ECE595 / STAT598: Machine Learning I
Lecture 22 Is Learning Feasible?

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Learning Theory

- Welcome to Part 3 of ECE595 / STAT598!
- Here is what we have learned:
    - Part 1: The machine learning pipeline
    - Part 2: Classification methods
- What are we going to do in Part 3?
    - Now that we have a method, then what?
    - Will it do well? How well?
    - Will it fail? When?
    - Complex model = better?
    - More sample = better?
    - Can every problem be solved by learning?
    - When do you overfit?
    - How to avoid overfit?

Outline

Today's Lecture:

- What constitutes a learning problem?
  - Training and testing samples
  - Target and Hypothesis function
  - Learning Model
- Is learning feasible?
  - An example
  - The power of probability
- Training versus Testing
  - In-sample error
  - Out-sample error
  - Probability bound

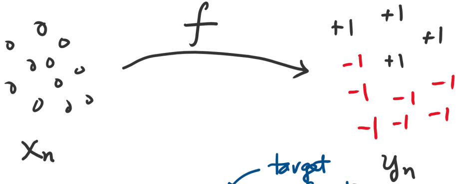
Reference:

- Learning from Data, chapter 1.3

Dataset

Let us first talk about a dataset:

- Input vectors: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$
- Labels: $y_1, \ldots, y_N$
- Training set: $\mathcal{D}$
- Target function f: Maps $\boldsymbol{x}_n$ to y_n
- Target function is always **unknown** to you



target function

$$y_n = f(x_n)$$

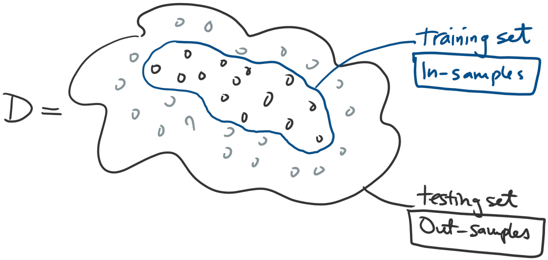
label

input

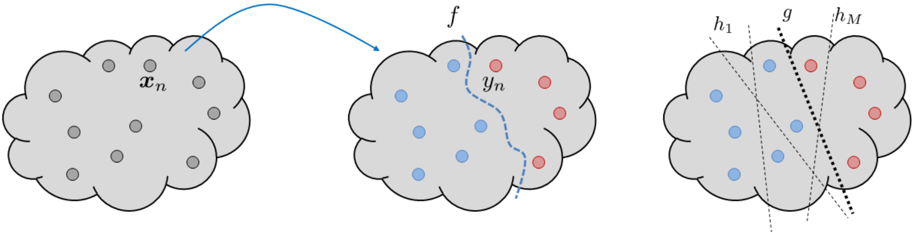
Training and Testing Set

Let us first talk about a dataset:

- **In-sample**: Samples that are inside the training set
- **Out-sample**: Samples that are outside the training set

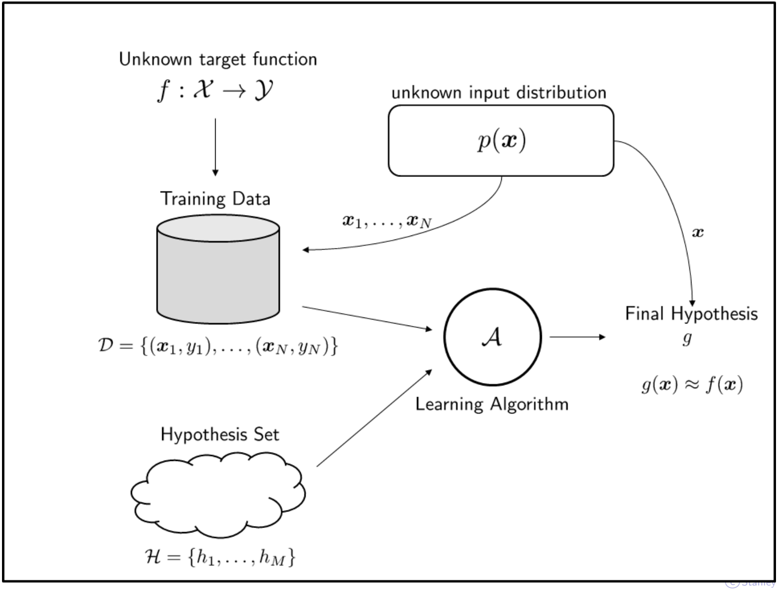


Hypothesis Fucntion



- Hypothesis set: $\mathcal{H} = \{h_1, \ldots, h_M\}$: Possible decision boundaries
- Algorithm: Picks h_m from $\mathcal{H}$
- Final hypothesis: g: The one you found

Learning Model



Outline

Today's Lecture:

- What constitutes a learning problem?
  - Training and testing samples
  - Target and Hypothesis function
  - Learning Model
- Is learning feasible?
  - An example
  - The power of probability
- Training versus Testing
  - In-sample error
  - Out-sample error
  - Probability bound

Is Learning Feasible?

In-sample and **Out-sample**:

- In-sample: Training Data
- Out-sample: Testing Data

When can we claim "learning is feasible"?

Suppose we have a training set $\mathcal{D}$, can we learn the target function f?

- "Learn" means: I use the data you give me to come up with an f
- "Successful" means: All in-samples are correctly predicted
- And all out-samples are also correctly predicted
- If YES, then we are in business.
- Learning is feasible!
- If NO, then we can go home and sleep.
- There is just no way to learn f from $\mathcal{D}$.

Example

- Let $\mathcal{X} = \{0, 1\}^3$
- Each $\boldsymbol{x} \in \mathcal{X}$ is a binary vector
- E.g., $\boldsymbol{x} = [0,\ 0,\ 1]^T$ or $\boldsymbol{x} = [1,\ 0,\ 1]^T$
- How many possible vectors are there? $2^3 = 8$
- Call them $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8$
- There is a target function f
- f maps every $\boldsymbol{x}$ to a y
- $y \in \{+1, -1\}$
- E.g., $f([0, 0, 1]) = +1$, $f([0, 1, 1]) = -1$, etc.
- How many possible f's?
- You can think of f as a 8-bit vector
- E.g., $f = [+1, -1, -1, -1, +1, +1, +1, -1]$.
- So there are $2^8 = 256$ possible f's.

Example

- We have 8 input vectors: $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8\}$
- We have 256 hypotheses: $\mathcal{H} = \{h_1, \ldots, h_{256}\}$
- Is learning feasible?
- Give me a subset $\mathcal{D} \subset \mathcal{X}$, can I find a hypothesis $g \in \mathcal{H}$ such that $g = f$?
- Suppose here is what you are given: $\circ = -1$, $\bullet = +1$. You know 6 out of 8. These are the training data.

$\boldsymbol{x}_n$			y_n
0	0	0	$\circ$
0	0	1	$\bullet$
0	1	0	$\bullet$
0	1	1	$\circ$
1	0	0	$\bullet$
1	0	1	$\circ$
1	1	0	?
1	1	1	?

Possibility 1

$\boldsymbol{x}_n$			y_n
0	0	0	$\circ$
0	0	1	$\bullet$
0	1	0	$\bullet$
0	1	1	$\circ$
1	0	0	$\bullet$
1	0	1	$\circ$
1	1	0	$\circ$
1	1	1	$\circ$

- One 1's will give me $\bullet$; Others give me $\circ$
- So the last two entries should be $\circ$

Possibility 2

$\mathbf{x}_n$			y_n
0	0	0	$\circ$
0	0	1	$\bullet$
0	1	0	$\bullet$
0	1	1	$\circ$
1	0	0	$\bullet$
1	0	1	$\circ$
1	1	0	$\circ$
1	1	1	$\bullet$

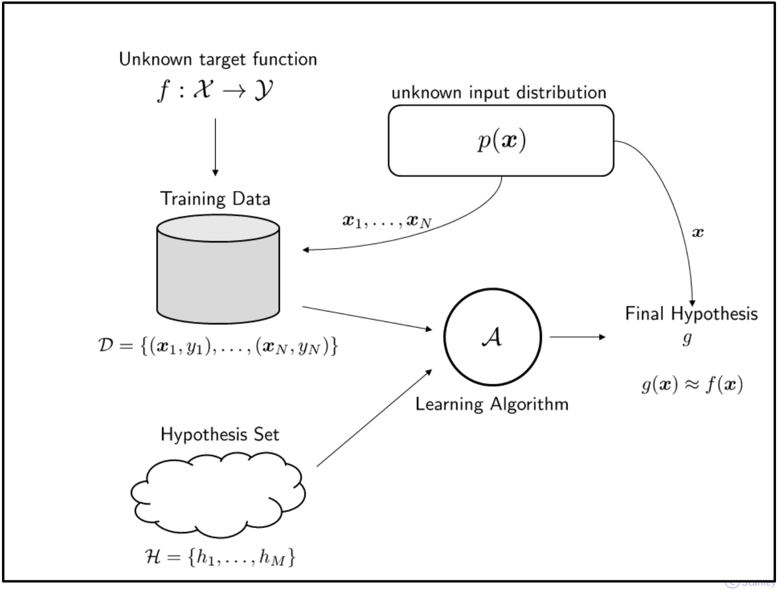
- Odd numbers of 1's give me $\bullet$
- Even numbers of 1's give me $\circ$
- So [1 1 0] should be $\circ$
- So [1 1 1] should be $\bullet$

All the Possibilities

	$\boldsymbol{x}_n$		y_n	g	f_1	f_2	f_3	f_4
0	0	0	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
0	0	1	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
0	1	0	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
0	1	1	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
1	0	0	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
1	0	1	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
1	1	0		$\circ/\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$
1	1	1		$\circ/\bullet$	$\circ$	$\circ$	$\bullet$	$\bullet$

- f_1, f_2, f_3, f_4 are the only hypotheses you need to consider
- You just don't know which one out of the four to choose!
- You won't do better the random guess.
- So you haven't learned anything from the training data.
- Learning is infeasible.

The Power of Probability



Outline

Today's Lecture:

- What constitutes a learning problem?
  - Training and testing samples
  - Target and Hypothesis function
  - Learning Model
- Is learning feasible?
  - An example
  - The power of probability
- Training versus Testing
  - In-sample error
  - Out-sample error
  - Probability bound

In-Sample Error

- Let $\boldsymbol{x}_n$ be a *training* sample
- h: Your hypothesis
- f: The unknown target function
- If $h(\boldsymbol{x}_n) = f(\boldsymbol{x}_n)$, then say training sample $\boldsymbol{x}_n$ is correctly classified.
- This will give you the **in-sample error**

Definition (In-sample Error / Training Error)

Consider a training set $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, and a target function f. The **in-sample error** (or the training error) of a hypothesis function $h \in \mathcal{H}$ is the empirical average of $\{h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)\}$:

$$E_{\text{in}}(h) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} [\![h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)]\!], \tag{1}$$

where $[\![\cdot]\!] = 1$ if the statement inside the bracket is true, and $= 0$ if the statement is false.

Out-Sample Error

- Let $\boldsymbol{x}$ be a *testing* sample drawn from $p(\boldsymbol{x})$
- h: Your hypothesis
- f: The unknown target function
- If $h(\boldsymbol{x}) = f(\boldsymbol{x})$, then say testing sample $\boldsymbol{x}$ is correctly classified.
- Since $\boldsymbol{x} \sim p(\boldsymbol{x})$, you need to compute the probability of error, called the **out-sample error**

Definition (Out-sample Error / Testing Error)

Consider an input space $\mathcal{X}$ containing elements $\boldsymbol{x}$ drawn from a distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$, and a target function f. The **out-sample error** (or the testing error) of a hypothesis function $h \in \mathcal{H}$ is

$$E_{\text{out}}(h) \stackrel{\text{def}}{=} \mathbb{P}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})], \tag{2}$$

where $\mathbb{P}[\cdot]$ measures the probability of the statement based on the distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$.

In-sample VS Out-sample

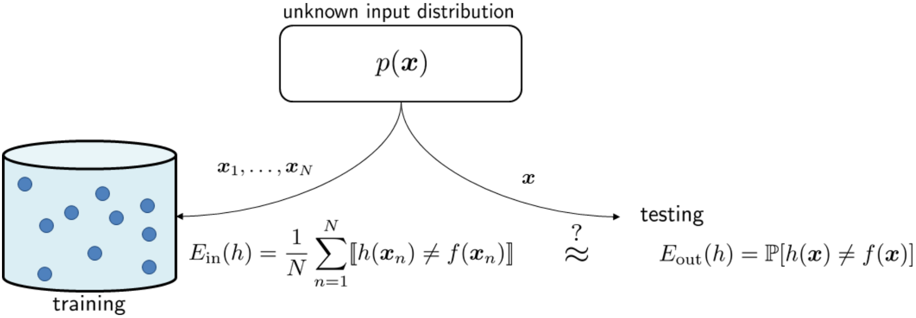
In-Sample Error

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N} [\![h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)]\!]$$

Out-Sample Error

$$\begin{aligned} E_{\text{out}}(h) &= \mathbb{P}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})] \\ &= \underbrace{[\![h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)]\!]}_{=1} \mathbb{P}\Big\{ h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n) \Big\} \\ &\qquad + \underbrace{[\![h(\boldsymbol{x}_n) = f(\boldsymbol{x}_n)]\!]}_{=0} \Big(1 - \mathbb{P}\Big\{ h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n) \Big\} \Big) \\ &= \mathbb{E}\Big\{ [\![h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)]\!] \Big\} \end{aligned}$$

The Role of $p(x)$



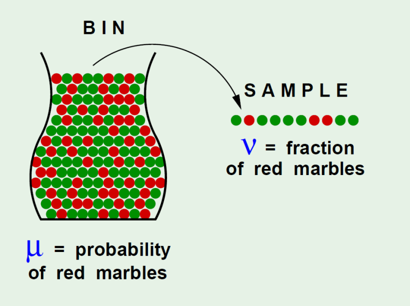
- Learning is feasible if $\boldsymbol{x} \sim p(\boldsymbol{x})$
- $p(\boldsymbol{x})$ says: Training and testing are related
- If training and testing are unrelated, then hopeless – the deterministic example shown previously
- If you draw training and testing samples with different bias, then you will suffer

When Will $E_{\text{in}} = E_{\text{out}}$?

Theorem (Hoeffding Inequality)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables such that $0 \leq X_n \leq 1$ and $\mathbb{E}[X_n] = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{n=1}^{N} X_n - \mu\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}. \tag{3}$$

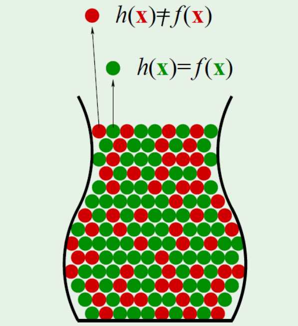


When Will $E_{\text{in}} = E_{\text{out}}$?

- To us, the inequality can be stated as

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}.$$

- N = number of training samples
- ϵ = tolerance level
- Hoeffding is applicable because $[\![h(\boldsymbol{x}) \neq f(\boldsymbol{x})]\!]$ is either 1 or 0.



Appendix