Some principles of machine learning

Machine Learning in Molecular Science

Prof. Michael Shirts

July 22nd, 2024

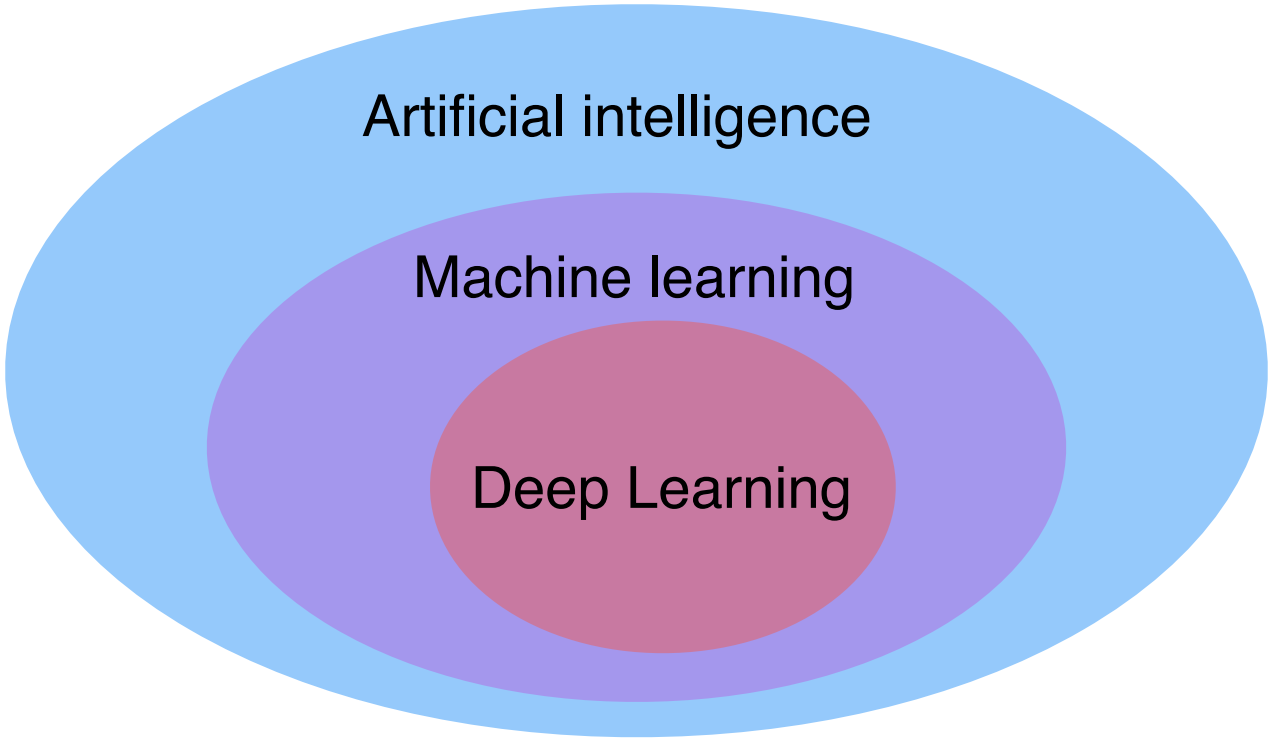
Welcome, everyone!

- There's a large diversity of backgrounds here
- Material might be too fast/too slow
- I'll be patient with you, you be patient with me
- Find a partner to work together on some of the exercises

Hello!

- Michael Shirts
- Professor at University of Colorado Boulder
- Department Chemical & Biological Engineering
- Specialty is statistical mechanics molecular dynamics, and soft materials
- Running, reading sci-fi
- Kids and cats!

Some different classifications



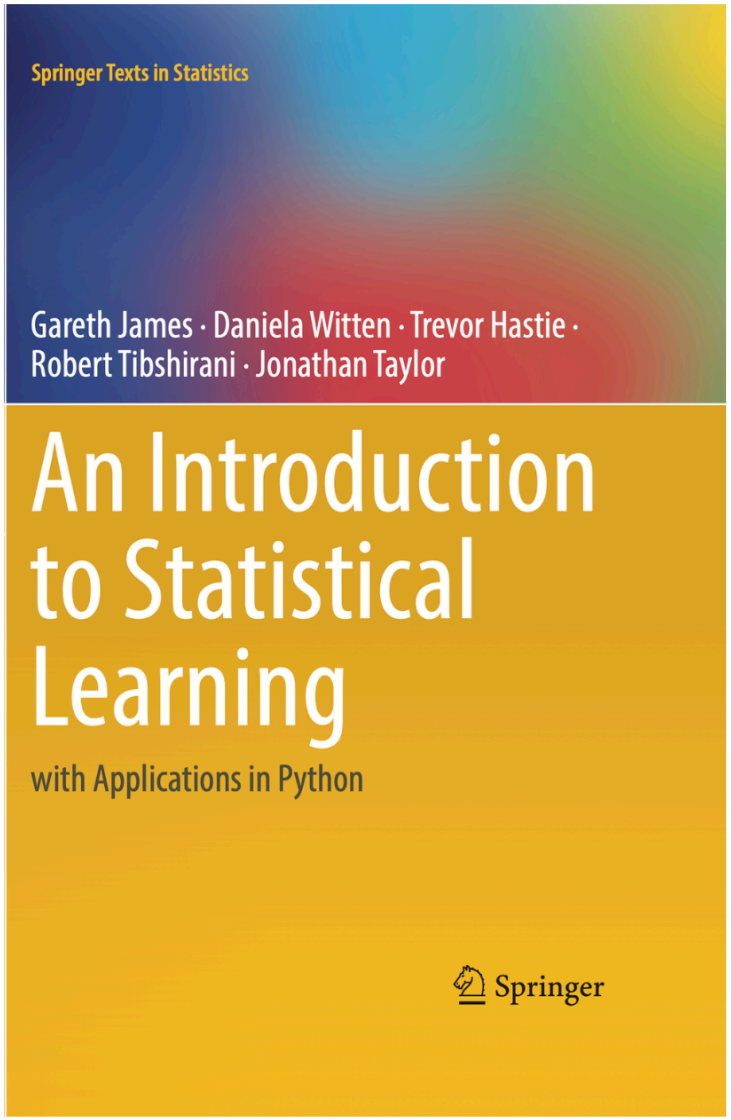
Deep Learning = **Generally**, a neural net with many layers
Does not imply the learning itself is "deep"
One of the most incredible success stories in history of tech marketing

Some abbreviations

- AI = artificial intelligence
- ML = machine learning
- DL = deep learning
- NN = neural networks
- GNNs = graph neural networks
- NNPs = neural network potentials

Resources

- Most of these concepts are taken from:
- *An Introduction to Statistical Learning* (ISL)
- I think it's best introduction to the **fundamental concepts** used in machine learning in a non-hyped way.
- Free to download! Python examples available!



What are we studying this week

Machine learning

vs.

Statistical learning

"Method of data analysis that automates the building of models which are derived from algorithmic pattern-recognition applied to existing data"

What is called "machine learning" today is really just taking the statistical tools that have existed for decades, and scaling them to data sets 10^3 to 10^6 times larger!

Supervised vs. Unsupervised Learning

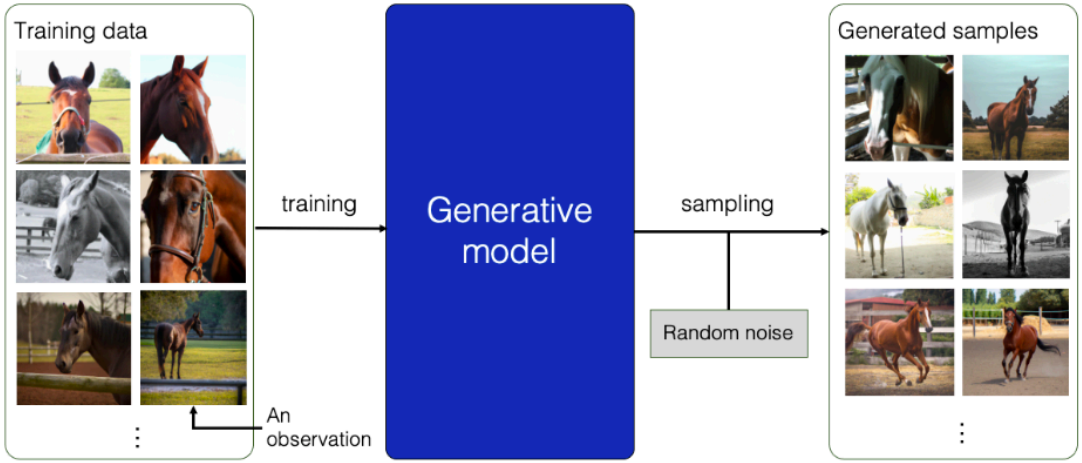
- Supervised learning:
  - Building a statistical model for predicting an output based on inputs.
  - We are given outputs for some of inputs. What are the outputs for other inputs?
  - Supervised - there's a supervisor who can tell you something about the data.
- Unsupervised learning:
  - Here is some data. Are there relationships or categorizations within the data that tell us something about the data?
  - There are no outputs

Prediction vs. Inference

- Prediction:
  - I give you some input and outputs to train.
  - If I then give you new inputs, can you accurately predict the outputs?
- Inference:
  - Asks questions like:
    - Which predictors are most associated with the response?
    - What is the relationship between the response and each predictor?
- Tends to be deemphasized in machine learning/ deep learning, since the predictors become very complicated and non-intuitive.
  - Area of intense interest right now: "Interpretable ML"

Generative Modeling

- Generate NEW data that "looks like" sample data.
  - Defining "looks like" is **the** most difficult thing here . . .



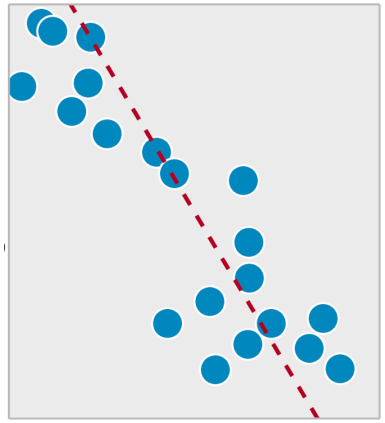
- Dall-E 2
  - CLIP and diffusion models
- ChatGPT 4
  - Large Language Model using self-attention
- Covered more at end of the week

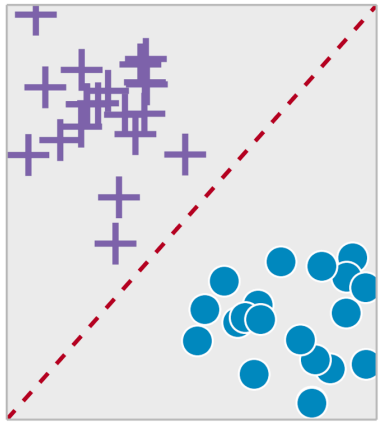
Reinforcement Learning

- Learn a process (a "policy") that goes from initial states to final states
  - Rewards and punishments based on the expected outcomes of the process
  - The policy is updated to maximize the reward
- Unlike supervised learning, we don't have an answer for each input
- Example:
  - a bot learning to play chess or Go
- Won't cover this week

Regression vs. Classification

- Both supervised learning tasks
- Regression
  - Learning a predictive model that gives continuous output
  - Linear regression:
    - If we assume y(x) = ax+b, which a and b are best implied by the data?
    - Equivalently, how would we predict y for new x?
- Classification
  - Learning a model that make a discrete prediction:
    - Is this image a cat or not a cat?
    - Which of n groups does this fall into?





Some more terminology

- "Features": the inputs you want to use
  - "Featurize": how to choose inputs for data
  - Very important to molecular learning!
- "Expressive features": Features that have some correlation to the outcome
- "Observations" or "labels": outputs
- "Score": How good is my model?
  - Also called "loss function"

Parametric vs. Nonparametric Regression

- Parametric regression
  - Your model has fixed parameters
    - Example: Linear or multilinear regression: parameters are the linear coefficients

- Nonparametric regression:
  - No fixed functional form
    - Example: Kernel density approximation, where you put a Gaussian down wherever you take a sample
  - Neural nets (kind of)
  - Usually, more data is needed for nonparametric
  - Less prone to bias
  - Nonparametric regression DOES have hyperparameters that you chose that at the beginning

We need to not overfit i.e.: fool ourselves

Richard Feynman, 1974 Caltech Commencement Address

The first principle is that you must not fool yourself — and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.

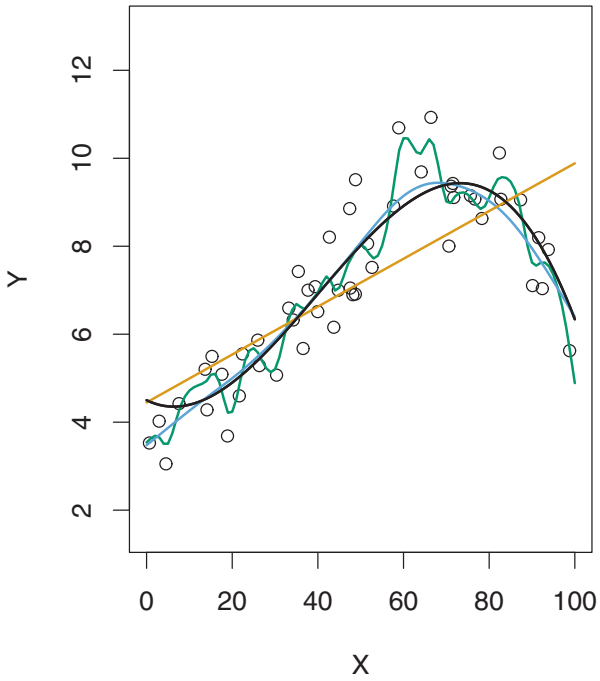
This long history of learning how not to fool ourselves — of having utter scientific integrity — is, I'm sorry to say, something that we haven't specifically included in any particular course that I know of. We just hope you've caught on by osmosis.

One of the most important underlying themes in learning ML will be to learn now to not fool yourself

Unfortunately, often accurate

https://xkcd.com/1838/

Flexibility vs. Robustness



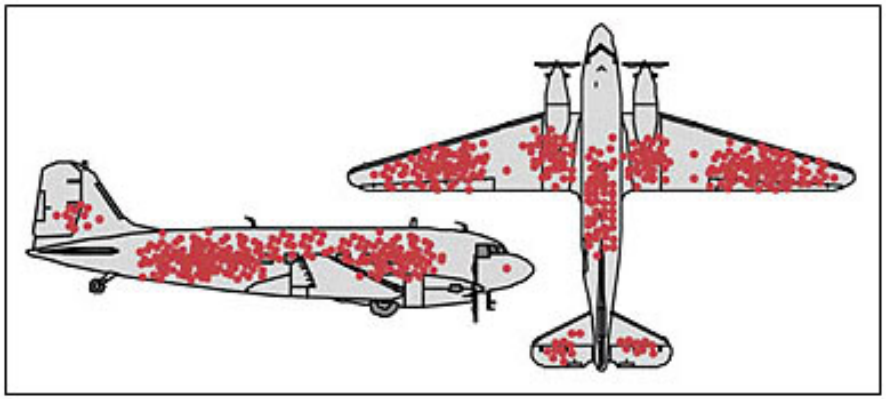
With four parameters I can fit an elephant,
and with five I can make him wiggle his trunk.
John von Neumann
(one of the fathers of modern computing)

Q: How many parameters does typical deep learning model have?

If there is some underlying correlation hidden in the data,
deep learning will find it.
If there is **no** underlying correlation hidden in the data,
deep learning will find it.

Bias due to the data set you train your model with

- Bias: you have encoded some dependence into your model from the data used to generate your model
  - Association bias: A hiring model that assumes that new hires should look like the people you have already hired.
  - Exclusion bias: Red marks are where bullets damage was found on planes. Where should you reinforce the planes?
  - Measurement bias: You train your model on images from a different brand of camera.
  - Observer bias: training is done on labels collected by biased observers



Credit: Cameron Moll

The bias/variance tradeoff

- Variance:
  - The amount the model predictions would change if we picked a different random input set
- Bias:
  - The amount the model differs from the real results because it is not a good enough model.
- Bias/Variance tradeoff
  - More complex models have **more variance** but **less bias**
    - It fits the data better, but predictions will change with training on different inputs
  - Simpler models have **less variance** but **more bias**
    - Less sensitive to changes in the inputs, but are less likely to be great models