ECE595 / STAT598: Machine Learning I

Lecture 32.1: Validation

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

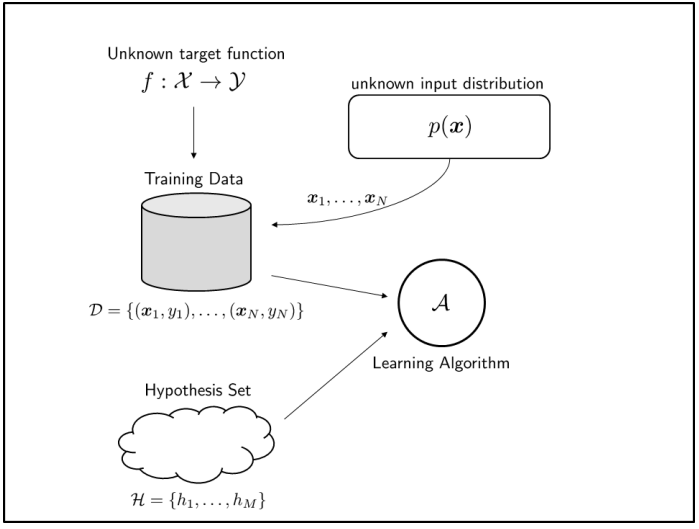
Outline

- Lecture 30 Overfit
- Lecture 31 Regularization
- Lecture 32 Validation

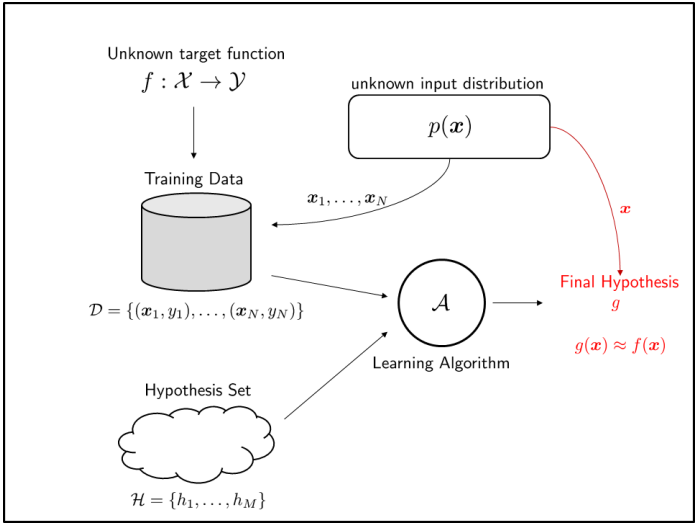
Today's Lecture:

- Validation
  - Concept of validation
  - Properties of validation error
- Model Selection
  - Basic idea
  - Case study
- Validation in Regularization
  - Cross validation
  - Parameter selection

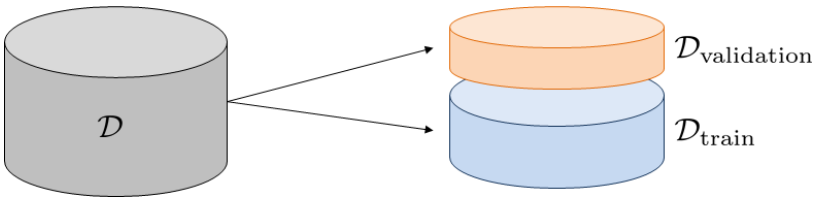
Evaluating Your Model



Evaluating Your Model



Validation Set



- What does $\mathcal{D}_{\text{val}}$ buy you?
- Generalization bound using $\mathcal{D}_{\text{val}}$?
- How to use $\mathcal{D}_{\text{val}}$?
- Validation vs Cheating
- Cross Validation

The Role of Validation

- Recall the generalization error:

$$E_{\text{out}}(h) = E_{\text{in}}(h) + \underbrace{\text{overfitpenalty}}_{\text{regularization suppresses this term}}$$

- How about validation?

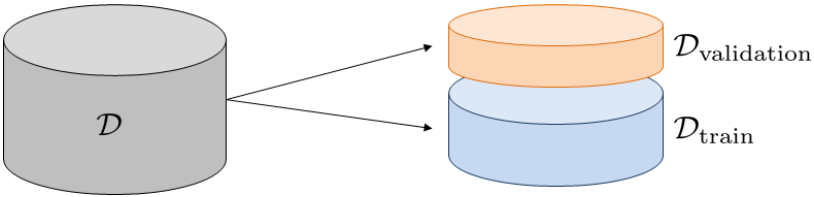
$$\underbrace{E_{\text{out}}(h)}_{\text{validation estimates this term}} = E_{\text{in}}(h) + \text{overfitpenalty}$$

- Is it the same as testing?

$$\underbrace{E_{\text{out}}(h)}_{\text{testing estimates this term}} = E_{\text{in}}(h) + \text{overfitpenalty}$$

- Testing: You cannot use testing set at any stage of training.
- Validation: You can use validation to make choices during training.

Creating the Validation Set



- Data set: $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$. N samples.
- Validation set: $\mathcal{D}_{\text{val}}$. K samples.
- Training set: $\mathcal{D}_{\text{training}}$. $N - K$ samples.
- If you run the learning algorithm on $\mathcal{D}_{\text{train}}$, you obtain

$$g^- \in \mathcal{H}$$

- g^-: a hypothesis learned by "subtracting" some samples
- g^- is not necessarily the final hypothesis you eventually report

What does validation tell us?

Goal: Define the validation error $E_{\text{val}}(g^-)$, and analyze its statistical properties.

- The **validation error** is

$$E_{\text{val}}(g^-) = \frac{1}{K} \sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \text{e}(g^-(\boldsymbol{x}_n), y_n)$$

- Average error over the *validation set*. $\text{e}(g^-(\boldsymbol{x}_n), y_n)$: Point-wise error.
- Classification:

$$\text{e}(g^-(\boldsymbol{x}), y) = [\![g^-(\boldsymbol{x}) \neq y]\!]$$

- Regression:

$$\text{e}(g^-(\boldsymbol{x}), y) = (g^-(\boldsymbol{x}) - y)^2$$

- Want to analyze the **mean** and **variance** of $E_{\text{val}}(g^-)$.

Property 1: Mean of $E_{\text{val}}(g^-)$

- Let us analyze the mean of $E_{\text{val}}(g^-)$.
- We want to show that the validation error $E_{\text{val}}(g^-)$ is an **unbiased estimate** of E_{out}
- That is, the expectation of $E_{\text{val}}(g^-)$ is E_{out}
- Here is why:

$$\mathbb{E}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)] = \mathbb{E}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \text{e}(g^-(\boldsymbol{x}_n), y_n)\right] \qquad \text{(definition)}$$

$$= \frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \mathbb{E}_{\mathcal{D}_{\text{val}}}\left[\text{e}(g^-(\boldsymbol{x}_n), y_n)\right] \qquad \text{(linearity)}$$

$$= \frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \mathbb{E}_{\boldsymbol{x}_n}\left[\text{e}(g^-(\boldsymbol{x}_n), y_n)\right] \qquad \mathcal{D}_{\text{val}} = (\boldsymbol{x}_n, f(\boldsymbol{x}_n))$$

$$= \frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} E_{out}(g^-) = E_{out}(g^-) \qquad \boldsymbol{x}_n \sim p(\boldsymbol{x})$$

Property 2: Variance of $E_{\text{val}}(g^-)$

- Define $\sigma^2_{\text{val}} = \text{Var}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)]$.
- How does σ^2_{val} depend on K?
- Let's do some calculation

$$\begin{aligned}
\sigma^2_{\text{val}} &= \text{Var}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \text{e}(g^-(\boldsymbol{x}_n), y_n)\right] && \text{(definition)} \\
&= \frac{1}{K^2}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \underbrace{\text{Var}_{\mathcal{D}_{\text{val}}}\left[\text{e}(g^-(\boldsymbol{x}_n), y_n)\right]}_{\stackrel{\text{def}}{=}\sigma^2(g^-)} && \text{(independence)} \\
&= \frac{1}{K^2}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \sigma^2(g^-) \\
&= \frac{1}{K}\sigma^2(g^-).
\end{aligned}$$

Property 2: Variance of $E_{\text{val}}(g^-)$

- If we consider a classification problem so that $\text{e}(g^-(\boldsymbol{x}), y) = [\![g^-(\boldsymbol{x}) \neq y]\!]$
- Then

$$\begin{aligned}
\sigma^2_{\text{val}} = \frac{1}{K}\sigma^2(g^-) &= \frac{1}{K}\text{Var}_{\mathcal{D}_{\text{val}}}\left[\text{e}(g^-(\boldsymbol{x}), y)\right] && \text{(definition)} \\
&= \frac{1}{K}\text{Var}_{\mathcal{D}_{\text{val}}}\left[[\![g^-(\boldsymbol{x}) \neq y]\!]\right] && \text{(classification)} \\
&= \frac{1}{K}\mathbb{P}[g^-(\boldsymbol{x}) \neq y](1 - \mathbb{P}[g^-(\boldsymbol{x}) \neq y]) && \text{(Bernoulli)}.
\end{aligned}$$

- Remark: If X is Bernoulli, then $\text{Var}[X] = p(1-p) \leq \frac{1}{4}$.
- Therefore, we can bound σ^2_{val} using

$$\sigma^2_{\text{val}} \leq \frac{1}{4K}.$$

- So as $K \to \infty$, $\sigma^2_{\text{val}} \to 0$.

Does $E_{\text{val}}(g^-)$ Generalize?

- $E_{\text{val}}(g^-)$ is a **random variable**. So it fluctuates.
- Mean: $\mathbb{E}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)]$.
- Variance: $\text{Var}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)]$.
- Previous slide: $\mathbb{E}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)] = E_{\text{out}}(g^-)$.
- So we should expect Hoeffding inequality to apply:

$$E_{\text{out}}(g^-) \leq E_{\text{val}}(g^-) + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

- Why? Recall Hoeffding inequality for *one* hypothesis:

$$E_{\text{out}}(h) \leq E_{\text{in}}(h) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

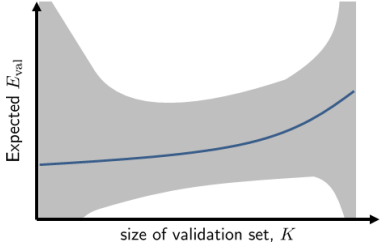
- So as K grows, $E_{\text{val}}(g^-)$ actually generalizes $E_{\text{out}}(g^-)$ very well.

Large K or Small K?

- No matter how you look at the result: Generalization bound or variance bound

$$\sigma_{\text{val}}^2 \leq \frac{1}{4K}.$$

- If $K \to \infty$, then $\sigma_{\text{val}}^2 \to 0$
- So large K is good.
- But can K be really really large?
- No. K for validation, $N - K$ for training.



Re-Using K

- Is it a waste if we can only use $N - K$ samples for training?
- No. You are *allowed* to reuse the K samples
- Use $\mathcal{D}_{\text{val}}$ to give an estimate of $E_{\text{val}}(g^-)$
- Use $E_{\text{val}}(g^-)$ as a guide to choose g
- Here is a pictorial illustration
- Rule of Thumb: $K = \frac{N}{5}$