ECON 340
Economic Research Methods

Div Bhagia

Lecture 25
Big Data & Machine Learning

Predictive vs Causal Inference

- Econometrics: Causal Inference

$$Y = \beta_0 + \beta_1 X + u$$

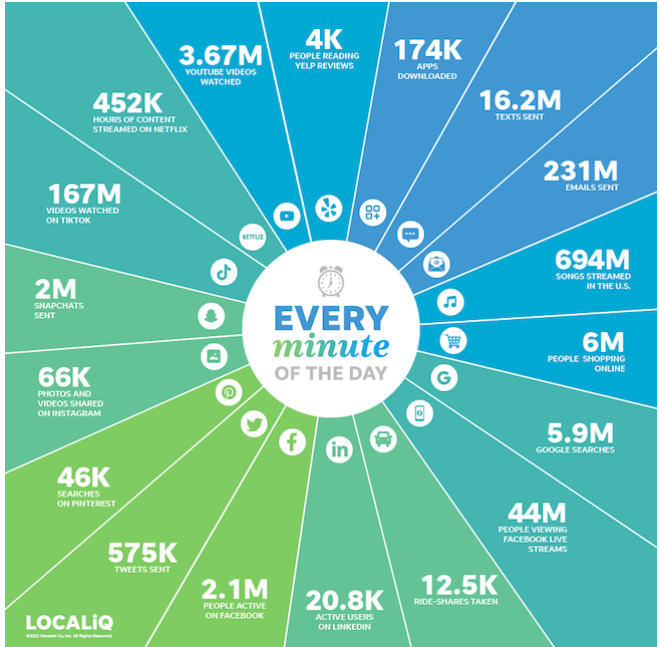
  β_1 is the causal impact of X on Y if $E(u|X) = 0$.

- Machine Learning (ML): Predictive Analytics
  - Want $\hat{Y}$ to be as close as possible to Y
  - Better with "big data"

- Distinction between the ML vs. "traditional" stats: prediction vs. unbiased estimation

Big Data and Machine Learning

- The term “big data” refers to data that is so large, fast, or complex that it’s difficult or impossible to process using traditional methods
  - → Not just lots of observations but also lots of variables
- Machine learning: set of algorithms for big data analytics
- Organizations collect data from a variety of sources
  - online purchases, scanner data, Uber analytics, smart sensors, aggregation of tweets on Twitter, Google searches, Yelp, Zillow, etc.


3.67M
YOUTUBE VIDEOS WATCHED
4K
PEOPLE READING YELP REVIEWS
174K
APPS DOWNLOADED
16.2M
TEXTS SENT
231M
EMAILS SENT
694M
SONGS STREAMED IN THE U.S.
6M
PEOPLE SHOPPING ONLINE
5.9M
GOOGLE SEARCHES
44M
PEOPLE VIEWING FACEBOOK LIVE STREAMS
12.5K
RIDE-SHARES TAKEN
20.8K
ACTIVE USERS ON LINKEDIN
2.1M
PEOPLE ACTIVE ON FACEBOOK
575K
TWEETS SENT
46K
SEARCHES ON PINTEREST
66K
PHOTOS AND VIDEOS SHARED ON INSTAGRAM
2M
SNAPCHATS SENT
167M
VIDEOS WATCHED ON TIKTOK
452K
HOURS OF CONTENT STREAMED ON NETFLIX
EVERY
minute
OF THE DAY
LOCALiQ

Machine Learning vs. Econometrics

- If the goal is prediction, ML methods beat econometrics (lasso, regression trees, random forests, etc.)
- However, more data cannot solve a causal inference problem. But that's ok!
- Lots of applications when prediction is useful.
  - Macro or financial forecasting
  - Predicting valuations for new products
  - Optimizing marketing campaigns
  - Others?

Semantics

Some language differences between statistics and ML:

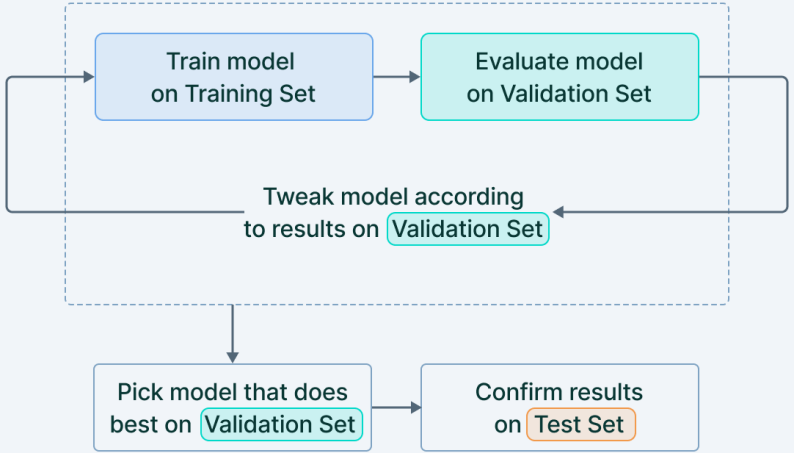
- instance = data point
- features = variables
- learning = fitting models to data
  - supervised learning: fit a function to a target (regression)
  - unsupervised learning: no target (density estimation), e.g., classification

Machine Learning

But what about statistical inference?

- How does the researcher know they are fitting true relationships to data and not those that have arisen spuriously from chance?
- Traditional null hypothesis significance testing is of limited use given millions of observations
- Solution: approximate out-of-sample fit using a training–validation-testing split of the underlying data

Training data/validation/test



Example: Spam Detection

- Spam Detection

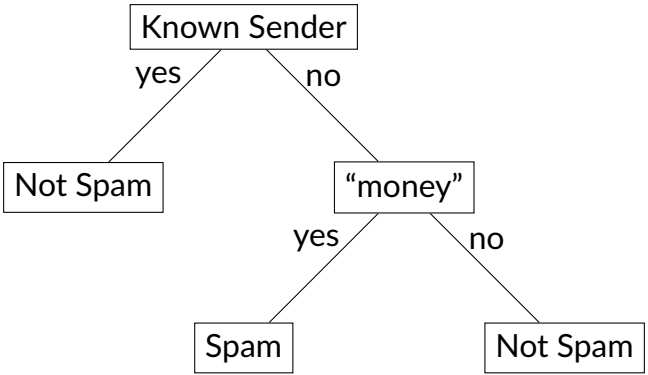
  Represent each message by features (e.g., keywords, spelling, etc.)

"money"	"Mr."	bad-spelling	known-sender	spam?
Y	Y	Y	N	Y
N	N	N	N	N
Y	N	Y	Y	N
N	Y	Y	N	Y
Y	N	N	N	Y
N	N	N	Y	N

Come up with rules: predict spam if...

An ML Algorithm: Decision Trees

With all the data, no need to fit a linear model, can be more flexible



Not necessary that all variables are relevant. ML pays attention to “feature selection.”

Machine Learning: Other Applications

- Face detection and recognition
- Weather prediction
- Diagnosing diseases
- Predict whether a user will click on an add
- Predict stock prices

Artificial Intelligence and Machine Learning

- Artificial intelligence: the general ability of computers to emulate human thought and perform tasks e.g. computer games, smart speakers, etc.
- How can a computer play a game?

Artificial Intelligence and Machine Learning

- Artificial intelligence: the general ability of computers to emulate human thought and perform tasks e.g. computer games, smart speakers, etc.
- How can a computer play a game?
  Moves are determined by an algorithm, which is designed to mimic human thought processes and decision-making.
- How to come up with this algorithm?
  - Traditional AI: pre-programmed directly using human judgement
  - Machine Learning: learn from data on past games

Tic Tac Toe

- You can play Tic Tac Toe (and much more complex games) with a computer
- Computer's intelligence is pre-determined by rules like
  - "if the opponent has two in a row, block them," or
  - "take the center square if it's available."
- These pre-programmed algorithms are "*Artificial Intelligence*"
- One way to come up with these rules is to just pre-program them directly using human judgment

Tic Tac Toe

- Another way to teach the computer to play a game is to use Machine Learning, in which the computer learns from data on past games
- In this approach, the ML model identifies patterns and strategies from the game data.
- For example, it might notice that taking the center square often leads to a win, or that blocking an opponent's potential line of three is a good defensive strategy.
- ML enables us to automate teaching computers to build AI

Natural Language Processing

- NLP: subfield of artificial intelligence (AI) and computational linguistics
- NLP is concerned with developing algorithms and models that allow computers to understand our language
- In NLP, ML is used to develop models that can learn from large amounts of text data and identify patterns and relationships in that data
- Uses: ChatGPT, Chatbots, translation, sentiment analysis, summarizing text

From the Horse's Mouth

ChatGPT on ChatGPT:

> *The model is trained on a large dataset of text, and during training, it learns to predict the probability of each word in a given context based on the words that came before it.*
>
> *When answering questions, ChatGPT generates a probability distribution over all possible responses and then selects the most likely response based on that distribution.*

What's next

- Final research paper due today
- Review class this Thursday
- Material for the final exam uploaded on the Course Website
- Final exam from 1–2.50 pm on Thursday
- Please fill the SOQs :)