Artificial Intelligence and Machine Learning

CS@CU

What is Artificial Intelligence?

- You know what it is—computer programs that "think" or otherwise act "intelligent"
  - The Turing test?
- What is "machine learning" (ML)?
  - It's simply one technique for AI—throw a lot of data at a program and let it figure things out
- What are "neural networks"?
  - A currently popular technique for ML

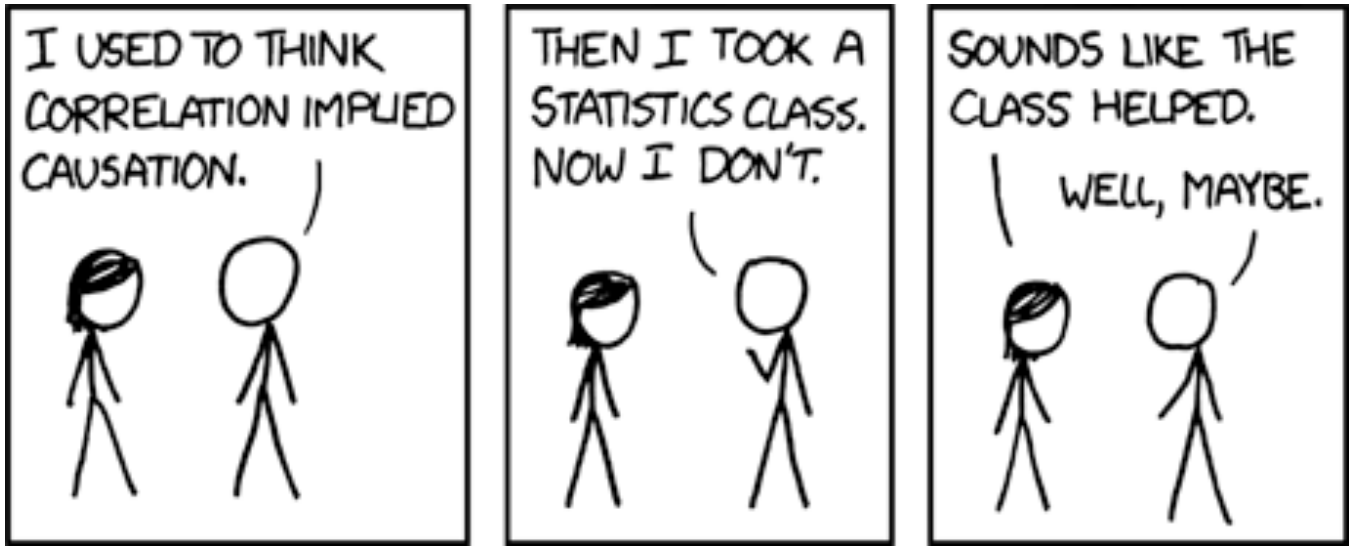
How Does ML Work?

- Lots of complicated math
- *Not* the way human brains with human neurons work
- To us, it doesn't matter—we'll treat it as an opaque box with certain properties

How ML Works

- You feed the program a lot of *training data*
- From this training data, the ML algorithm builds a model of the input
- New inputs are matched against the model
  - Examples: Google Translate, Amazon and Netflix's recommendation engines, speech and image recognition
- However—machine learning algorithms find *correlations*, not *causation*
  - It's not always clear why ML makes certain connections

Correlation versus Causation



https://xkcd.com/552/

Training Data

- Training data must represent the desired actual input space
- Ideally, the training records should be statistically *independent*
- If you get the training data wrong, the output will be biased
- To understand or evaluate the behavior of an ML system, you need the code *and* the data it was trained on
  - "Algorithm transparency" alone won't do it

Learning Styles

Supervised

- A human *labels* the training data according to some criteria, e.g., spam or not spam
- The algorithm then “learns” what characteristics make items more like spam or more like non-spam

Unsupervised

- Finds what items cluster together
- Useful for large datasets, where there is no ground truth, or where labels don’t matter
- What counts is *similarity*

Supervised: Image Recognition

- Feed it lots of pictures of different things
- Label each one: a dog, a plane, a mountain, etc.
- Now feed it a new picture—it will find the closest match and output the label

Unsupervised Learning

- Feed in lots of data *without* ground truth
- The algorithms find clusters of similar items; they can also find outliers—items that don't cluster with others
- They can also find probabilistic dependencies—if a certain pattern of one set of variables is associated with the values of another set, a prediction can be made about new items' values for those variables

Uses of Machine Learning

Recommendation Engines

- To recommend things to you, Amazon, Netflix, YouTube, etc., do not need to know what you buy or watch
- Rather, they just need to know that people who liked *X* also tended to like *Y* and *Z*.
- This is a classic example of unsupervised learning

An Amazon Recommendation

Customers Who Bought This Item Also Bought

Page 1 of 20

Samsung Galaxy S4 Charger 2.1Amp 2-Port Adapter for Travel Home Wall with 3 feet Micro…

★★★★★ 22

$9.99 Prime

iphone 6s Full Screen Protector, PLESON® iphone 6 6s Edge to Edge Full Screen Cover [3D…

★★★★★ 41

$14.99 Prime

Soniworks Compatible (2-Pack) Replacement Facial Cleansing Brush Heads, designed for…

★★★★★ 106

$11.95 Prime

Hard Rhino Creatine Monohydrate Micronized 200 Mesh Powder, 125 Grams

★★★★★ 241

$10.01 Prime

Miss Jordan Salt and Pepper Grinder Set. Elegant Stainless Steel Salt and Pepper Grinder Set…

★★★★★ 56

$23.00 Prime

Finding Terrorists?

- There are very, very few terrorists
- Where are you going to find enough training data?
- Almost certainly, any features the real terrorists have in common will be matched by very many other innocent people
- The algorithms can’t distinguish them

Finding Terrorists

- There are very, very few terrorists
- Where are you going to find enough training data?
- Almost certainly, any features the real terrorists have in common will be matched by very many other innocent people
- The algorithms can't distinguish them
- When humans do this, we call it profiling

A Recent Facebook Patent

United States Patent **10,607,154**
Sullivan , et al. **March 31, 2020**

Socioeconomic group classification based on user features

Abstract

An online system uses classifiers to predict the socioeconomic group of users of the online system. The classifiers use models that are trained using features based on global information about a population of users such as demographic information, device ownership, internet usage, household data, and socioeconomic status. The global information can be aggregated from market research questionnaires and provided to the online system. The classifiers input information about a user and output a probability that the user belongs to a given socioeconomic group. The input information is based on a user profile on the online system associated with the user as well as actions performed by the user on the online system. Thus, the online system can predict the user's socioeconomic group without using the user's income information. The online system can generate content for presentation to the user based on the predicted socioeconomic group.

ML Doesn’t Always Work the Way We Want it To…

Some Examples

- Biased training data
- Microsoft Tay
- Recidivism risk
- Targeted advertising
- More…

Watch Out for Biased Training Data!

Training data that doesn't represent actual data

- Google Photos misidentified two African-American men as gorillas
  - Jacky Alciné—a programmer and one of the people who was misidentified, "I understand HOW this happens; the problem is moreso on the WHY."
  - Likely cause: not enough dark-skinned faces in the training dataset
- Cultural biases by the trainers
  - Mechanical Turk workers are often used for labeling
- False positives and false negatives

Bias In, Bias Out

- Suppose you want an ML system to evaluate job applications
- You train it with data on your current employees
- The ML system will find applicants who “resemble” the current work force
- *If your current workforce is predominantly white males, the ML system will select white male applicants and perpetuate bias*

Microsoft Tay

- A Twitter “chatbot”
- Tay “talked” with people on Twitter
- What people tweeted to it became its training data
- It started sounding like a misogynist Nazi…

What Happened?

- People from 4Chan and 8Chan decided to troll it.
- With ML, vile Nazi garbage in, vile Nazi garbage out
- Microsoft didn't appreciate just what people would try.
- "Sinders is critical of Microsoft and Tay, writing that 'designers and engineers have to start thinking about codes of conduct and how accidentally abusive an AI can be.'" (*Ars Technica*)

Recidivism

- Several companies market “risk assessment tools” to law enforcement and the judiciary
- Do they work? Do they exhibit impermissible bias?
- A ProPublica study says that one popular one doesn’t work and does show racial bias: blacks are more likely to be seen as likely reoffenders—but the predictions aren’t very accurate anyway

What Happened?

- Inadequate evaluation of accuracy
- Using the program in ways not intended by the developers
- Proxy variables for race
- Using inappropriate variables, e.g., “arrests” rather than “crimes committed”

Hypertargeted Advertising

- It’s normal practice to target ads to the “right” audience
- ML permits very precise targeting—others can’t even see the ads
- Used politically—some research says that YouTube’s recommendation algorithms radicalize people
- Target managed to identify a pregnant 16-year-old—her family didn’t even know

Target

- People habitually buy from the same stores
- They tend to switch only at certain times, e.g., when a baby is born
- Target analyzed sales data to find leading indicators of pregnancy
- They then sent coupons to women who showed those indicators
- People found that creepy—so Target buried the coupons among other, untargeted stuff that they didn't really care if you bought

ChatGPT and Other Large Language Models

- *Not* intelligent!
- As with all ML systems, based on a large amount of training data
- They generate *probabilistic* sentences, based on the chance of the next word occurring in its training data
- Words are represented as *vector*—long lists of numbers (probably tens of thousands of numbers) for each word
- Words that are somehow related have vectors that are "near" each other—"dog" will be closer to "cat" than it is to "computer"
  - Arithmetic on vectors: "biggest" - "big" + "small" can produce a vector that is close to "smallest"
  - Can echo biases in training data: "doctor" - "man" + "woman" → "nurse"
- Detailed understanding of the output is often beyond human grasp

Privacy and ML

- ML algorithms can act on you without knowing who you are
- ML algorithms can link disparate datasets to identify you, even without common database keys
- ML algorithms can predict thing about you

And: these algorithms are often wrong—is that better or worse?

Questions?

Red-eyed vireo, Central Park, August 31, 2022