

7750: Mathematical Foundations of Machine Learning

A first-year PhD class that builds linear algebra and probability foundations for ML research



Instructor: Ashwin Pananjady (ashwinpm@gatech.edu)
TAs: Mouhyemen Khan (mouhyemen.khan@gatech.edu) and TBA

Expectations

- Mathematical maturity/curiosity/willingness to learn about foundational material and engage with proofs
- This class will **not** be a laundry list of methods to apply to your favorite problem. We will introduce some methods with the goal of rigorously understanding some basic data analysis pipelines.
- The most important takeaway for some of you might be to recognize that these ideas can help in designing new, principled machine learning methodology, or conversely, to recognize the immense opportunity that exists to place several modern machine learning techniques on a rigorous footing.
- This is a graduate-level class: We expect that you are here to learn something new and will conduct yourself as such. No redistribution of materials on external websites, no referencing of past assignments/solutions. Strict adherence to honor code.
- We reiterate our support for your well-being, and applaud your drive in choosing to learn something new this semester

Registration

- We are at classroom capacity (116).
- Enrollment fluctuates **a lot** during the first week. If you are on the waitlist for this class at the end of Friday, you may get in.
- Class is offered every Fall semester. Consider if you **need** to take the class this semester.
- The syllabus is up. Notes from previous offerings can be found online. This offering will proceed at a faster pace.
- HW0 should help you assess if you have/can pick up the background required for this class. HW1 will be released on Wednesday.

Complementary classes

- ISyE/CSE 6740: Much more exposure to **implementation** of ML methods on real data.
- ECE6254: More focus on the statistical aspects of ML as opposed to the modeling.
- CS 7643: Applied ML, with a focus on deep learning
- ECE8803 (Sequential decision making in ML): Mathematical foundations of online ML.
- ISyE 4803 (Foundations of Modern Data Science): Undergraduate special topics class with much of this material but taught at the undergraduate level with more real datasets.

Communication

- Canvas (all HW, solutions, notes, slides)
- Course website (syllabus and description, HW0): <https://sites.gatech.edu/ashwin-pananjady/7750-mathematical-foundations-of-machine-learning-fall-22/>
- **Piazza:** Sign up with if you haven't already; link is here: piazza.com/gatech/fall2022/eceecscseisye7750! This is the primary mode of communication outside class. *Post on Piazza if you have a question.* If it is pertinent for the staff only, create a private post. Do not give away answers on Piazza, but try to give your peers hints. We will have bonus credit for Piazza engagement.
- Be polite and professional to the staff and your peers in your interactions.

Schedule

- Lecture: 3.30-4.45pm MW, Weber SST Lecture Hall 1
 - Notes will provide all the info you need if you miss a class. If you must miss a class and require assistance over and above the notes, please come to OH.
- Instructor OH (online for the first week, TBA after):
5-5.30pm M and 8-9am Th
- TA OH and problem solving session (location TBA):
5.30-7pm on F

Grading

- Homework: 50%
- Midterm 1: 25% (tentatively Sep 26, in class)
- Midterm 2: 25% (tentatively Nov 21, in class)
- Bonus credit scattered throughout (e.g., HW0, Piazza participation)
- Standard grading system: (over 90% A, 80-89% B, etc.).

HW logistics

- 6-7 HW, due fortnightly. Will use best $N-1$ out of N graded HWs.
- Released on Wed, due on Tuesday at 11.59pm ET via Canvas upload. See syllabus for exceptions and tentative schedule.
- Collaboration encouraged, but all solutions (both to math and coding problems) must be written yourself and collaborators/ other sources acknowledged on the first page.
- HW grade will be based largely on **self-grade**, due 3 days after solutions are released. TAs will grade 1 problem chosen at random and provide feedback on whether you are being lenient/harsh on yourself.

HW0

- Released on Canvas already. Piazza has screenshots of individual problems to guide questions/discussion. Ask questions here!
- **Not a graded HW.** This is bonus, to brush up on your prerequisites (there are links to external resources if you need to revise). Worth 2/50 HW points just for submitting.
- In particular, this will get you set up with Python installation, etc. Coding in HW will be on Jupyter notebooks
- Solutions will be discussed briefly in class, and you will have a chance to ask any lingering questions

COVID

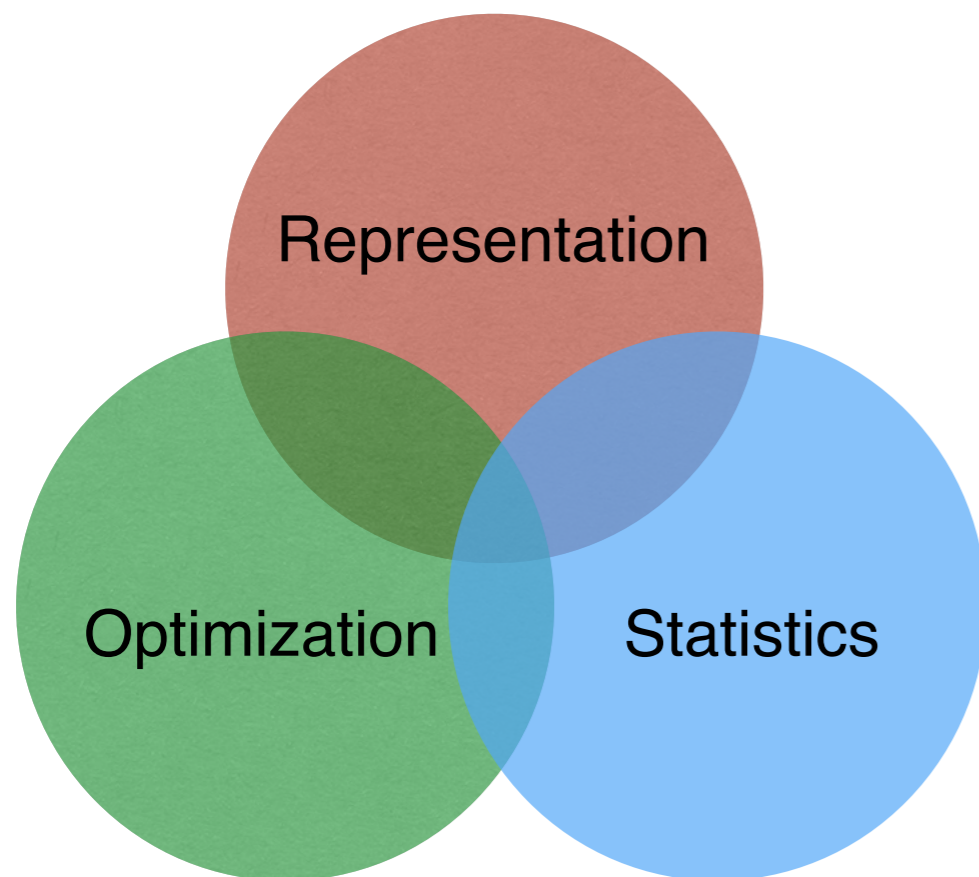
- I will be testing weekly
- Overall, I expect to see common sense exercised when dealing with covid
- My recommendation to you (I cannot mandate any of this):
 - Test if you have any symptoms or have had exposure
 - Do not come to class if you have symptoms/test positive
 - Socialize outdoors whenever possible and stay vigilant
 - Respect one another's choices when it comes to masking

Questions?

Machine learning/Data Science

Core tenet: There is some underlying phenomenon, and each datum presents an example of the phenomenon.

Model class: Articulation of plausible structure we expect to be satisfied by the phenomenon.



How do we choose model classes appropriate for the phenomenon?

Are there **computationally efficient** procedures that can fit useful models?

How much data do we need before we can make reliable inferences?

What is this class (not) about?

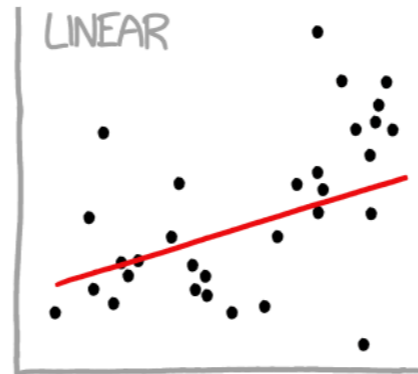
- Primary focus: **Rigorously understanding** ML foundations
- We will:
 - Use applications to motivate a few (progressively more complex) mathematical models for machine learning
 - Derive **rigorous** methodology with proof-based arguments and use simulations to guide our understanding
 - (Sometimes) Use real data to run basic experiments that solidify conceptual foundations
- We will not:
 - Provide a laundry list of methods and ask you to run them on many problems
 - Cover the mathematical foundations of neural networks in any detail, except for a few more mature facets

Why are your prereqs useful?

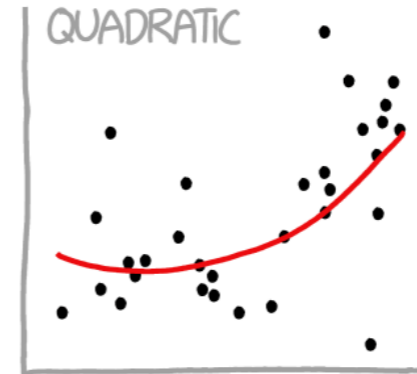
- Theory for modeling natural phenomena and making inferences from these models (Linear algebra/probability)
- What assumptions does this rely on? (Probability/stats)
- Once we have a family of models and data, how do we **fit** a specific model to data? (Linear algebra/calculus)
- Posing and understanding optimization problems and algorithms for specific ML tasks (Linear algebra/calculus)
- We will do all of this **rigorously**. We expect that you know/will learn how to write rigorous proofs.

Representation: How to model phenomenon?

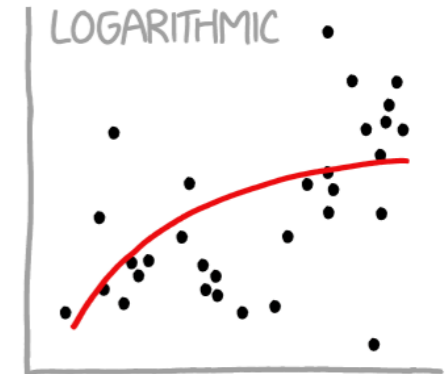
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



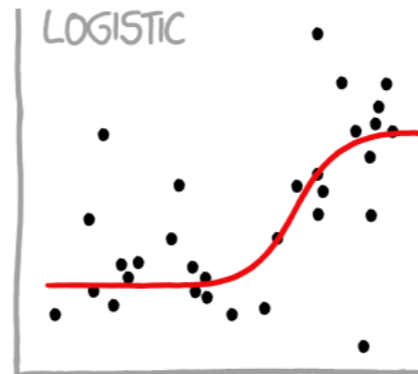
"HEY, I DID A
REGRESSION."



"I WANTED A CURVED
LINE, SO I MADE ONE"



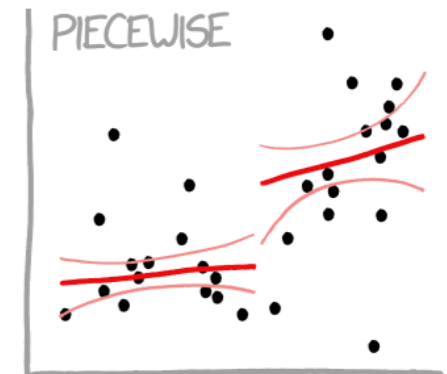
"LOOK, IT'S
TAPERING OFF!"



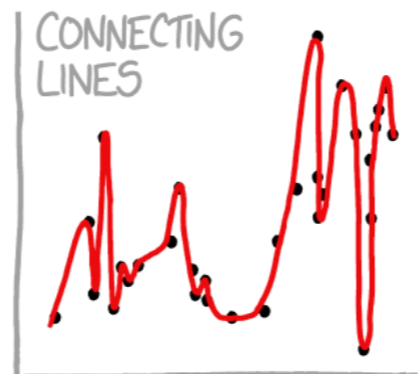
"I NEED TO CONNECT THESE
TWO LINES, BUT MY FIRST IDEA
DIDN'T HAVE ENOUGH MATH."



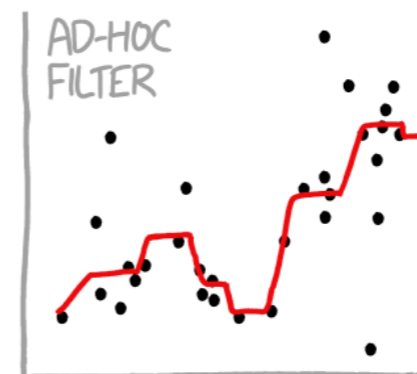
"LISTEN, SCIENCE IS HARD.
BUT I'M A SERIOUS
PERSON DOING MY BEST."



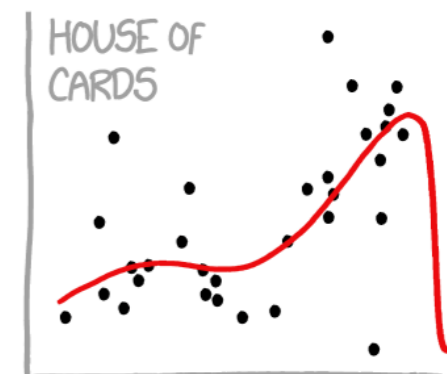
"I HAVE A THEORY,
AND THIS IS THE ONLY
DATA I COULD FIND."



"I CLICKED 'SMOOTH
LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW
TO CLEAN UP THE DATA.
WHAT DO YOU THINK?"



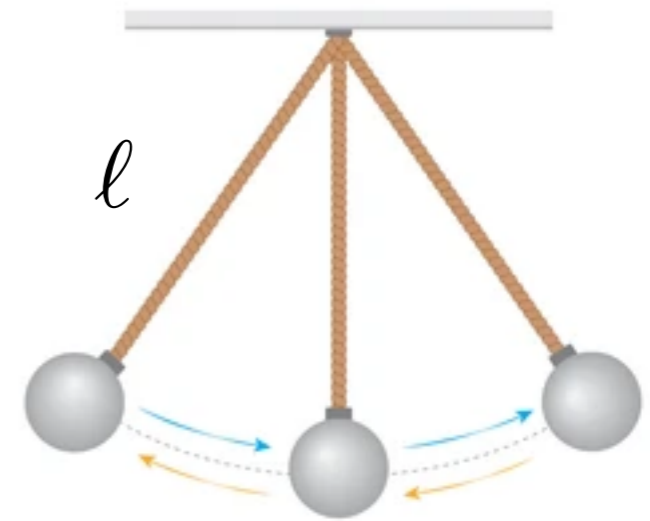
"AS YOU CAN SEE, THIS
MODEL SMOOTHLY FITS THE—
WAIT NO NO DON'T
EXTEND IT AAAAAA!!"

Regression: Predicting a scalar response

$$(x_1, x_2, \dots, x_d) \mapsto y$$

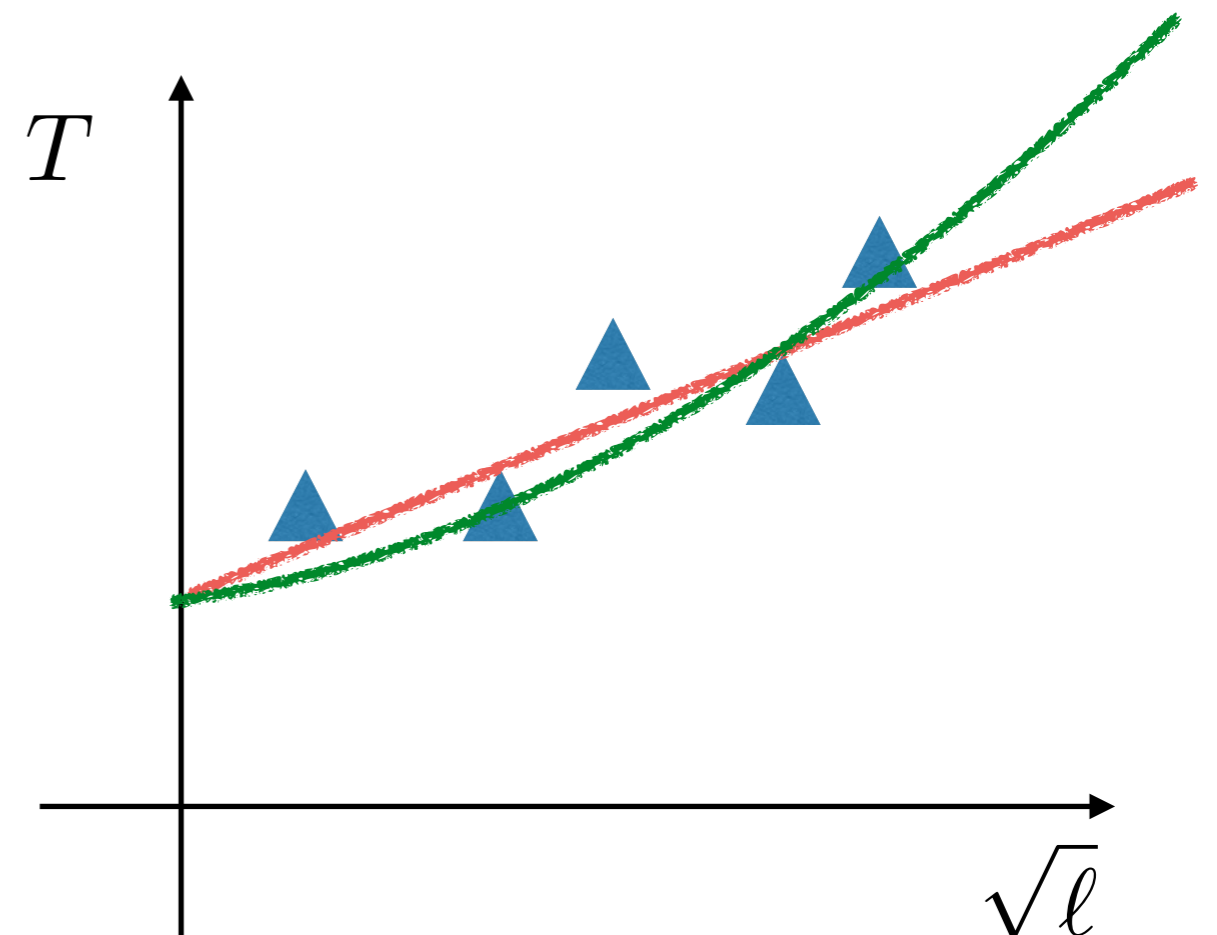
Predictors
Covariates
Features

Response



■ $T = \beta_1 \sqrt{l} + \alpha_1$

■ $T = \gamma_2 l + \beta_2 \sqrt{l} + \alpha_2$



**Easier to discover “correct”
pattern with more data**

Regression: Predicting a scalar response

$$(x_1, x_2, \dots, x_d) \mapsto y$$

Predictors
Covariates
Features

Response

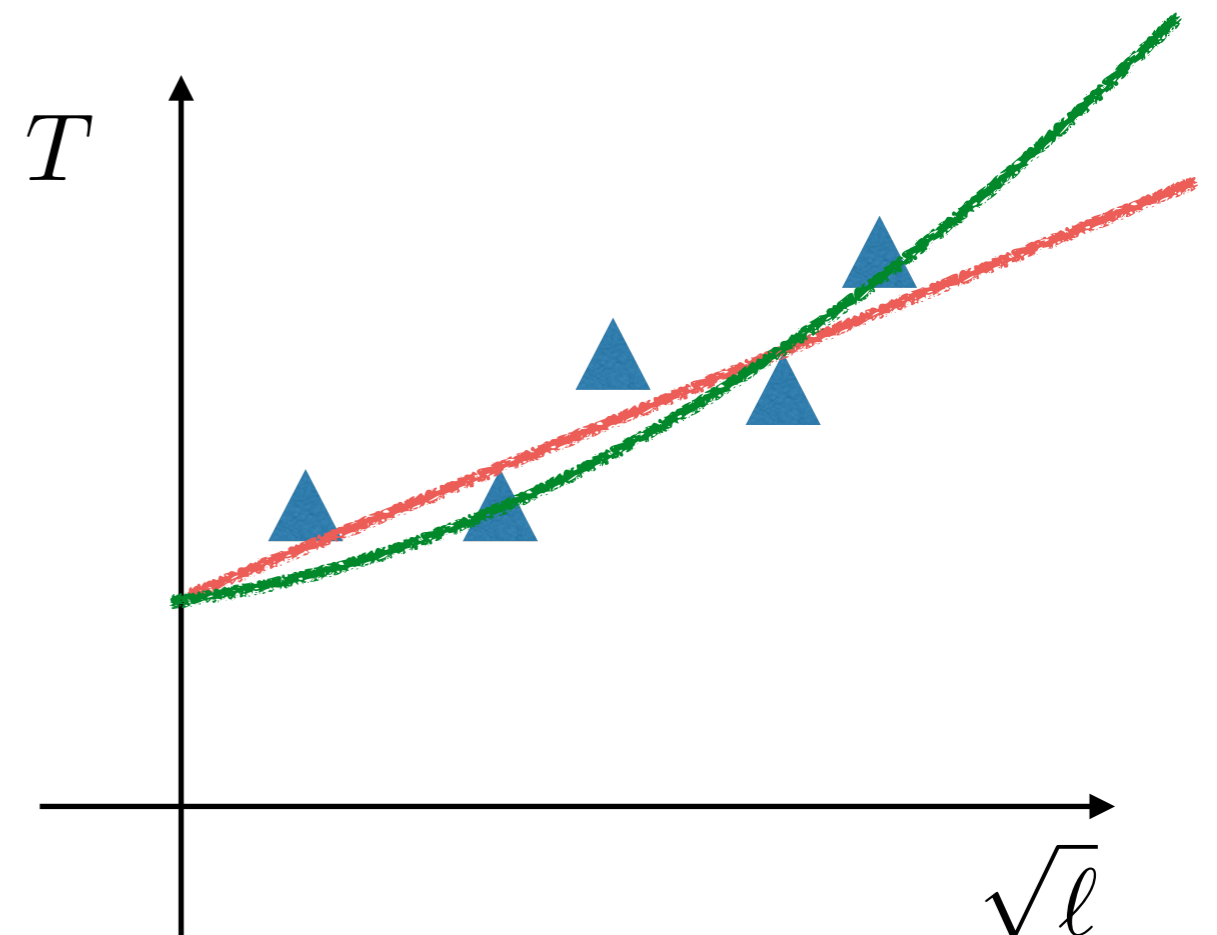
■ $T = \beta_1 \sqrt{l} + \alpha_1$

■ $T = \gamma_2 l + \beta_2 \sqrt{l} + \alpha_2$

- Posit “function class”

$$y \approx f(x_1, \dots, x_d) \quad f \in \mathcal{F}$$

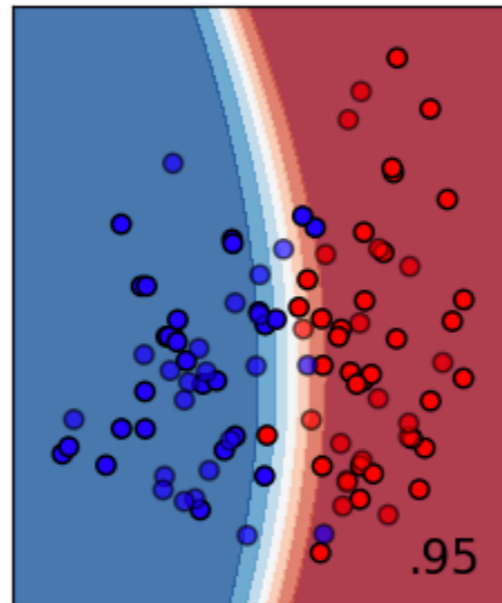
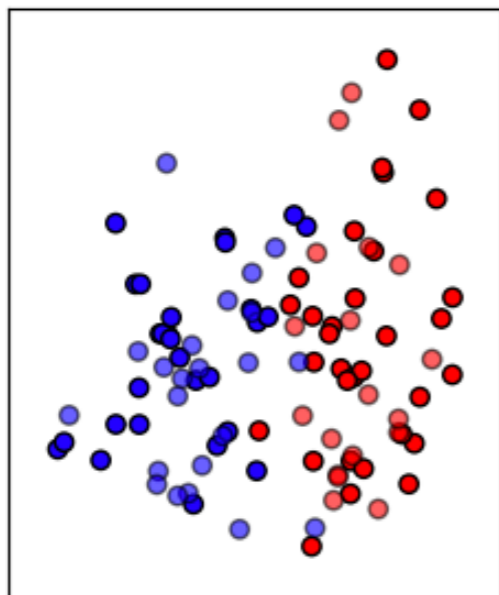
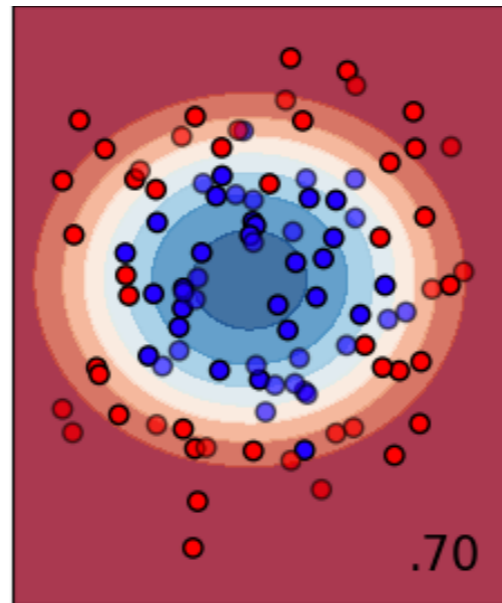
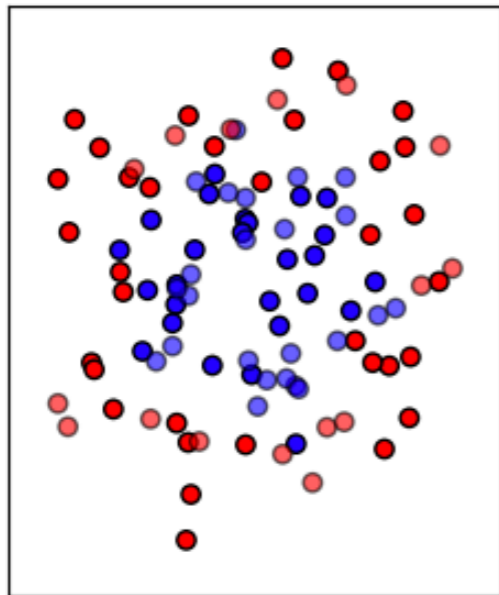
- Learn specific function by minimizing “distance” to data



Classification: Predicting a discrete “class”

$(x_1, x_2, \dots, x_d) \mapsto y$ Discrete, one of k categories

x_1

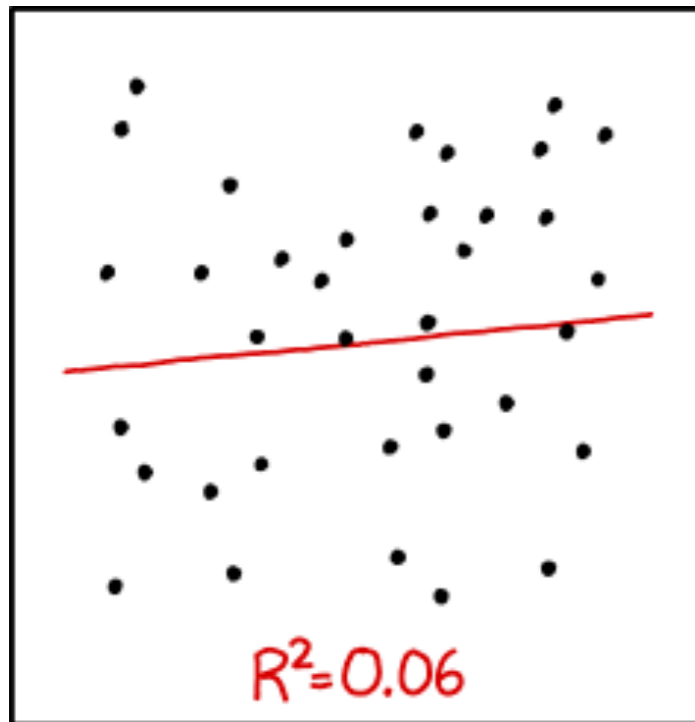


Both can be implemented
As linear separators!

x_2

What we will learn

- Understanding approximation properties of function classes for linear regression and its **infinite dimensional, nonlinear** relatives.
- Classification posed as **predicting probabilities**
- **Incorporating prior structural knowledge within the model**
- Will also cover **unsupervised** approaches



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Statistics: Tools for reliable inference

$$\mathbf{x} := (x_1, x_2, \dots, x_d) \mapsto y$$

Phenomenon

$$\{\mathbf{x}_i, y_i\}_{i=1}^n \quad \text{drawn i.i.d.}$$

Exemplars/data

$$\text{minimize}_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{x}, y} [\mathcal{L}(f; \mathbf{x}, y)]$$

$$\text{minimize}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\mathbf{w}; \mathbf{x}_i, y_i)$$

⋮

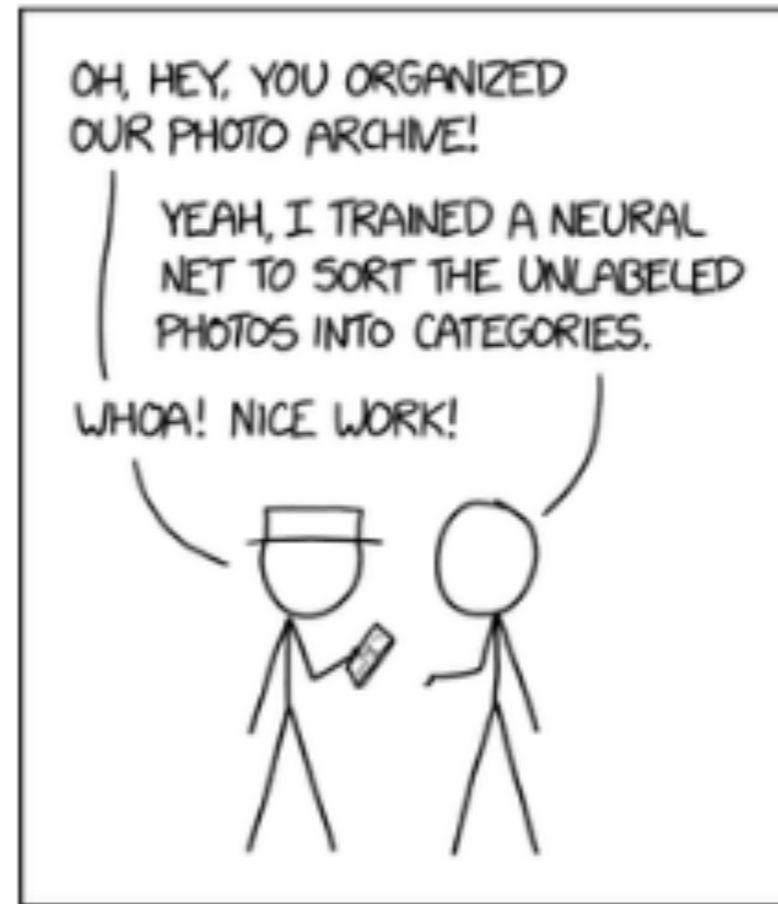
$$\text{minimize}_{\mathbf{w} \in \mathbb{R}^d} \mathbb{E}_{\mathbf{x}, y} [(y - \mathbf{x}^\top \mathbf{w})^2]$$

$$\text{minimize}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

What we will learn

- How to set “distance” function and assumptions on which it relies
- How much data do we need to trust output of the “learned” model?
- How can understanding the “sample complexity” of various approaches guide the design of ML methodology?
- (If there is time): Can we go beyond prediction to confident decision making?

Optimization: Methodology for (fast) model-fitting

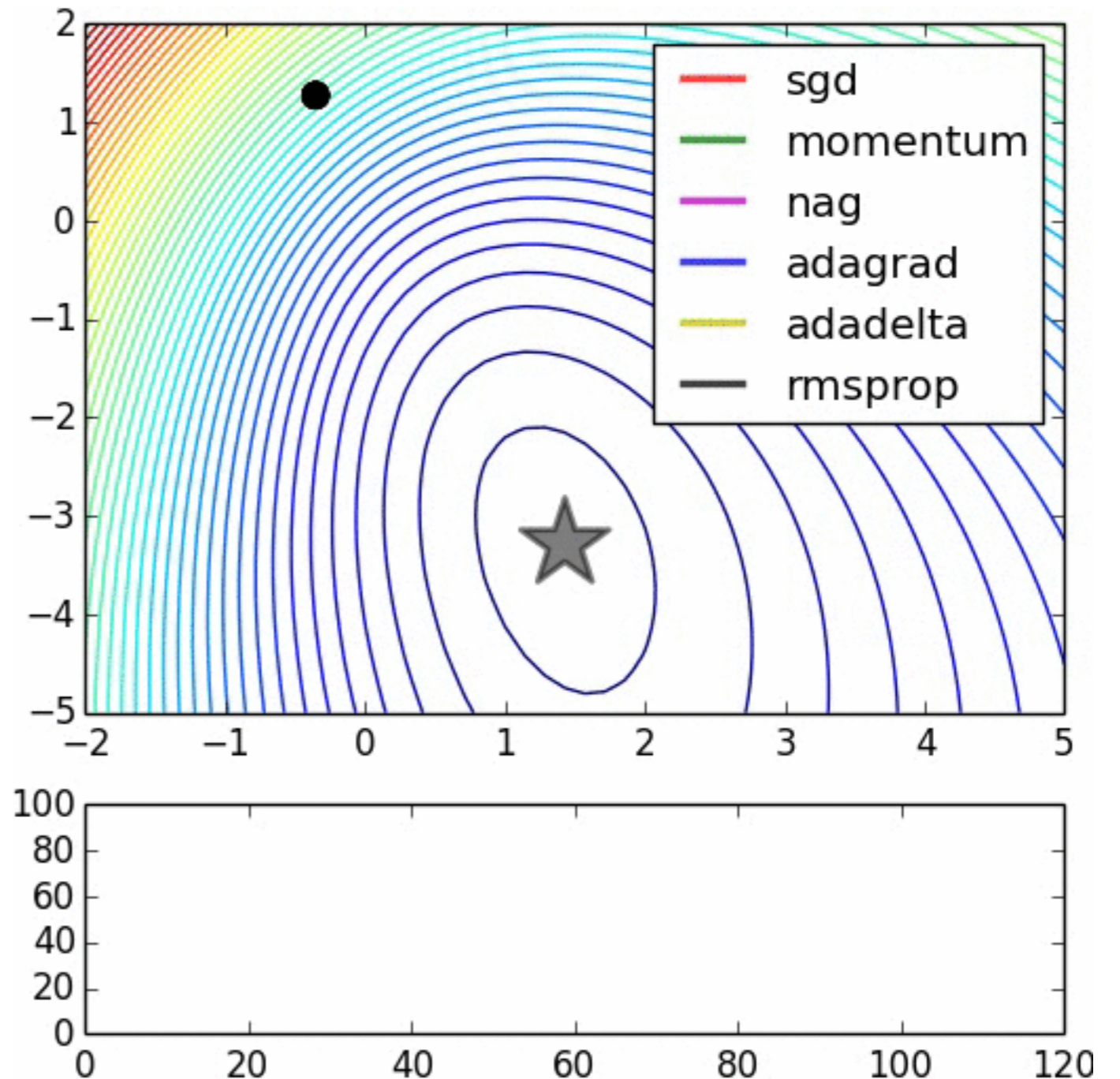


ENGINEERING TIP:
WHEN YOU DO A TASK BY HAND,
YOU CAN TECHNICALLY SAY YOU
TRAINED A NEURAL NET TO DO IT.

$$\text{minimize}_{\mathbf{w}} \quad \mathcal{L}(\mathbf{w}; \mathbf{X}, \mathbf{y})$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}; \mathbf{X}, \mathbf{y})$$

Many variants



The above is a GIF animation.

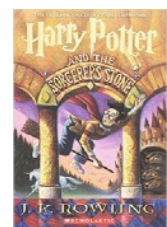
See <http://www.denizyuret.com/2015/03/alec-radfords-animations-for.html>

What we will learn

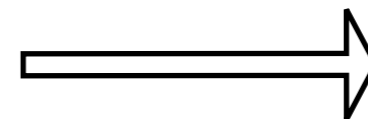
- Posing and solving optimization problems for model fitting
- How to formulate optimization problems for classification and deal with their apparent discrete nature
- How fast do various families of (stochastic) optimization algorithms converge on these problems?

Where is supervised ML (regression) used?

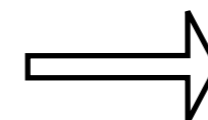
$$(x_1, x_2, \dots, x_d) \mapsto y$$



(100 10 50 ... 20 3000)
← Current warehouse inventory → Target inventory

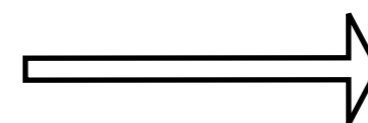


Web data on a product from the past month



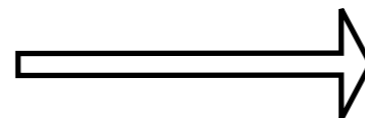
Demand next week

Two-week data from COVID testing center:
Tests, people in each age bracket, positives



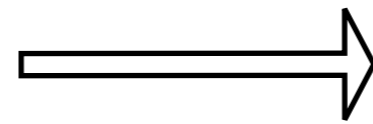
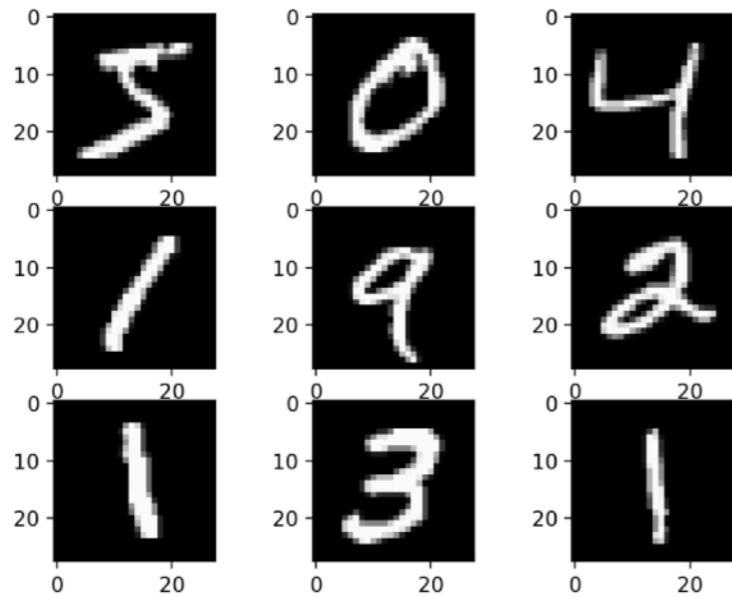
Positivity rate next week

Sensor inputs

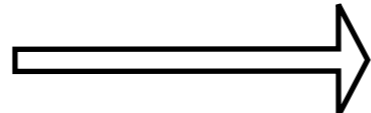


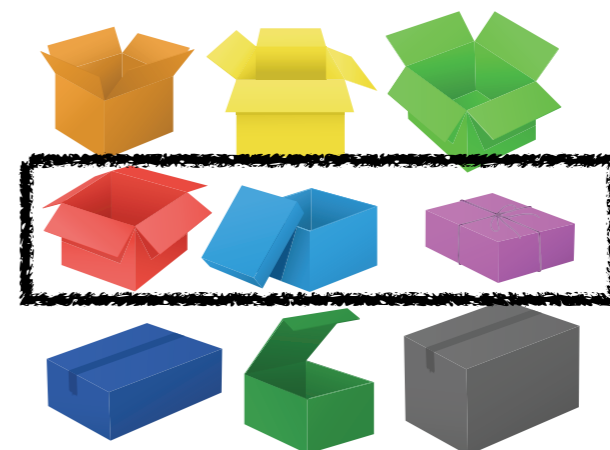
Controls
(Possibly a vector)

Where is supervised ML (classification) used?



One of ten digits

Information about a user  Movie recommendation



Summary

- We will introduce mathematical foundations of probabilistic modeling, optimization-based model-fitting, and statistical inference, focusing on linear algebraic ways of thinking.
- The class will require the maturity to appreciate mathematical arguments and write proofs. There will be plenty of exposure in lecture/HW/exams to this mode of thinking.