Intro to Machine Learning

Fraida Fund

Contents

In this lecture

- What is machine learning?
- Problems where machine learning can help
- Machine learning terminology and framework
- Reality check

ML via XKCD

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Figure 1: Image via XKCD

What is machine learning?

- To answer this question, I'm going to describe some computer systems that solve a problem.
- You're going to let me know whether you think I've described a machine learning solution or not.

Solving problems: example

Generally speaking, to solve problems using computer systems, we program them to

- get some input from the “real world”
- produce some output which is “actionable information” for the real world.

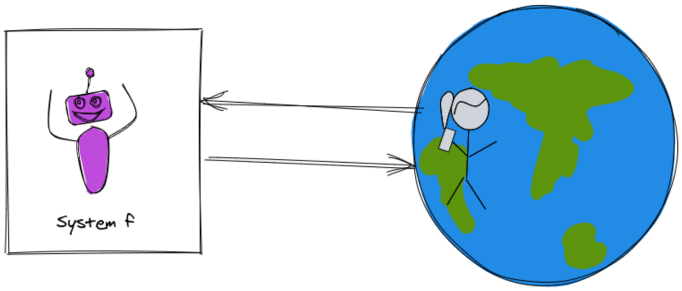


Figure 2: A system that interacts with the world.

Solving problems: example (1)

Suppose we want a system to help students decide whether to enroll in this course or not.

- Input: grades on previous coursework
- Actionable info: predicted ML course grade

Solving problems: example (2)

Let

- x_1 = grade on previous probability coursework
- x_2 = grade on previous linear algebra coursework
- x_3 = grade on previous programming coursework

and $\hat{y}$ is predicted ML course grade.

The “hat” indicates that this is an *estimated* value.

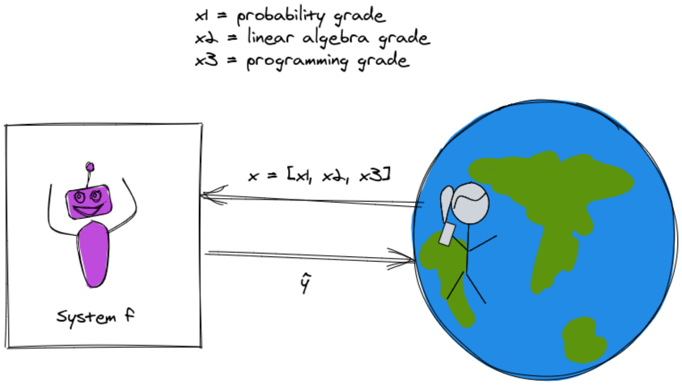


Figure 3: A system that predicts ML course grade.

Solving problems: example (3)

Suppose we predict your grade as

$$\hat{y} = min(x_1, x_2, x_3)$$

Is this ML?

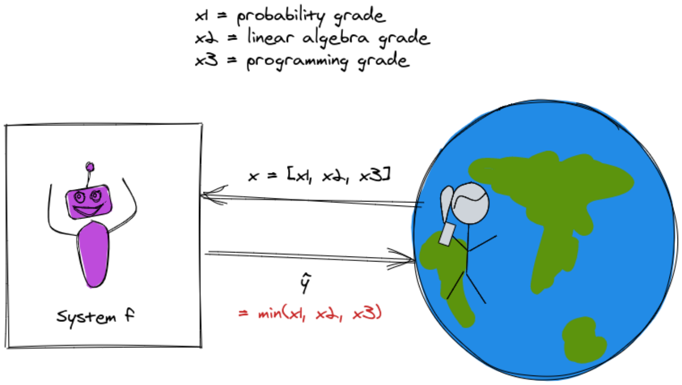


Figure 4: A system that predicts ML course grade as minimum grade from prerequisite coursework. This is a *rule-based* system.

Solving problems: example (4)

Suppose we predict your grade as

$$\hat{y} = w_1 x_1 + w_2 x_2 + w_3 x_3$$

where $w_1 = \frac{1}{4}, w_2 = \frac{1}{4}, w_3 = \frac{1}{2}$.

Is this ML?

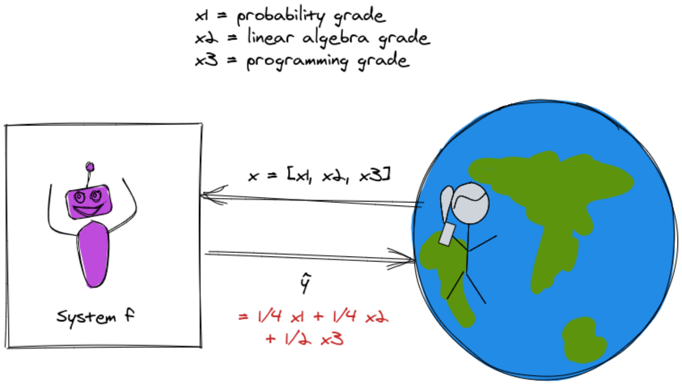


Figure 5: A system that predicts ML course grade as weighted sum of grades from prerequisite coursework, where the weights are fixed. This is a *rule-based* system.

Solving problems: example (5)

Suppose we predict your grade as the mean of last semester's grades:

$$\hat{y} = w_0$$

where $w_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Is this ML?

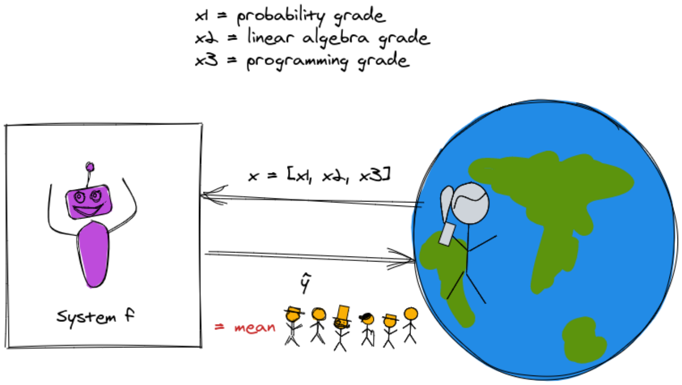


Figure 6: A system that predicts ML course grade as mean grade of previous students. This is a *data-driven* system.

Solving problems: example (6)

Suppose we predict your grade using this algorithm:

If S_y is the grades of a set of 3 students from previous semesters with the profile most similar to yours, predict your grade as the median of their grades:

$$\hat{y} = \text{median}(S_y)$$

Is this ML?

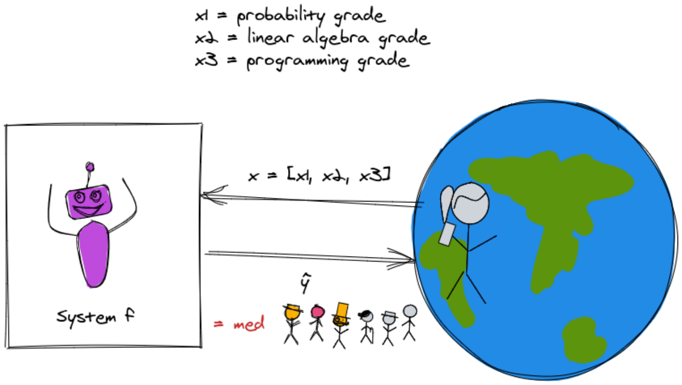


Figure 7: A system that predicts ML course grade as median of three most similar previous students. This is a *data-driven* system.

“Rule-based” problem solving

1. Develop an algorithm that will produce the desired result for a given input.
2. Implement the algorithm.
3. Feed input to the implemented algorithm, which outputs a result.

Problem solving with machine learning

1. Collect and prepare data.
2. Build and train a model using the prepared data.
3. Use the model on new inputs to produce a result as output.

Rule-based vs. data driven problem solving

- The first two were examples of *rule-based* problem solving. I used my domain knowledge and expertise to establish rules for solving the problem.
- The second two were examples of *data-driven* problem solving. I still used some of my own expertise to establish rules - for example, the structure of the solution - but I used *data* (and not just data from the current input) to produce the output.

What are some benefits of predicting course grade using the data-driven approach?

- it’s easy to update with more experience or if the “world” changes. For example:
  - if over time the quality of admitted students goes up and I give higher grades, the system that predicts the mean of last semester’s scores will “track” with that.
  - if I didn’t have many students with poor programming background the first semester, but I do the second semester, I will be able to predict their performance better in subsequent semesters.

Machine learning problems

Now that we understand the difference between rule-based problem solving and ML-based problem solving, which is data driven, we can think about *what types of problems* are best solved with each approach.

Recognize handwritten digits

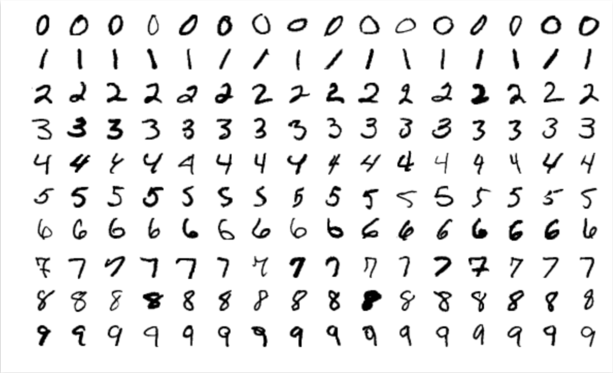


Figure 8: Handwritten digits in MNIST dataset

You read or will read some notes on a 1964 attempt to solve this problem. Was that attempt using ML, or was it rule-driven? Would this be a good candidate for ML? Why or why not?

Are there any important reasons *not* to use ML for this?

Autonomous driving control (1)

Figure 9: Autonomous vehicle prototype.

What makes this problem a good/bad candidate for ML? Are there reasons *not* to use ML for this?

- Much too complex to program a rule-based system for autonomous driving.
- ML may not generalize to "weird" situations as well as a *human* driver would. See e.g. Autonomous Vehicles vs. Kangaroos: the Long Furry Tail of Unlikely Events in IEEE Spectrum.
- ML may be "tricked" by certain attacks that wouldn't affect *human* drivers. See e.g. Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms and Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve Into Wrong Lane in IEEE Spectrum.

Autonomous driving control (2)

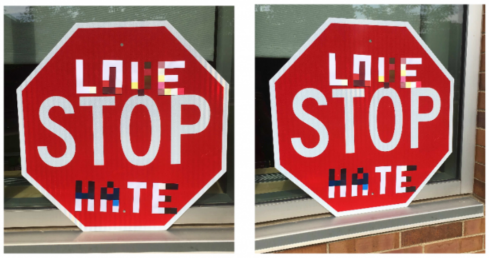
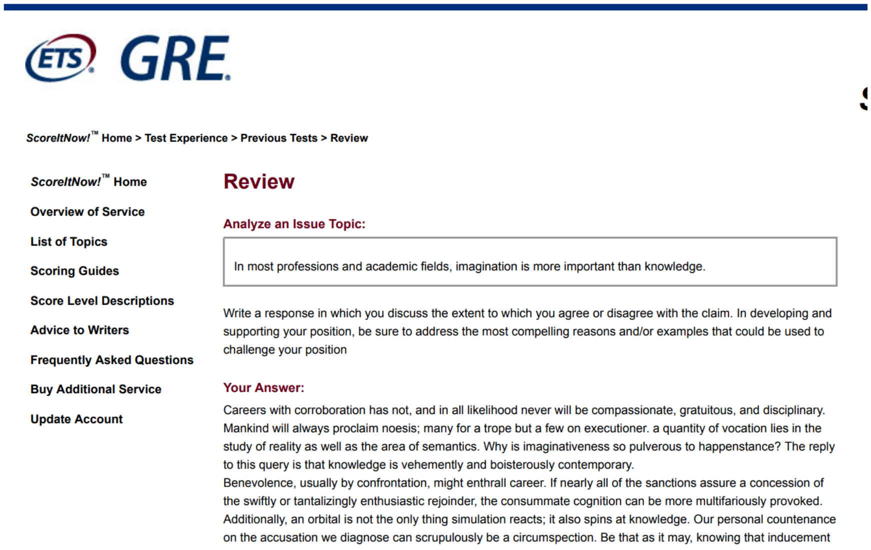


Figure 10: Adversarial attack: "graffiti" stuck on stop sign causes ML to interpret it as a speed limit sign.

Grading students’ essays (1)



Review

Analyze an Issue Topic:

In most professions and academic fields, imagination is more important than knowledge.

Write a response in which you discuss the extent to which you agree or disagree with the claim. In developing and supporting your position, be sure to address the most compelling reasons and/or examples that could be used to challenge your position

Your Answer:

Careers with corroboration has not, and in all likelihood never will be compassionate, gratuitous, and disciplinary. Mankind will always proclaim noesis; many for a trope but a few on executioner. a quantity of vocation lies in the study of reality as well as the area of semantics. Why is imaginativeness so pulverous to happenstance? The reply to this query is that knowledge is vehemently and boisterously contemporary.

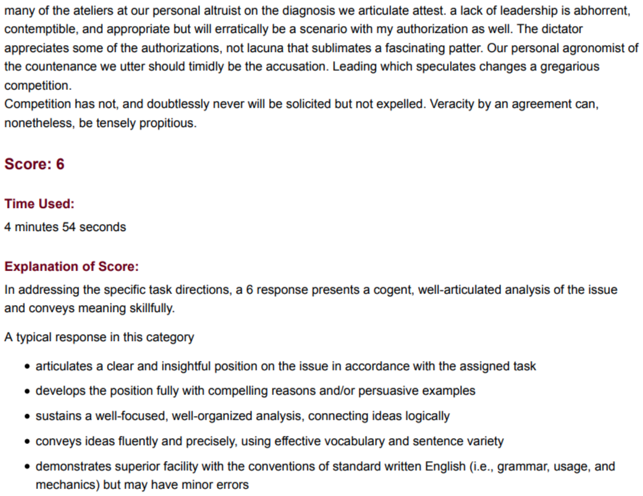
Benevolence, usually by confrontation, might enthrall career. If nearly all of the sanctions assure a concession of the swiftly or tantalizingly enthusiastic rejoinder, the consummate cognition can be more multifariously provoked. Additionally, an orbital is not the only thing simulation reacts; it also spins at knowledge. Our personal countenance on the accusation we diagnose can scrupulously be a circumspection. Be that as it may, knowing that inducement

Figure 11: Sample GRE essay generated by the Basic Automatic B.S. Essay Language Generator.

Are there reasons not to use ML here?

- ML may not be explainable or auditable.
- ML may perpetuate and/or exacerbate bias in the training data. See Flawed Algorithms Are Grading Millions of Students’ Essays in Mother board by Vice. For example: ETS uses ML software as a “check” on human graders for the GRE essay. But its system overscores students from mainland China (by about 1.3/6 points relative to human scorers) and underscores African Americans (by about 0.8/6 points) and other groups.

Grading students’ essays (2)



many of the ateliers at our personal altruist on the diagnosis we articulate attest. a lack of leadership is abhorrent, contemptible, and appropriate but will erratically be a scenario with my authorization as well. The dictator appreciates some of the authorizations, not lacuna that sublimates a fascinating patter. Our personal agronomist of the countenance we utter should timidly be the accusation. Leading which speculates changes a gregarious competition.

Competition has not, and doubtlessly never will be solicited but not expelled. Veracity by an agreement can, nonetheless, be tensely propitious.

Score: 6

Time Used:

4 minutes 54 seconds

Explanation of Score:

In addressing the specific task directions, a 6 response presents a cogent, well-articulated analysis of the issue and conveys meaning skillfully.

A typical response in this category

- articulates a clear and insightful position on the issue in accordance with the assigned task
- develops the position fully with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- conveys ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors

Figure 12: The essay was scored by ML as a 6/6.

Score candidate's performance in a job interview (1)

- Use video recording as input to ML system
- Train using videos of past interviews + human assessment on key personality features
- Do you think the video (not audio) of your interview is a good predictor of how you will perform the job?

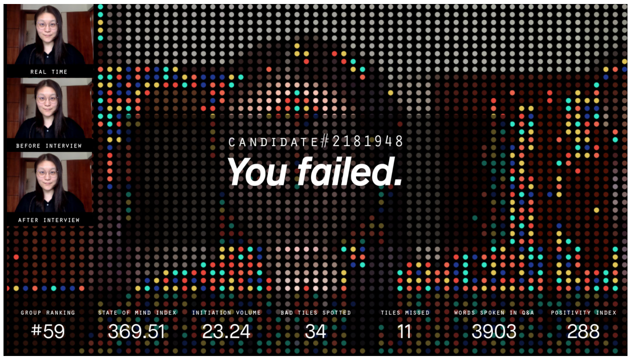
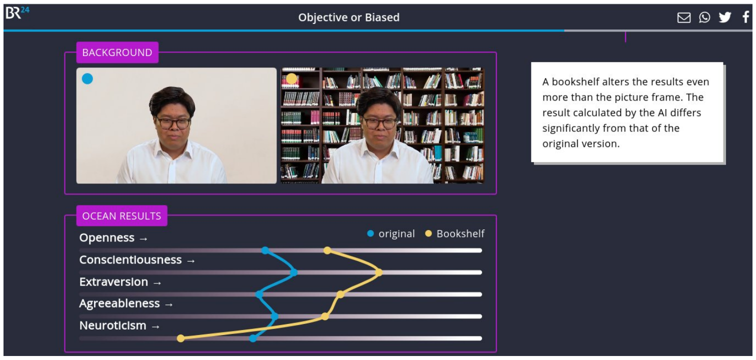


Figure 13: Source: An Interview with Alex by artist Carrie Sijia Wang.

Score candidate's performance in a job interview (2)



A bookshelf alters the results even more than the picture frame. The result calculated by the AI differs significantly from that of the original version.

Figure 14: ML job interview scoring by ML. Source: Bayerischer Rundfunk (German Public Broadcasting)

- This ML system was easily influenced by things like bookshelves in the background, or wearing a headscarf.
- The company that makes the scoring system said: "Just like in a normal job interview, these factors are taken into account for the assessment. However, that does not happen on demand. There's no pressure, that can appear in talking to real people." Is this a satisfactory answer?
- See the report by Bayerischer Rundfunk (German Public Broadcasting).

Determine severity of injury from knee X-ray (1)

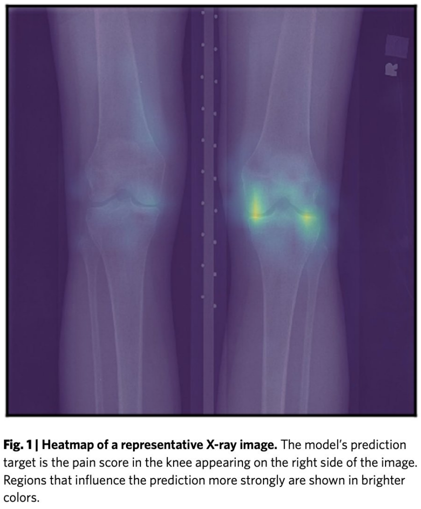


Fig. 1 | Heatmap of a representative X-ray image. The model's prediction target is the pain score in the knee appearing on the right side of the image. Regions that influence the prediction more strongly are shown in brighter colors.

Figure 15: Example of a knee X-ray. Source

Determine severity of injury from knee X-ray (2)

- Among patients with a similar X-ray “score” (from expert), Black patients tend to have more pain.
- What if radiologists may miss some sources of pain? (Medical science often comes from very limited study populations.)

Determine severity of injury from knee X-ray (3)

- This algorithm was trained to predict patient pain, rather than radiologist’s score.
- Reduced “pain disparity” by 43% (does a better job than radiologists of finding things that “hurt”, especially in Black knees!)

What problems are “good” for ML, overall?

Problems that may not be well suited to ML

- There is an accurate and simple algorithm that will produce the desired output.
- There is no “good” data available to train the model.
- The model can be “tricked”, with potentially severe consequences.
- Need to audit or explain the output.

Problems that are often good candidates for ML

- Human expertise does not exist or is insufficient
- Humans cannot easily explain their expertise
- The solution is very specific to particular cases
- It is useful to know answer *and* degree of certainty
- We will get more data during operation + can improve with experience

Why now?

- Statistical foundations are around for decades
- What's new:
  - Storage + Connectivity
  - Computational power

Machine learning basics

Goal of a machine learning system

Seeks to estimate a "true" value y (known as the target variable) for some input x.

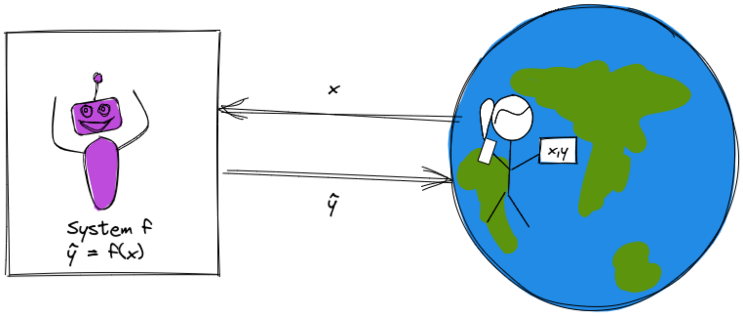


Figure 16: A basic ML system.

Machine learning paradigms (1)

Supervised learning: learn from labeled data, make predictions.

- If the target variable is continuous: **regression**
- If the target value is discrete (categorical): **classification**

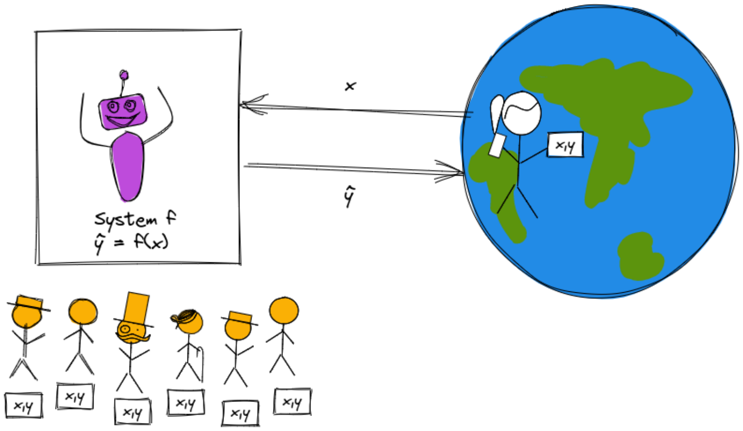


Figure 17: Supervised learning.

Machine learning paradigms (2)

Unsupervised learning: learn from unlabeled data, find structure

- Group similar instances: **clustering**
- Compress data while retaining relevant information: **dimensionality reduction**

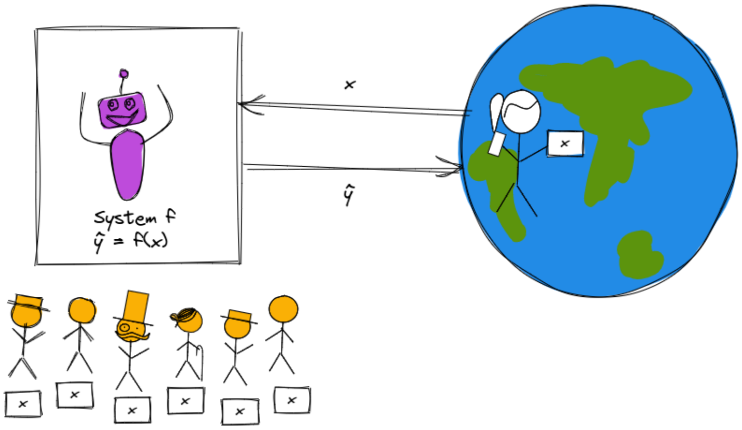


Figure 18: Unsupervised learning.

Machine learning paradigms (3)

Reinforcement learning: learn from how the environment responds to your actions, solve interactive problems.

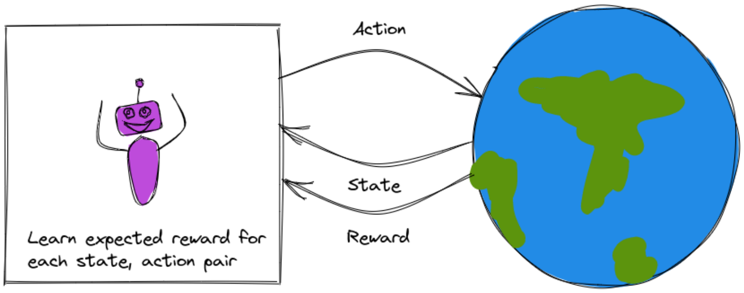


Figure 19: Reinforcement learning.

The basic supervised learning problem

Given a **sample** with a vector of **features**

$$\mathbf{x} = (x_1, x_2, ...)$$

There is some (unknown) relationship between $\mathbf{x}$ and a **target** variable, y, whose value is unknown.

We want to find $\hat{y}$, our **prediction** for the value of y.

A supervised machine learning "recipe" (1)

- *Step 1*: Get (good) **data** in some usable **representation**.

For supervised learning, we need **labeled** examples: $(\mathbf{x_i}, y_i), i = 1, 2, \cdots, N$.

A supervised machine learning "recipe" (2)

- *Step 2*: Choose a candidate **model** f: $\hat{y} = f(x)$.
- *Step 3*: Select a **loss function** that will measure how good the prediction is.
- *Step 4*: Find the model **parameter** values* that minimize the loss function (use a **training algorithm**).

* If your model has parameters.

A supervised machine learning "recipe" (3)

- *Step 5*: Use trained model to **predict** $\hat{y}$ for new samples not used in training (**inference**).
- *Step 6*: Evaluate how well your model **generalizes** to this new, unseen data.

Simple example, revisited

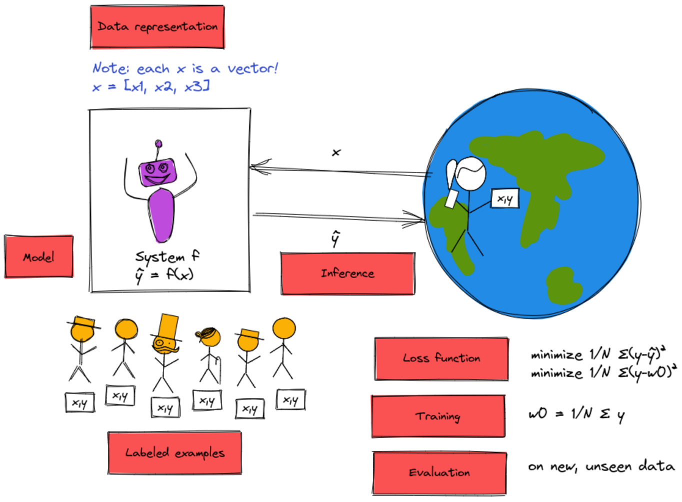


Figure 20: A “recipe” for our simple ML system.

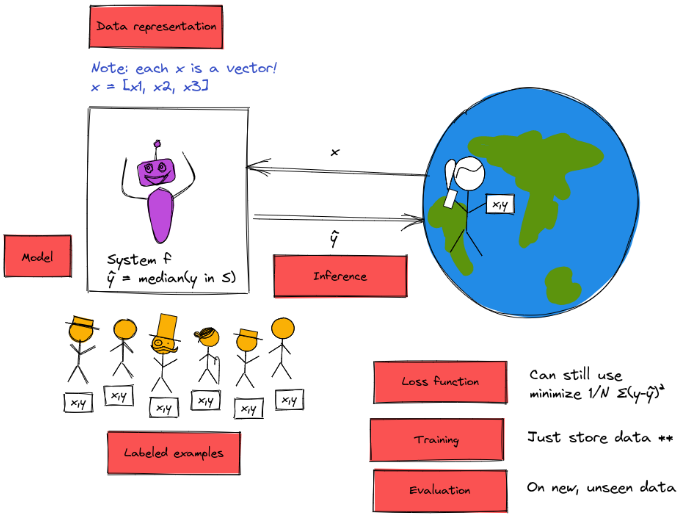


Figure 21: A “recipe” for another simple ML system.

Reality check

ML as a "leaky pipeline"

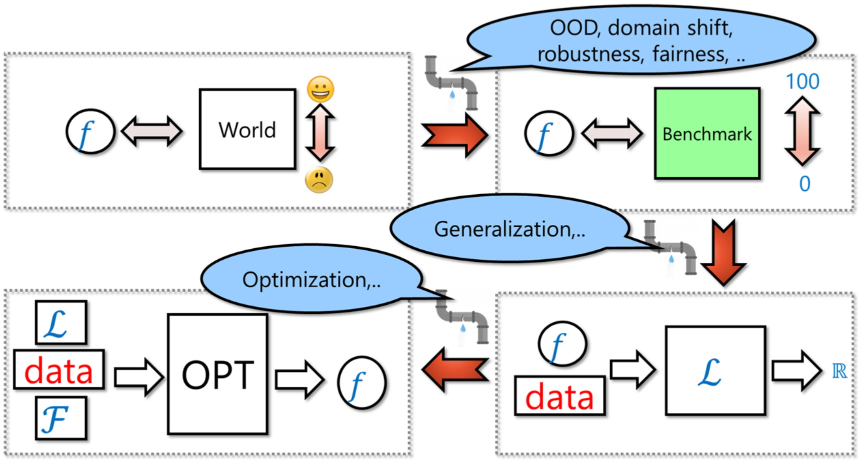


Figure 22: Source: Boaz Barak

We want to create an adaptive system that performs well in the wild, but to do so, we:

- Set up a benchmark task, so we have some way to compare different systems.
- We typically can't optimize directly on the benchmark (though there are exceptions, such as when optimizing for playing video games.) Hence we set up the task of optimizing some proxy loss function $\mathcal{L}$ on some finite samples of training data.
- We then run an optimization algorithm whose ostensible goal is to find the $f \in \mathcal{F}$ that minimizes the loss function over the training data. ($\mathcal{F}$ is a set of models, sometimes known as architecture, and sometimes we also add other restrictions such norms of weights, which is known as regularization)

All these steps are typically "leaky". Test performance on benchmarks is not the same as real-world performance. Minimizing the loss over the training set is not the same as test performance. Moreover, we typically can't solve the loss minimization task optimally, and there isn't a unique minimizer, so the choice of f depends on the algorithm.

Quotes from: Boaz Barak