ECE595 / STAT598: Machine Learning I

Lecture 22.3: Training versus Testing

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

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

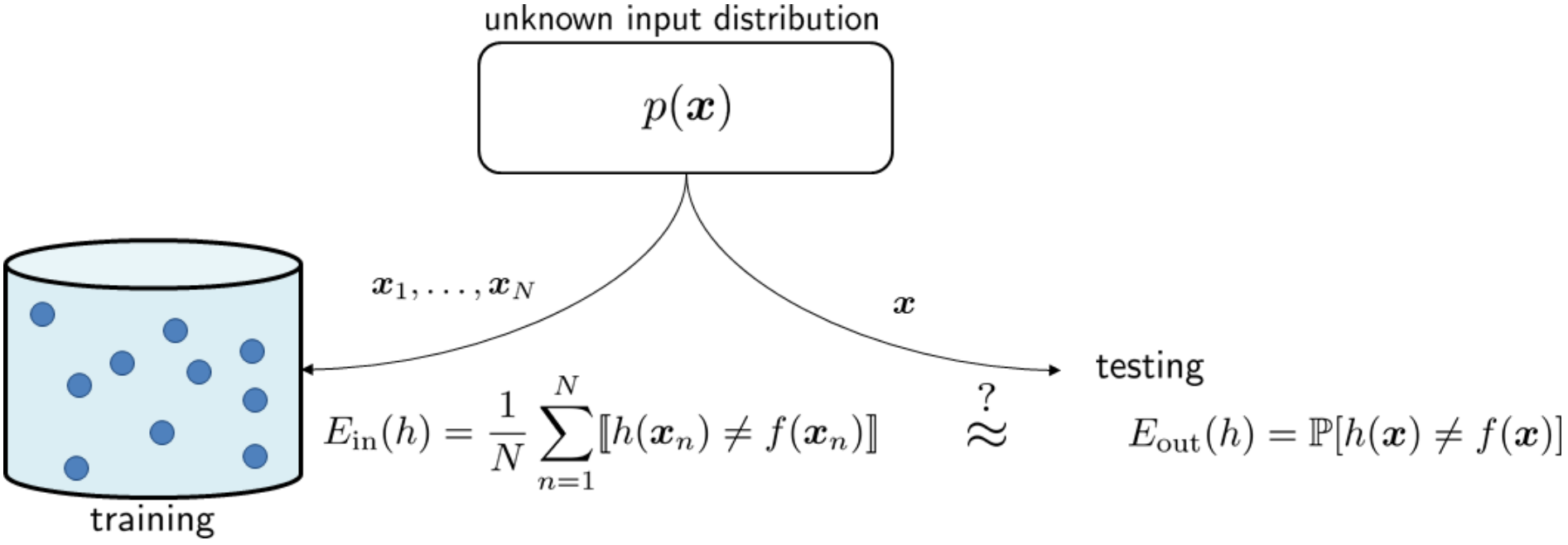
In-Sample Error

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

Out-Sample Error

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

The Role of $p(x)$



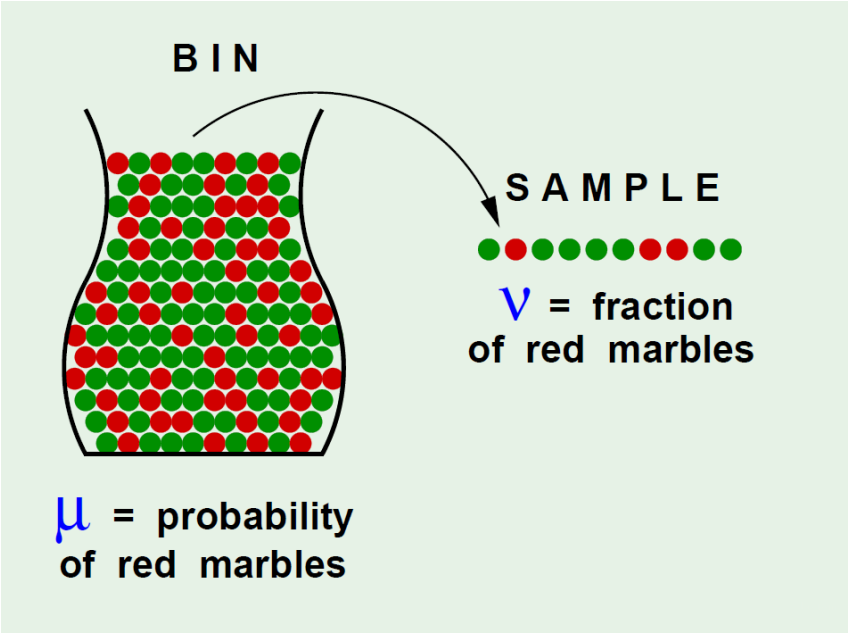
- Learning is feasible if $\boldsymbol{x} \sim p(\boldsymbol{x})$
- $p(\boldsymbol{x})$ says: Training and testing are related
- If training and testing are unrelated, then hopeless – the deterministic example shown previously
- If you draw training and testing samples with different bias, then you will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