

CMSC848P Advanced Topics in Machine Learning Theory

Lecture 1: Logistics & Introduction

Han Shao

About Me

- Assistant Professor of Computer Science
- I work primarily on **machine learning theory, economics and computation**, and topics at the intersection of these fields.
- <https://www.cs.umd.edu/~hanshao/>
- hanshao@umd.edu

Course Information

My office hours: Monday 4:45–5:30,
IRB 5132

Teaching assistant: Shashaank Aiyer

TA office hours: Thursday 3:30 -4:30,
AVW 4117

Course website

Lecture notes, and syllabus

cs.umd.edu/~hanshao/cmsc848p-26.html

Canvas

Announcements, assignments, and Q&A

Coursework and Grading

Homework

35%

Midterm

25%

Final

25%

Final Project

15%

Homework and Exams

HOMework

- 4-5 homework assignments
- A question set will be posted at least two weeks before the deadline.
- At the beginning of class on the due date, we will randomly select one or two questions for you to complete in class.
- No devices or notes are allowed during the in-class assessment.
- Your homework grade will be based solely on your in-class submission.
- If you cannot attend, notify us at least 48 hours in advance.

EXAMS

- Exams will consist of proof-based questions.
- For medical absences, a doctor note is required.
- You can secure a B by putting sufficient effort into the homework and exams.
- You must complete the final project to earn an A or A+.

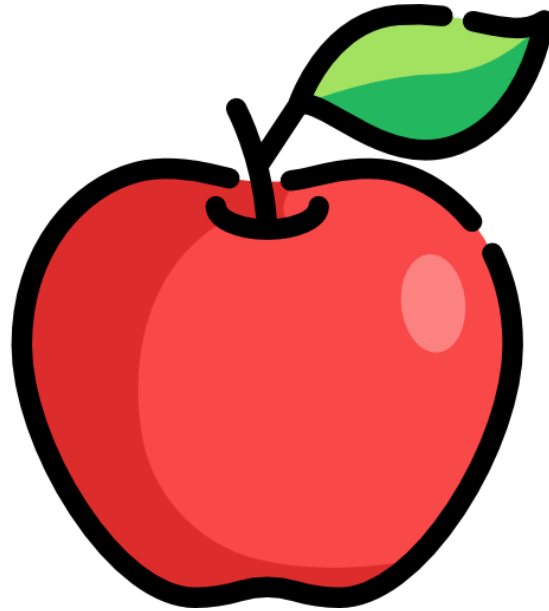
Goals of machine learning theory



Develop and analyze models to understand:

- what kinds of tasks we can hope to learn, and from what kind of data,
- what types of guarantees might we hope to achieve,
- other common issues that arise.

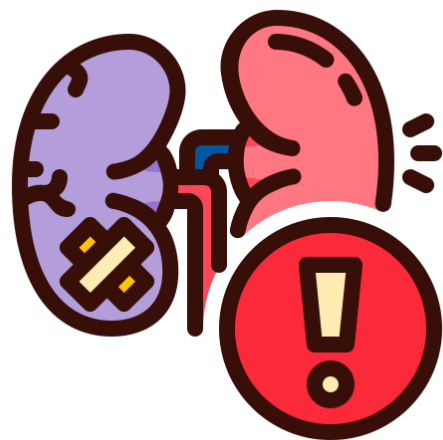
Are Apples Tasty?



Are Apples Tasty?

(color, size)	(green, small)	(red, large)	(red, small)	...
Tasty?	X	✓	✓	

Which Treatment Would You Choose?



Kidney Stone
Treatments

If we observe only the treatment:

Treatment	Successful Cases / Total	Success Rate
A: Open surgery	273 / 350	78%
B: Minimally invasive	289 / 350	83%

Conclusion: Choose Treatment B.

Now Add One Feature: Stone Size

The two treatments were used on very different mixtures of patients.

Stone Size	A: Open Surgery	B: Minimally Invasive
Small	81 / 87 = 93%	234 / 270 = 87%
Large	192 / 263 = 73%	55 / 80 = 69%

Treatment A has the higher success rate **for both stone sizes.**

Same data, different features → different conclusion.

x = treatment vs. x = (treatment, stone size)

How Should We Collect Feedback for an LLM?

Prompt: “Draft a cover letter for a machine learning internship.”

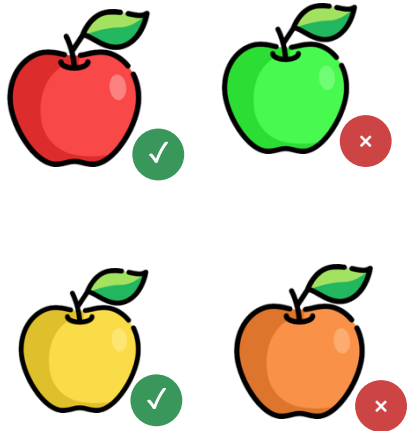
Goal: *Train an LLM assistant for students.*

Data Type	Example Observation	Signal
Demonstration	A human writes an ideal cover letter. (prompt, ideal response)	A target output
Rating	A human scores one draft: 4/5. (prompt, response, score)	An absolute quality score
Pairwise Comparison	A human chooses Draft A over Draft B. (prompt, A, B, $A \succ B$)	A relative preference

Pairwise comparisons are often easier and more reliable for people to provide than absolute ratings.

Training: From Data to a Predictor

Training Sample S



Learning Algorithm A

Predictor

$$h : X \rightarrow \{0, 1\}$$

Apple $\rightarrow$ tasty / not tasty

Central questions:

- (A) How can we generate rules that fit the observed data?
- (B) How confident are we that they will perform well on future data?

Natural Formalization

DATA-GENERATING PROCESS

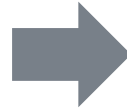


Fix one taster.

$$x_1, \dots, x_m \stackrel{\text{i.i.d.}}{\sim} D$$

$$y_i = f(x_i) \in \{0, 1\}, \quad S = \{(x_i, y_i)\}_{i=1}^m$$

D is the population of future apples.



LEARNING



$$A(S) = h \in \mathcal{H}$$

Use the sample S to choose h .

Ideally, $\text{err}_S(h)$ is small.

TRAINING ERROR (OBSERVED)

$$\text{err}_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}[h(x_i) \neq y_i]$$

What we can measure from S

TRUE ERROR (FUTURE)

$$\text{err}_D(h) = \Pr_{x \sim D} [h(x) \neq f(x)]$$

Performance on a new apple

GOAL

$$\Pr_{S \sim D^m} [\text{err}_D(A(S)) \leq \varepsilon] \geq 1 - \delta$$

For now: a noiseless, realizable model.

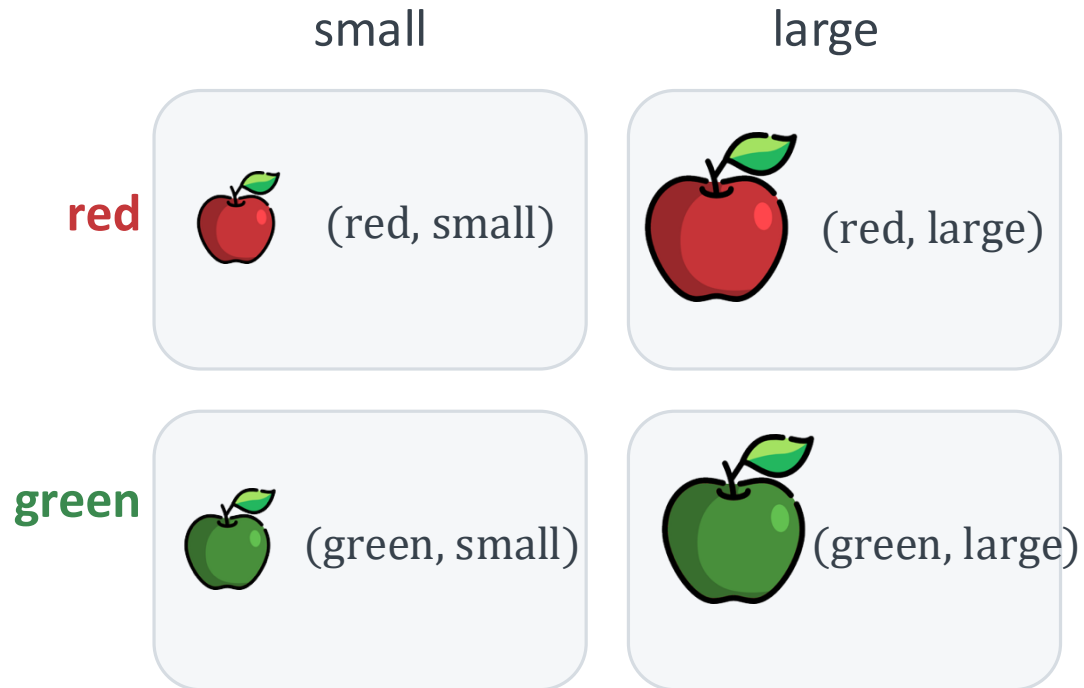
accuracy ε

confidence $1 - \delta$

A Finite Apple Hypothesis Class

Toy feature representation: color $\in \{\text{red}, \text{green}\}$, size $\in \{\text{small}, \text{large}\}$

Hypothesis class $H = \{h : X \rightarrow \{0,1\}\}$: all binary labelings of the four types



4 possible input types

$$|X| = 2 \times 2 = 4$$

16 possible binary predictors

$$|H| = 2^4 = 16$$

Learning rule: choose a consistent $h \in H$.

Consistent = zero training error. Since $f \in H$, one exists.

Confidence and Sample Complexity

Assume $\mathbf{f} \in \mathbf{H}$, and the learner returns a predictor with zero training error.

1 FIX A BAD PREDICTOR

$$\text{err}_D(h) \geq \varepsilon$$

A random apple exposes a mistake with probability at least ε .

2 COULD IT FOOL US?

$$\Pr[\text{err}_S(h) = 0] \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}$$

All m training apples must miss its errors.

3 UNION BOUND OVER $\mathbf{H}$

$$\Pr[\exists \text{ bad consistent } h] \leq |\mathcal{H}|e^{-\varepsilon m} = 16e^{-\varepsilon m}$$

There are only $|\mathcal{H}| = 16$ predictors that might fool us.

Make the failure probability at most δ :

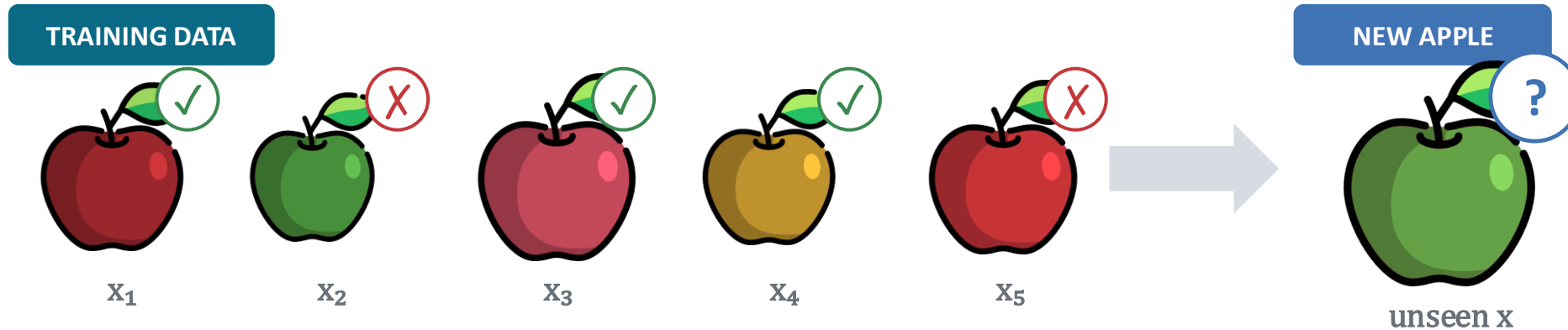
$$m \geq \frac{\ln |\mathcal{H}| + \ln(1/\delta)}{\varepsilon} = \frac{\ln 16 + \ln(1/\delta)}{\varepsilon}$$

Example: $\varepsilon = 0.10$ and $\delta = 0.05 \Rightarrow m \geq 58$ is sufficient.

With probability at least $1 - \delta$, every consistent predictor has true error $< \varepsilon$.

But No Two Apples Are Exactly Alike

$$\mathbf{x} = (\text{hue, firmness, sweetness, acidity, size, ...}) \in \mathbb{R}^d$$



h_1

Fits every observed apple

Prediction: The new apple is tasty. ✓

h_2

Fits every observed apple

Prediction: The new apple is not
tasty. ✗

Both have zero training error but disagree on the new apple.

Memorization is not generalization: The sample does not determine labels on unseen apples.

Occam's Razor: Learning Needs Assumptions

One possible assumption: The world is well approximated by a simple hypothesis.

MEMORIZATION

A separate rule for each observed apple

if $x = x_1$: tasty
if $x = x_2$: not tasty
if $x = x_3$: tasty
:
otherwise: ?

0
training
error

HIGH
complexity

SIMPLE RULE

One rule applies to many apples

$$h(x) = \mathbb{1}[\text{color}(x) = \text{red}]$$

green apple
predict: not tasty

red apple
predict: tasty

Occam's principle: Among hypotheses that explain the data, prefer the simpler one.

$$\text{true error} \lesssim \text{training error} + \text{complexity penalty}$$

Learning theory asks how to define complexity—and how many samples make this guarantee valid.

After Training: Do We Need a Test Set?

THEORY

Generalization from the training sample

In our finite, realizable apple model, enough i.i.d. training data make every consistent predictor accurate with high probability.

$$\widehat{\text{err}}_S(h) = 0 \implies \text{err}_D(h) \leq \varepsilon$$

This conclusion uses the assumptions: a fixed hypothesis class, the same distribution at training and deployment, and a sufficiently large sample.

A separate test set is not required for this theorem.

TESTING

An independent empirical estimate

Reserve a test sample $T \sim D^n$ that was not used to choose or tune h .

$$\widehat{\text{err}}_T(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[h(x_i) \neq y_i]$$

For a fixed final predictor h , with probability at least $1-\delta$:

$$|\widehat{\text{err}}_T(h) - \text{err}_D(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}$$

Testing measures performance; it does not create generalization.

Key rule: if test data influence the choice of h , they become validation data—not a clean final test set.

Course Project



The course project is research-oriented.

You will conduct a research project on **generative AI/LLM evaluation**.

Steps

1. Identify the aspect you want to evaluate.
2. Formulate it as a mathematical problem.
3. Analyze the problem using tools you learn from the class.
4. Possibly validate your conclusions using real-world data.