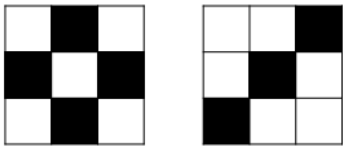
CSE 417T: Introduction to Machine Learning

Lecture 2: Generalization

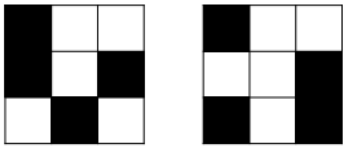
Henry Chai

08/30/18

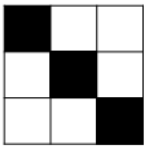
Puzzle



$f(x) = +1$

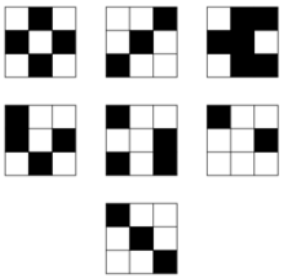


$f(x) = -1$



$f(x) = ???$

An Answer

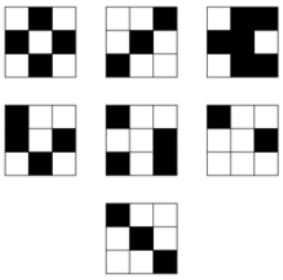


$$h(x) = \begin{cases} +1 \text{ if symmetric} \\ -1 \text{ otherwise} \end{cases}$$

$$\downarrow$$

$$h\left(\right) = +1$$

An Answer

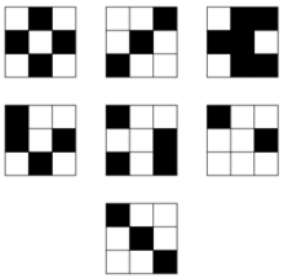


$$h(x) = \begin{cases} +1 \text{ if top left is white} \\ -1 \text{ otherwise} \end{cases}$$

$\downarrow$

$h($  $) = -1$

An Answer

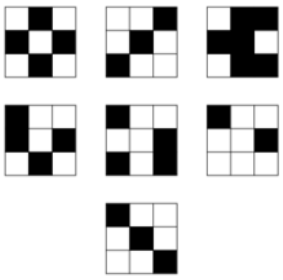


$$h(x) = \begin{cases} +1 \text{ if } x \in \{ \quad \} \\ -1 \text{ otherwise} \end{cases}$$

$$\downarrow$$

$$h\left(\quad \right) = -1$$

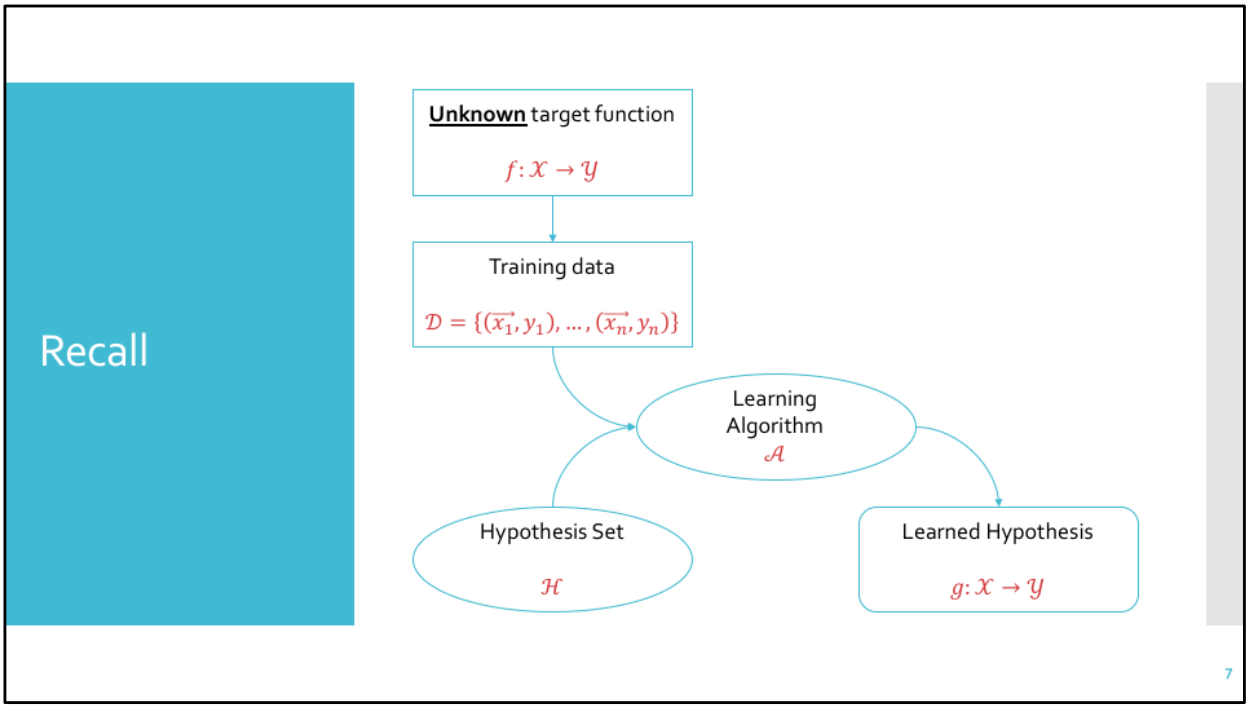
An Answer



$$h(x) = \begin{cases} +1 \text{ if } x \in \{ \text{ } \} \\ -1 \text{ otherwise} \end{cases}$$

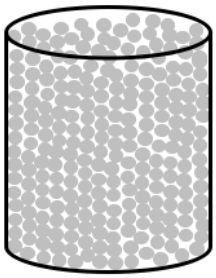
$$\downarrow$$

$$h\left(\text{ } \right) = +1$$



Unless your training data spans the input space (an impossibility for infinite input spaces), one can never say how well a hypothesis will do on inputs not in the training data with absolute certainty

Analogy



μ = fraction of red marbles in bin

n i. i. d. samples

ν = fraction of red marbles in sample

Does $\nu = \mu$?

Does ν say anything about μ?

Samples are drawn independently from the same probability distribution

Hoeffding's Inequality

- μ = fraction of red marbles in bin
- ν = fraction of red marbles in a sample of size n

- $P\{|\nu - \mu| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$

- As n increases, RHS decreases
- As ϵ decreases, RHS increases

The probability is w.r.t. the distribution over the input space. This statement is true for all n and ϵ; note that the right hand side (RHS) does not depend on μ

Connection to Learning

 = input space ($\mathcal{X}$)

● = point in the input space ($\vec{x}$)

●●●●●●●●●● = training data ($\mathcal{D}$)

● = a point classified correctly by a specified hypothesis h

● = a point classified incorrectly by a specified hypothesis h

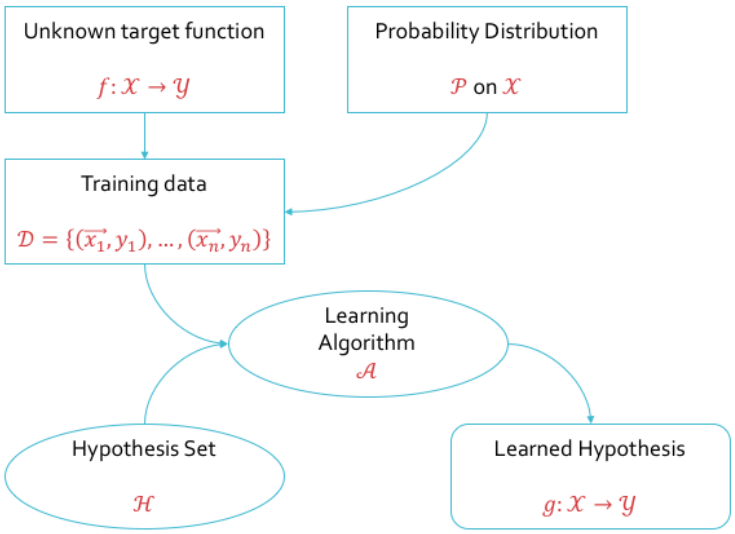
ν = fraction of training data classified incorrectly by h ($E_{in}(h)$)

μ = fraction of points in the input space classified incorrectly by h ($E_{out}(h)$)

$$P\{|E_{in}(h) - E_{out}(h)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$$

If h does well on the training set, then there's a pretty good chance that h generalizes well i.e. h will do well on points not in the training set. Or rather, it says that the probability that h's performance on points not in the training set and h's performance on points in the training set are different is low.

Formal Setup



11

Validation

= input space ($\mathcal{X}$)

= point in the input space ($\vec{x}$)

= training data ($\mathcal{D}$)

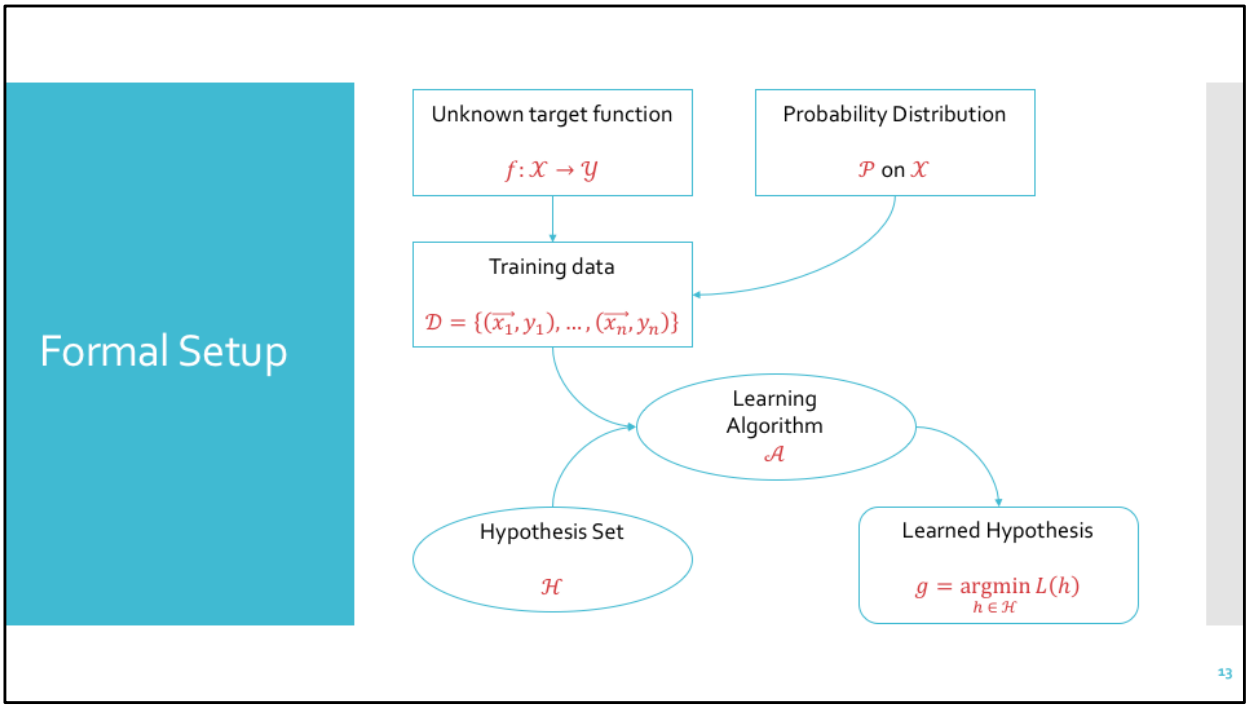
= a point classified correctly **by a specified hypothesis** h

= a point classified incorrectly **by a specified hypothesis** h

ν = fraction of training data classified incorrectly **by** h ($E_{in}(h)$)

μ = fraction of all possible data classified incorrectly **by** h ($E_{out}(h)$)

$$P\{|E_{in}(h) - E_{out}(h)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$$



The problem is that g isn't some random hypothesis in $\mathcal{H}$ but rather it's a cherry-picked hypothesis, one chosen according to some very specific methodology.

Another Analogy

- If you toss a fair coin 20 times, the probability that it comes up heads 20 times is $2^{-20} \approx 1e\text{-}6$
- If you toss 2^{20} fair coins 20 times each, the probability that at least one coin comes up heads 20 times is $\approx 1 - \frac{1}{e} \approx 0.63$

Another Analogy

- Take any coin that came up all heads and apply Hoeffding's inequality

- $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$
- $P\{|\text{Fraction of tails in 20 trials} - \text{Fraction of tails in } \infty \text{ trials}| > \epsilon\} \leq 2e^{-40\epsilon^2}$
- $P\{|0 - P\{\text{this coin coming up tails}\}| > \epsilon\} \leq 2e^{-40\epsilon^2}$
- $P\left\{P\{\text{this coin coming up tails}\} > \frac{1}{4}\right\} \leq 2e^{-2.5} \approx 0.15$

The probability that this coin is fair is pretty low, almost zero. But we started from the premise that all the coins were fair and we know that really what's happening is if you flip enough coins, then sooner or later you're going to get lucky and find a coin that comes up all heads. But that doesn't mean we should suspect that that coin will always come up heads

Hoeffding's Inequality (Corrected)

- Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, \dots, h_m\}$

$$P\{|E_{in}(g) - E_{out}(g)| > \epsilon\}$$

$$\leq P\left\{\bigcup_{j=1}^{m} |E_{in}(h_j) - E_{out}(h_j)| > \epsilon\right\}$$

$$\leq \sum_{j=1}^{m} P\{|E_{in}(h_j) - E_{out}(h_j)| > \epsilon\}$$

$$\leq \sum_{j=1}^{m} 2e^{-2\epsilon^2 n} = 2(m)e^{-2\epsilon^2 n}$$

Hoeffding's Inequality (Corrected)

- Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, \dots, h_m\}$
- $E_{in}(g)$ = in-sample error of best hypothesis in $\mathcal{H}$
- $E_{out}(g)$ = out-of-sample error of best hypothesis in $\mathcal{H}$

- $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2(m)e^{-2\epsilon^2 n}$

- As n increases, RHS decreases
- As ϵ decreases, RHS increases
- As m increases, RHS increases