Embedded Ethics Module
Recommender System
Objectives

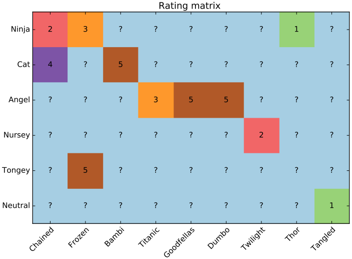
Today

- **Topic:** objective functions for recommender systems
- Two parts
    - **Part 1:** technical challenges in moving beyond regression and classification
    - **Part 2:** ethical challenges, and philosophical tools for reasoning about them

Recap and Motivation

Recap: Netflix Challenge

- We can view collaborative filtering as a matrix completion problem.



Rating matrix

	Chained	Frozen	Bambi	Titanic	Goodfellas	Dumbo	Twilight	Thor	Tangled
Ninja	2	3	?	?	?	?	?	1	?
Cat	4	?	5	?	?	?	?	?	?
Angel	?	?	?	3	5	5	?	?	?
Nursey	?	?	?	?	?	?	2	?	?
Tongey	?	5	?	?	?	?	?	?	?
Neutral	?	?	?	?	?	?	?	?	1

- In addition to the learning algorithm, it is important to consider the data and the objective function.

Recommender Systems

- Other kinds of recommendation systems include search engines and social media feeds.
- What are some difficulties you'd run into if you tried to use a Netflix-style algorithm to organize a user's social media feed?

Recommender Systems

- If you were designing an ML algorithm to organize a user's social media feed, what other information might you use?

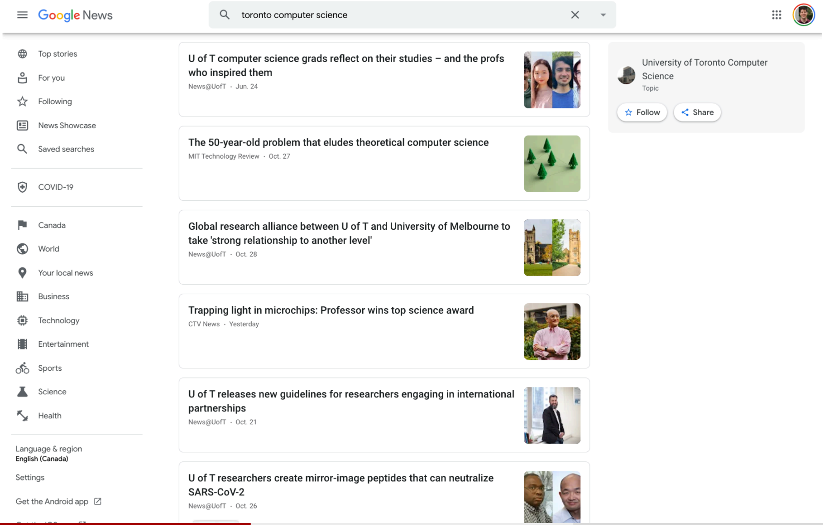
- As a supervised learning problem, what would be the inputs, and what would be the targets?

Warmup: Open up your social media feeds - shout out some of the topics in the posts that you find.

Challenge 1: Inferring User Preferences

Challenge 1: Inferring User Preferences

- Google News was an early example of training a model to predict clicks.



Google News

toronto computer science

- Top stories
- For you
- Following
- News Showcase
- Saved searches

- COVID-19

- Canada
- World
- Your local news
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Language & region
English (Canada)

Settings

Get the Android app

U of T computer science grads reflect on their studies – and the profs who inspired them
News@UofT · Jun. 24

The 50-year-old problem that eludes theoretical computer science
MIT Technology Review · Oct. 27

Global research alliance between U of T and University of Melbourne to take 'strong relationship to another level'
News@UofT · Oct. 28

Trapping light in microchips: Professor wins top science award
CTV News · Yesterday

U of T releases new guidelines for researchers engaging in international partnerships
News@UofT · Oct. 21

U of T researchers create mirror-image peptides that can neutralize SARS-CoV-2
News@UofT · Oct. 26

University of Toronto Computer Science
Topic

☆ Follow
Share

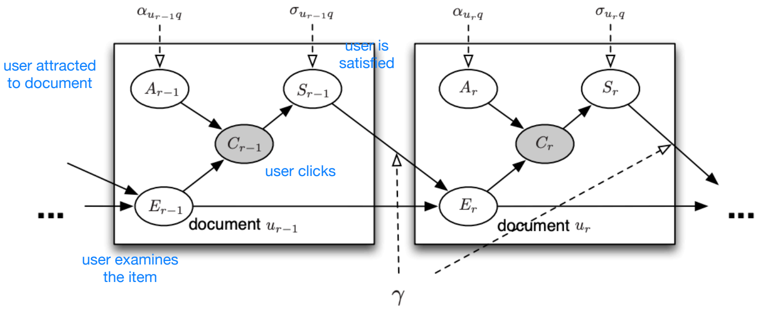
Challenge 1: Inferring User Preferences

- Why are clicks a useful signal?

- What are some problems with optimizing for clicks?

Challenge 1: Inferring User Preferences

- Here is a Bayesian network designed to model user behavior for a search engine.
  - We covered Bayes nets briefly when we discussed naïve Bayes.
- Nodes represent random variables, and edges represent direct influences. Shaded = observed.
- Want to infer user satisfaction (S).



Chuklin et al., “Click models for web search”

Challenge 1: Inferring User Preferences

- User preferences aren't just a matter of reactions to individual items, but also of the user's overall experience.
- Many web services optimize for a criterion called engagement.
  - User's frequency, intensity, or depth of interaction with a product over some time period
  - Not a technical term, but a business term, instantiated in different ways by different companies
  - E.g. Gmail: percentage of active users who visited the site on 5 or more days during the past week Rodden et al., "Measuring the user experience on a large scale"
  - E.g. Facebook: time spent on site, meaningful social interactions `https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/`
- This is not directly optimized by an ML algorithm (as far as I know), but is used to evaluate changes to the system.
  - Sort of analogous to how logistic regression minimizes cross-entropy loss but you might tune hyperparameters based on accuracy.

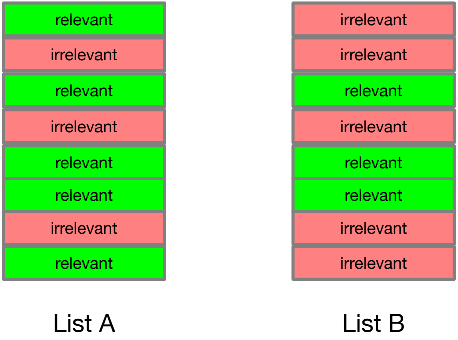
Challenge 1: Inferring User Preferences

- The choice of what to optimize for can have ethical implications.
- The recently published Facebook Papers reveal a lot about unintended consequences of algorithm design
  - My aim isn't to pick on Facebook here. They found these harms and worked to fix them!
- Early years: optimizing for likes and clicks ⇒ clickbait
- Optimizing for time spent reading/watching ⇒ favored professional over organic content
- 2017: service changed to reward comments & emojis ⇒ most successful political posts were the polarizing ones
  - Some political parties consciously shifted their messaging to be much more negative
  - Facebook eventually rolled back this change for health and politics
- `https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215`

Challenge 2: Evaluating Structured Outputs

Challenge 2: Evaluating Structured Outputs

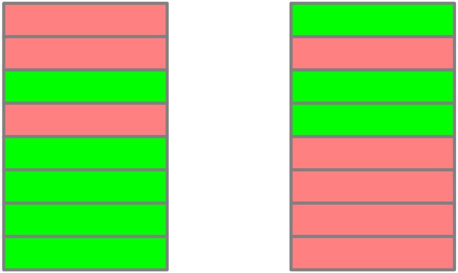
- Most of this class has focused on classification, where there is a natural metric to use (accuracy).
- In this case, we'd like to produce a feed (an ordered list of items). Problems where we want to predict a structured object are known as structured prediction.
- For now, assume that all items are either relevant or irrelevant.
- Which of the following lists is preferable?



List A	List B
relevant	irrelevant
irrelevant	irrelevant
relevant	relevant
irrelevant	irrelevant
relevant	relevant
relevant	relevant
irrelevant	irrelevant
relevant	irrelevant

Challenge 2: Evaluating Structured Outputs

- One basic measure is precision: the fraction of items which are relevant.
- Which of the following lists is preferable?



List A

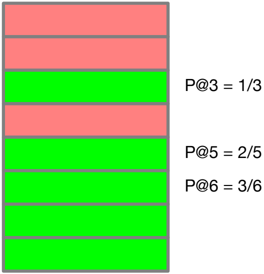
List B

Challenge 2: Evaluating Structured Outputs

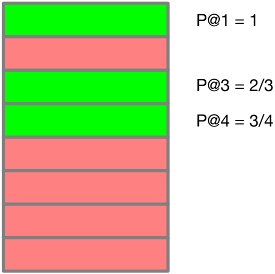
- Precision@K: Precision for the list up to the Kth item.
- Average Precision (AP): average of Precision@K, where K is taken as the indices of the first N relevant items.
  - Moving a relevant item from position 2 to position 1 is worth more points than moving it from position 8 to position 7.
- Mean Average Precision (MAP): mean of the AP over multiple queries.
- Note: in different application areas, there are different (but related) definitions of AP/MAP.

Challenge 2: Evaluating Structured Outputs

An example of calculating AP with $N = 3$.



$$\text{MAP} = \frac{1}{3}\left(\frac{1}{3} + \frac{2}{5} + \frac{3}{6}\right) \approx 0.41$$



$$\text{MAP} = \frac{1}{3}\left(1 + \frac{2}{3} + \frac{3}{4}\right) \approx 0.81$$

Challenge 2: Evaluating Structured Outputs

What other factors might you consider in evaluating a list of recommendations?

Towards Ethics

- We've been discussing challenges that arise when defining optimization objectives beyond the basic classification and regression settings.
- So far, we've focused on challenges of building a useful and engaging system.
- But what we choose to optimize for can have unintended consequences. The rest of the lecture focuses on thinking about optimization objectives from an ethical standpoint.

Machine learning and the rise of large models

- Our lecture today focused on recommender systems.
- In the next 5-10 years, we will see a dramatic rise in the use of large language models and generative models of realistic images,
- These models are good enough that they produce outputs that are indistinguishable from human generated content.

Input into GPT-3

The following text was input into the model:
Summarize this for a second-grade student: Decision tree learning is a method commonly used in data mining. . . . This process of top-down induction of decision trees (TDIDT). This is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.
Given the above, the model was asked to predict the next few sentences. The provided text was half a page in length and taken verbatim from Wikipedia and was half a page in length

Output from GPT-3

Decision tree learning is a method used to create a model that predicts the value of a target variable based on several input variables. A decision tree is a simple representation for classifying examples. A tree is built by splitting the source set, constituting the root node of the tree, into subsets – which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions.

DALL-E

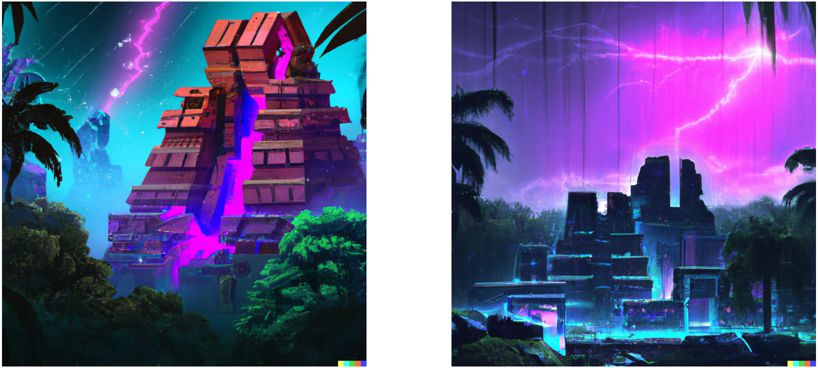


Figure: Synthetic image generated by caption “An synthwave style ancient city in a lush rainforest with a backdrop of moonlight and lightning”

Ethical concerns with large generative models

- Trained on open source data from the web – often (via maximum likelihood estimation) on content that is sexist, racist and misogynist.
- The resulting predictive content can then be biased.
- Legal implication of training models on data from the internet.

Acknowledgements

This module was created as part of an Embedded Ethics Education Initiative (E3I), a joint project between the Department of Computer Science[1] and the Schwartz Reisman Institute for Technology and Society[2], University of Toronto.

Licensed by the University of Toronto Embedded Ethics Education Initiative and Roger Grosse under the Attribution-NonCommercial-ShareAlike 4.0 International license.
To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/.

Instructional Team:
Roger Grosse, Steven Coyne, Emma McClure

Faculty Advisors:
Diane Horton[1], David Liu[1], and Sheila McIlraith[1,2]

Department of Computer Science
Schwartz Reisman Institute for Technology and Society
University of Toronto

Computer Science
UNIVERSITY OF TORONTO

UNIVERSITY OF
TORONTO

SCHWARTZ REISMAN INSTITUTE
FOR TECHNOLOGY AND SOCIETY