al., IROS 2020
ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes, Li, Tucker, et al., ICRA 2021

Slides adapted from Yisong Yue
Metric Elicitation

Determine the fairness and performance metric by interacting with individual stakeholders

See Hiranandani et. al., “Fair Performance Metric Elicitation”

Metric elicitation from stakeholder groups

See Robertson et. al., “Probabilistic Performance Metric Elicitation”

Empirical evaluation

See Hiranandani et. al., “Metric Elicitation; Moving from Theory to Practice”
Why elicit metric preferences?

Useful when tradeoffs are inherently stakeholder dependent i.e., human preferences are the best approach to measuring what tradeoffs matter.

Important for socio-technical tradeoffs e.g., fairness vs performance, privacy vs fairness, ...
Cooperative Inverse Decision Theory (CIDT)

Imitator (R) seeks to learn decision rule matching Demonstrator (H) preferences
Can be formalized as an assistance game (Hadfield-Menell et al., 2016)

Challenge: Highlights description-experience gap in measuring preferences

Robertson et. al. “Cooperative inverse decision theory for uncertain preferences” 2023
Recommendation systems

User item preferences to recommend new items

Both passive (offline data) and active querying (contextual bandits)

Often ratings, ranking or thumbs up/down feedback
Reinforcement Learning (RL) from Human Preferences


Flying helicopters using imitation learning and inverse reinforcement learning (IRL)

Learning for Control from Multiple Demonstrations, Adam Coates, Pieter Abbeel, and Andrew Y. Ng. ICML, 2008.
Batch active preference learning for RL

E Bıyık, D Sadigh, "Batch Active Preference-Based Learning of Reward Functions", 2nd Conference on Robot Learning (CoRL), Zurich, Switzerland, Oct. 2018.
Reward hacking in inverse RL
Design of tools for eliciting feedback from humans often has to tradeoff several factors

- **Cognitive load/effort:** Human friendly vs model-friendly feedback
- **Truthfulness:** What if there is no “correct” answer?
- **Accuracy:** What of human mistakes? What is the role of expertise?
Recurring assumptions and discussion

- Assuming human rationality (~existence of a deterministic reward function).
- Human preference often expressed as discrete choice, models often have strong parametric assumptions.
- Limited work on the role of human biases, do they matter?
- Limited work on aggregation in learning applications (lots of work in mechanism design).
- RL and active learning emphasize careful querying, while language applications are have less focus on active querying. Does it matter?
Course Goals

• Topics course covering (some) foundations and applications of learning from human preferences. Somewhat focus on breadth/coverage vs. depth

• **Foundations:** Judgement, decision making and choice, biases (psychology, marketing), discrete choice theory, mechanism design, choice aggregation (micro-economics), human-computer interaction, ethics

• **Machine learning and statistics:** Modeling, active learning, bandits

• **Applications:** recommender systems, language models, reinforcement learning, AI alignment

• **Note:** lecture schedule is tentative and topics/speakers may change
Guest Lectures

Dorsa Sadigh
Robotics

Vasilis Syrgkanis
Mechanism Design

Jason Hartline
Mechanism Design

Diyi Yang
NLP, Ethics

Merrie Morris
HCI, Ethics

Nathan Lambert
NLP, RLHF

Noah Goodman
Psychology

Jonathan Levav
Marketing

S. Wheeler
Marketing

Pat Langley
Human Computing
“Topics course” here means

- **In-class presentation**: Reading and presenting papers
- **Scribe** for lectures
- **Active feedback** and peer grading
- **Course project** i.e., research project that is relevant to this topic

- Course staff will help **curate and evaluate** this work.
Prerequisites

CS 221 (AI) or CS 229 (ML) or equivalent

You are expected to …

• Be proficient in Python (most project will include a programming component)

• Be comfortable with machine learning concepts, e.g., train/dev test set, model fitting, function class, loss functions.

• Writing assignments will likely require latex
Grading

• **Project (50%)**:  
  • Proposal (5%)  
  • Final manuscript (25%)  
  • Poster (10%)  
  • Blog post (10%)

• **Presentation + Class notes (40%)**:  
  • In-class presentations (20%)  
  • Scribe notes (20%)

• **Peer Grading (10%)**:  
  • Feedback for presentations (5%)  
  • Feedback for projects (5%)

• **Class participation (Extra credit up to 5%)**  
  • Ed Question Answering  
  • Q/A participation in class

• **No homework, no exams**
Project

- Proposal (5%)
- Final manuscript (25%)
- Poster (10%)
- Blog post (10%)

**Project scope:** any topic related to what is covered in class, i.e., make clear how in related to “ML from human preferences”. Examples:
  - Simulation platforms that capture human biases or group dynamics
  - Mechanism design for learning from group preferences
  - Applications to new problems, e.g., generative models
  - Mathematical and statistical foundations
  - Well designed user studies
In-class Presentation

• You are expected to present material for one class with a group
• We will release a signup schedule
• **Expectation**: start with provided references, then add relevant material based on our interests
Scribe

- You are expected to scribe for two lectures (group scribing)
  - One scribe for the class where you present material
  - One scribe for an invited lecture (will release a signup schedule soon)
- One week for the first draft, then iterate with the course staff
- **Expectation**: a summary of the lecture content with depth on technical details and relevant references
Peer grading

• You are expected to peer grade two components
  • In-class presentation
  • Project feedback, similar to a paper review.

• One week for feedback submission

• We will provide a rubric soon

• **Expectation**: feedback that helps your peers improve and helps you think critically about the material
In-Person Expectation

• The class is designed to be interactive and is not remote-friendly
• You will need to be in person for any components where you are providing feedback or scribe
• Slides should be available soon after lecture for review
Computing credits

• Anticipating credits on Google cloud.
• May be able to include credits for foundation models (pending)
Late Assignments

• Each student will have a total of 4 free late (calendar) days. Final project papers cannot be turned in late under any circumstances.

• Once these late days are exhausted, any work turned in late will be penalized 10% per late day.

• If a group's assignment is late n days, then each group member is charged n late days.

• Late days are never transferrable between students.
Welcome to CS 329H

• **Today:** Introduction and overview
• **Next lecture:** Discrete choice and human preference models

• **HW0:**
  • Complete the pre-course survey (see Ed)
  • Signup for scribes
  • Signup for paper presentations / lectures (soon!)
  • Start thinking of project ideas (we are also soliciting ideas, should be available Oct 6)