ECE595 / STAT598: Machine Learning I
Lecture 32 Validation

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

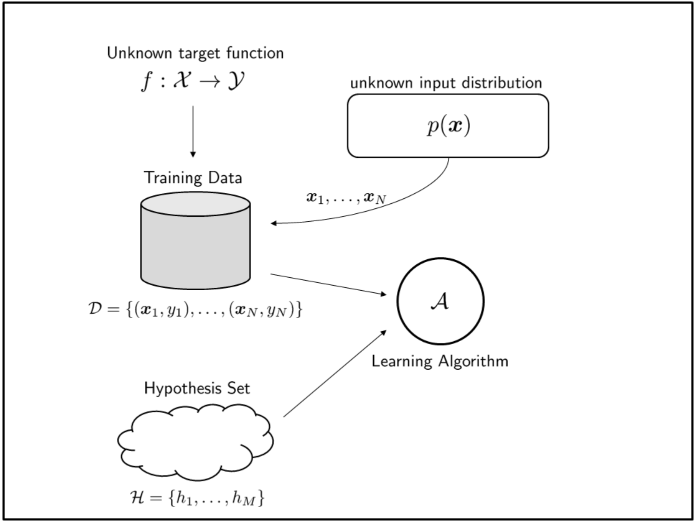
Outline

- Lecture 31 Overfit
- Lecture 32 Regularization
- Lecture 33 Validation

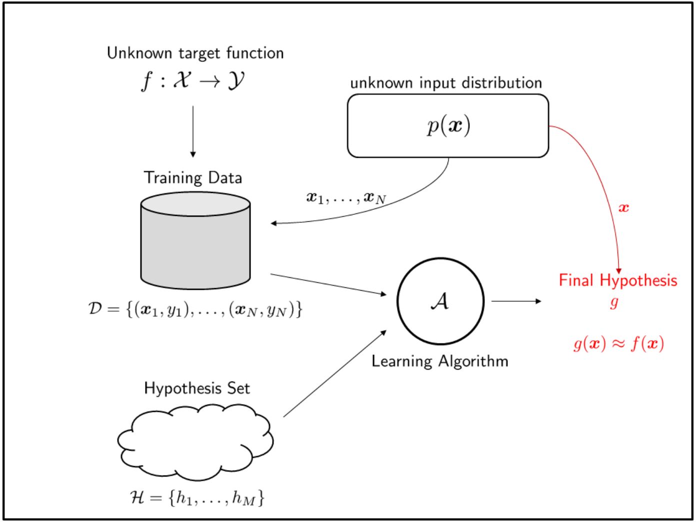
Today's Lecture:

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

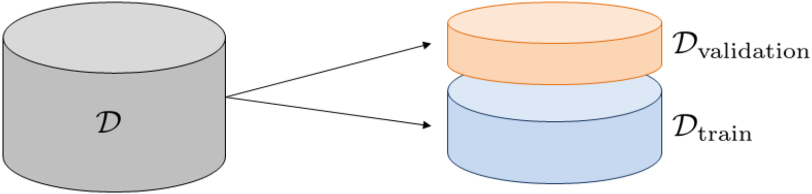
Evaluating Your Model



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:

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

- How about validation?

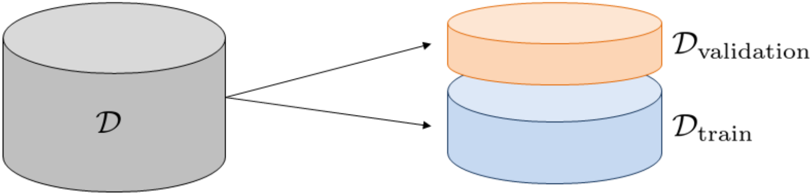
$$\underbrace{E_{\text{out}}(h)}_{\text{validation estimates this term}} = E_{\text{in}}(h) + \text{overfit penalty}$$

- Is it the same as testing?

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

- 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

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

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

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

- Regression:

$$\mathrm{e}(g^-(\mathbf{x}), y) = (g^-(\mathbf{x}) - y)^2$$

- 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:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)] &= \mathbb{E}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \mathrm{e}(g^-(\boldsymbol{x}_n), y_n)\right] && \text{(definition)} \\
&= \frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \mathbb{E}_{\mathcal{D}_{\text{val}}}\left[\mathrm{e}(g^-(\boldsymbol{x}_n), y_n)\right] && \text{(linearity)} \\
&= \frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} \mathbb{E}_{\boldsymbol{x}_n}\left[\mathrm{e}(g^-(\boldsymbol{x}_n), y_n)\right] && \mathcal{D}_{\text{val}} = (\boldsymbol{x}_n, f(\boldsymbol{x}_n)) \\
&= \frac{1}{K}\sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{val}}} E_{out}(g^-) = E_{out}(g^-) && \boldsymbol{x}_n \sim p(\boldsymbol{x})
\end{aligned}
$$

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

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

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

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

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

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

- Remark: If X is Bernoulli, then $\mathrm{Var}[X] = p(1-p) \leq \frac{1}{4}$.
- 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:

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

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

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

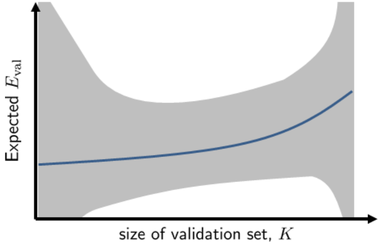
- So as K grows, $E_{\text{val}}(g^-)$ actually generalizes $E_{\text{out}}(g^-)$ very well.

Large K or Small K?

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

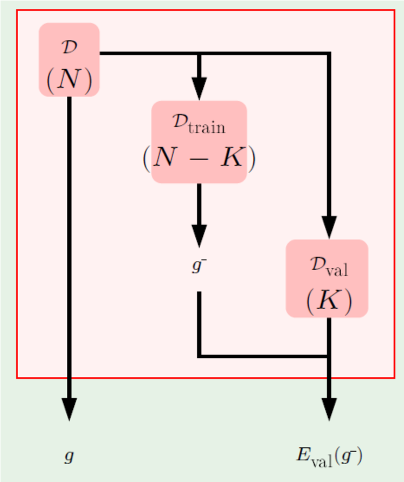
$$\sigma_{\text{val}}^2 \le \frac{1}{4K}.$$

- 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: $K = \frac{N}{5}$



Outline

- Lecture 31 Overfit
- Lecture 32 Regularization
- Lecture 33 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

Validation for Model Selection

- Consider a set of M models: $\mathcal{H}_1, \ldots, \mathcal{H}_M$
- E.g., linear / quadratic / logistic, etc
- E.g., linear model with different regularization parameters, etc
- How to choose the model?
- Use $\mathcal{D}_{\text{train}}$ to train $g_1^-, \ldots, g_M^-$.
- Evaluate

$$E_m = E_{\text{val}}(g_m^-),$$

  for $m = 1, \ldots, M$.
- E_m is an **unbiased estimate** of the out-sample error $E_{\text{out}}(g_m^-)$.
- Select the one with the minimum validation error:

$$m^* = \underset{m}{\text{argmin}} \quad E_m$$

- The model $\mathcal{H}_{m^*}$ is the best model

Generalization Bound for Model Selection

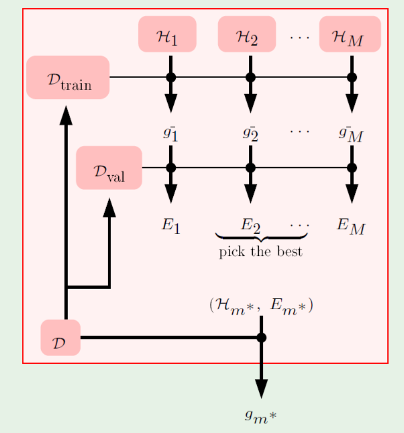
- If you choose $g^{-}_{m^*}$ from $g^{-}_1, \ldots, g^{-}_M$
- You are effectively considering

$$\mathcal{H}_{\text{val}} = \{g^{-}_1, \ldots, g^{-}_M\}.$$

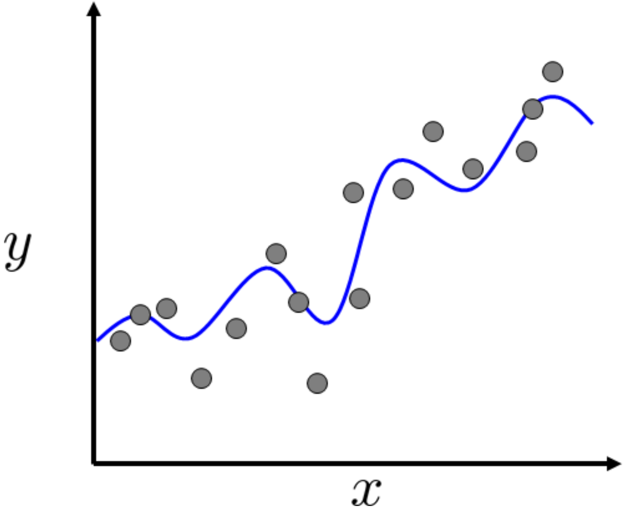
- So the price you need to pay in the generalization bound is

$$E_{\text{out}}(g^{-}_{m^*}) \leq E_{\text{val}}(g^{-}_{m^*}) + \mathcal{O}\left(\sqrt{\frac{\log M}{K}}\right).$$

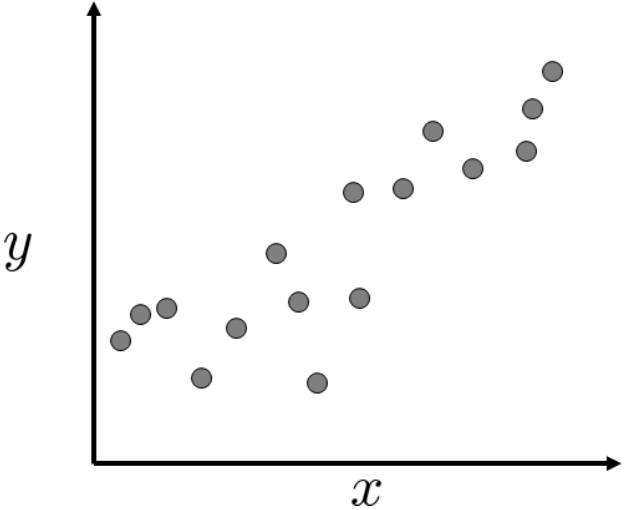
- Use $g^{-}_{m^*}$ as the final hypothesis?
- No. Should choose $\mathcal{H}_{m^*}$, and train with N samples.



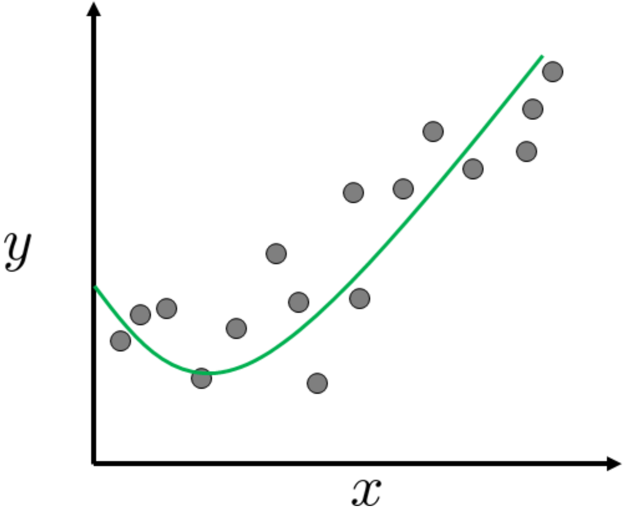
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



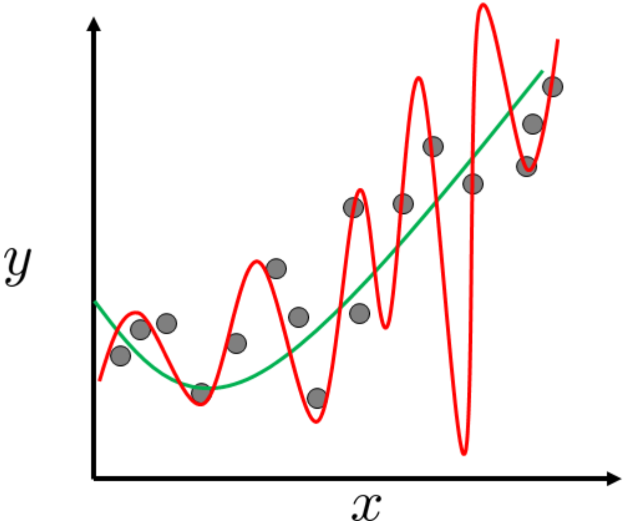
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



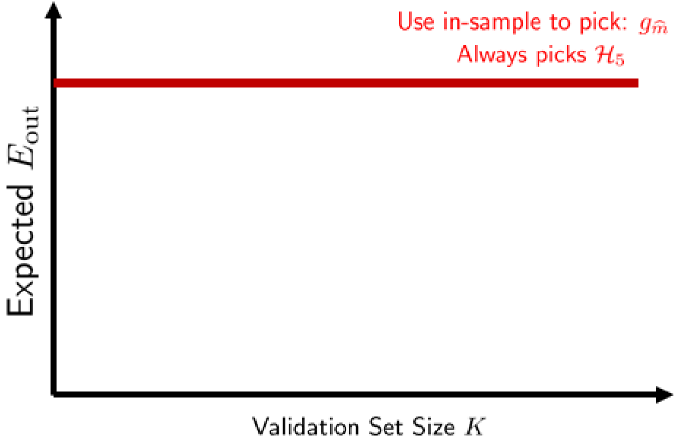
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



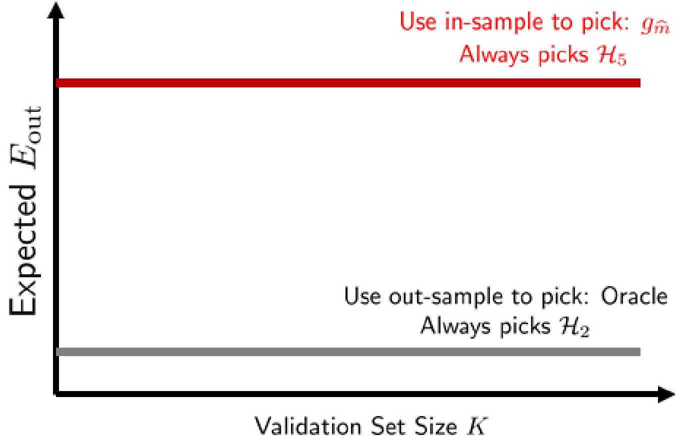
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



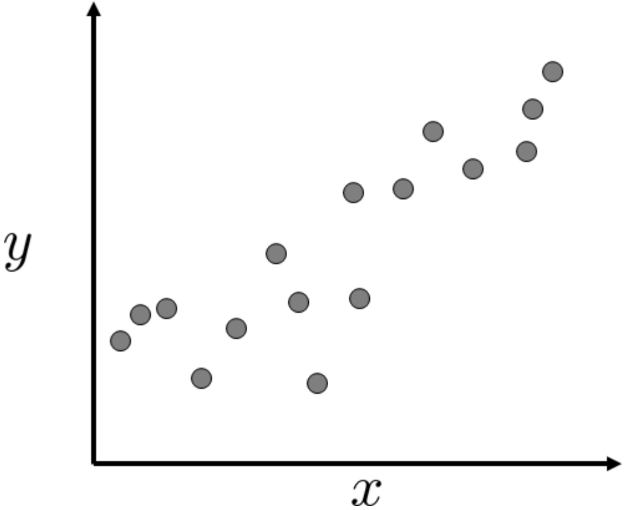
Expected Error



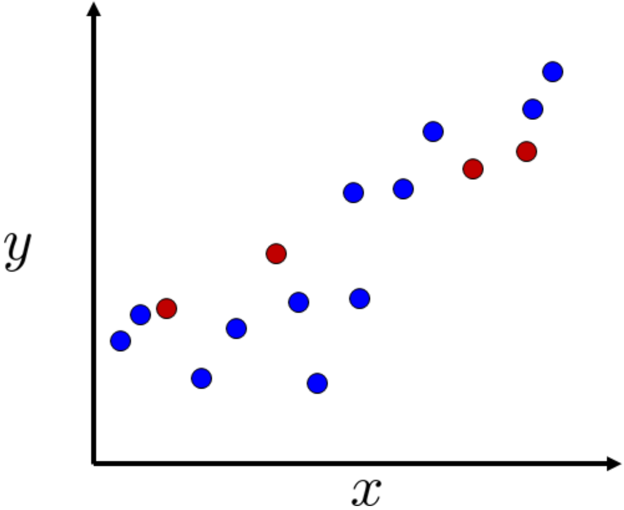
Expected Error



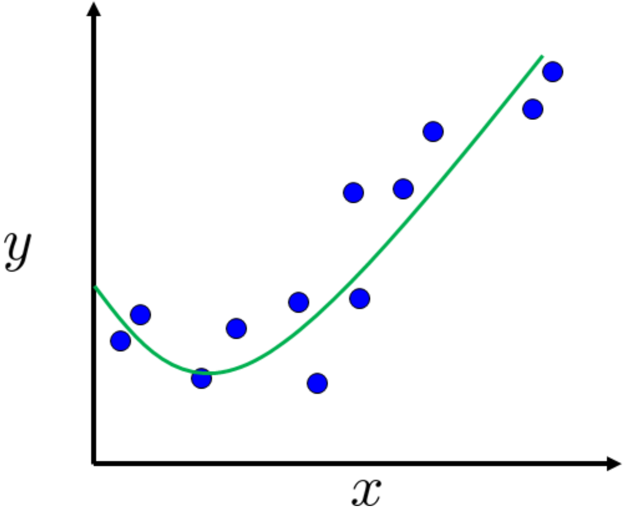
Validation



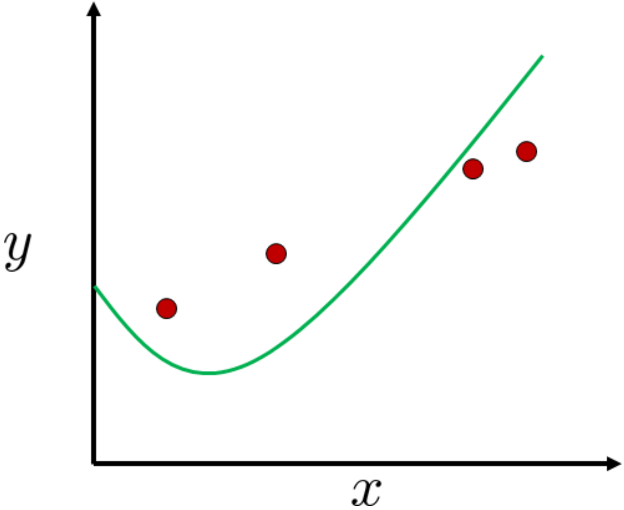
Validation



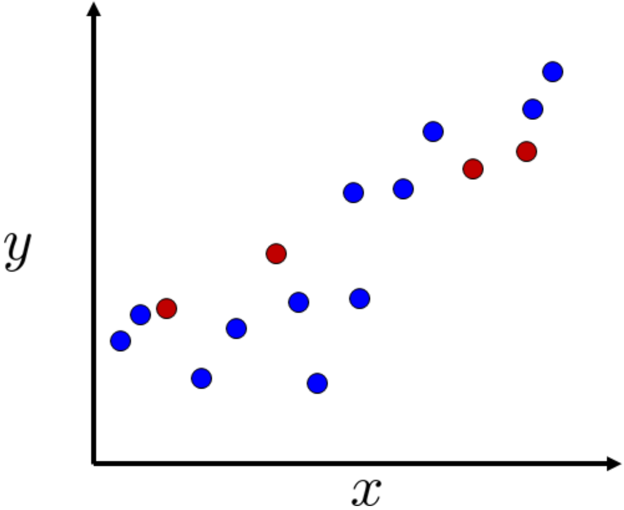
Validation



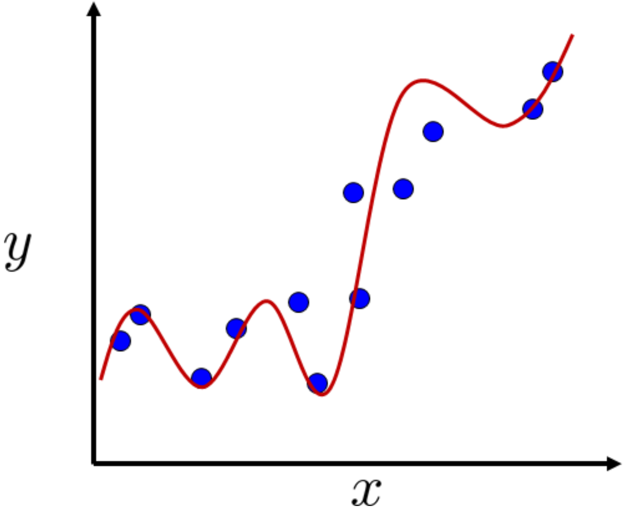
Validation



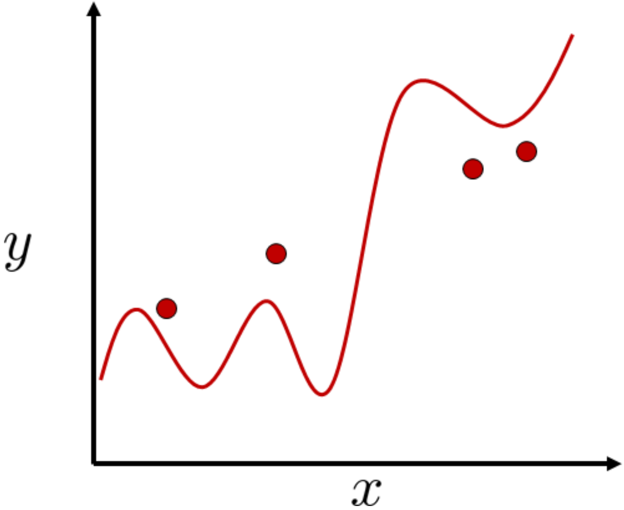
Validation



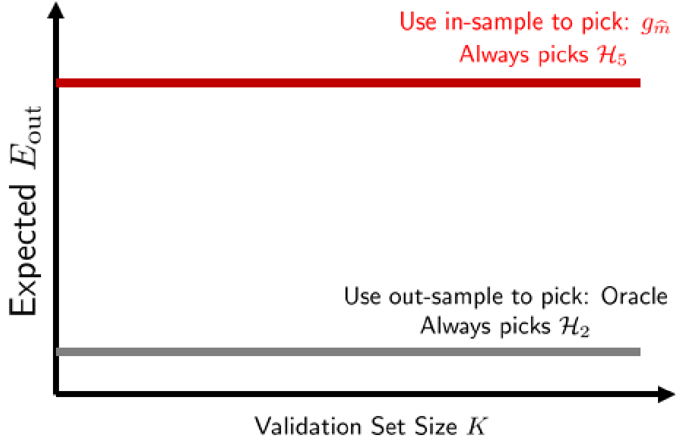
Validation



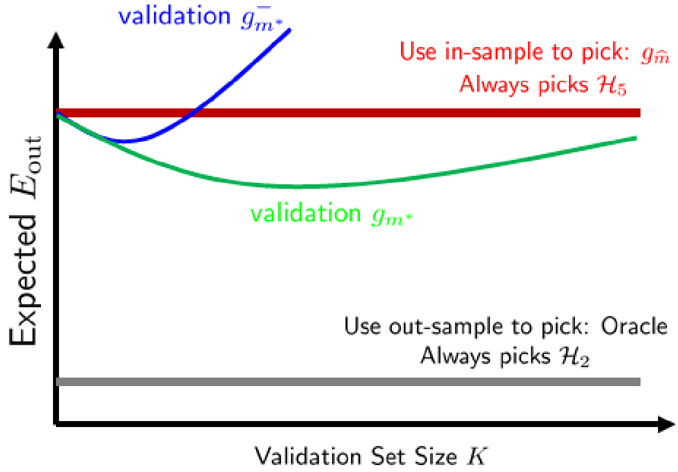
Validation



Expected Error



Expected Error



Observations

Validation and $N - K$ samples for training:

- $\mathbb{E}[E_{out}(g^-_{m^*})]$ drops and then rise.
- Compared to in-sample, $\mathbb{E}[E_{out}(g^-_{m^*})]$ uses a few samples to validate.
- This gives a good estimate of out-sample error.
- As K increases, the estimate improves. So $\mathbb{E}[E_{out}(g^-_{m^*})]$ drops.
- If K is too large, then only $N - K$ samples for training.
- Poor training makes $\mathbb{E}[E_{out}(g^-_{m^*})]$ rise.

Validation and N samples for training:

- $\mathbb{E}[E_{out}(g_{m^*})]$ will be lower.
- Because you have chosen the best.

Therefore, you should always recycle the validation data for training the final hypothesis.

Outline

- Lecture 31 Overfit
- Lecture 32 Regularization
- Lecture 33 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

Cross Validation

- A principled way to estimate the out-sample error, without suffering from small K problem.
- Consider the **leave-one-out** approach.
- Let the data set be

$$\mathcal{D}_n = (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{n-1}, y_{n-1}), \cancel{(\boldsymbol{x}_n, y_n)}, (\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_N, y_N)$$

- Remove the n-th training sample
- Learn the hypothesis function

$$g_n^- = \text{learn from } \mathcal{D}_n.$$

- Let error

$$\mathrm{e}_n \overset{\text{def}}{=} E_{\text{val}}(g_n^-) = \mathrm{e}(g_n^-(\boldsymbol{x}_n), y_n).$$

- Remark: e_n is based on a single data point $(\boldsymbol{x}_n, y_n)$.

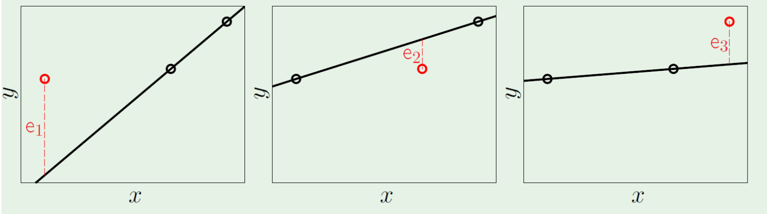
Cross Validation

- This will give you

$$e_1, e_2, \ldots, e_N$$

- Let's compute the average

$$E_{\text{cv}} = \frac{1}{N}\sum_{n=1}^{N} e_n.$$



- Validation: Use K samples to validate
- Cross-Validation: Recycle the N samples to validate

Cross-Validation for Linear Regression

- Recall the linear regression model:

$$\boldsymbol{w}^* = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

- How to estimate the optimal λ?
- Let

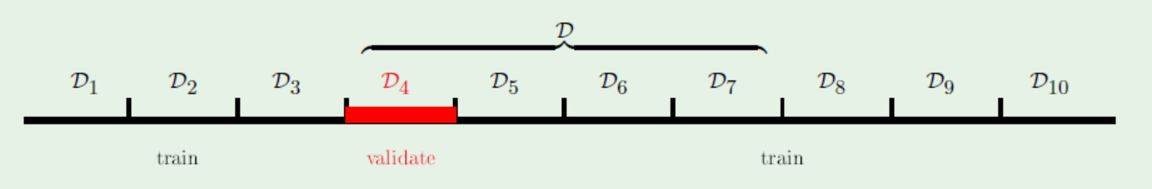
$$\begin{aligned}\boldsymbol{H}(\lambda) &= \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T \\ \widehat{\boldsymbol{y}} &= \boldsymbol{H}\boldsymbol{y}\end{aligned}$$

- Compute the cross validation score:

$$E_{\mathrm{cv}} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\widehat{y}_n - y_n}{1 - H_{n,n}(\lambda)}\right)^2$$

- $H_{n,n}(\lambda) = \boldsymbol{x}_n^T(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{x}_n$. (See textbook Problem 4.26.)
- Pick λ that minimizes E_{cv}

V-fold validation



- Leave one out: N training sessions. Each session has $N - 1$ points.
- In practice: Partition the dataset into V sessions.
- Each session has N/V points.
- Train using $\mathcal{D} \backslash \mathcal{D}_V$.
- Test using $\mathcal{D}_V$.
- Rule of Thumb: $V = 10$. 10-fold cross-validation.

Summary

- Validation says: Break the dataset into testing and validation.
- Use validation set to help selecting models and parameters.
- Then reuse the data to report the final hypothesis.
- Can also use cross-validation to get a better estimate of E_{out}.
- Never use testing data for validation.

Reading List

- Yaser Abu-Mustafa, Learning from Data, Chapter 4.3

Appendix

Unbiasedness of E_{cv}

- Why care? If yes, then we can use E_{cv} to estimate E_{out}
- Recall $g^{(\mathcal{D})}$. The out-sample error for $g^{(\mathcal{D})}$ is

$$E_{\mathrm{out}}(N) = \mathbb{E}_{\mathcal{D}}\left[E_{\mathrm{out}}(g^{(\mathcal{D})})\right].$$

- $E_{\mathrm{out}}(N)$: Overall out-sample error average over all possible training sets
- $E_{\mathrm{out}}(N)$: Function of N. If you have more training samples, then you have lower error
- We can show that

$$\begin{aligned} E_{\mathrm{out}}(N) &\stackrel{?}{=} \mathbb{E}_{\mathcal{D}}[E_{\mathrm{cv}}] \\ &= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\sum_{n=1}^{N} \mathrm{e}_n\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathcal{D}}\left[\mathrm{e}_n\right] = \mathbb{E}_{\mathcal{D}}\left[\mathrm{e}_n\right]. \end{aligned}$$

Unbiasedness of E_{cv}

- So what is $\mathbb{E}_{\mathcal{D}}[\mathrm{e}_n]$?

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\mathrm{e}_n] &= \mathbb{E}_{\mathcal{D}_n,(\boldsymbol{x}_n,y_n)}[\mathrm{e}_n] && \color{red}{\text{decouple } \mathcal{D}} \\
&= \mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{(\boldsymbol{x}_n,y_n)}[\mathrm{e}(g_n^-(\boldsymbol{x}_n), y_n)] \\
&= \mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{(\boldsymbol{x}_n,y_n)}[E_{\mathrm{val}}(g_n^-)] \\
&= \mathbb{E}_{\mathcal{D}_n} E_{\mathrm{out}}(g_n^-) && \color{red}{\text{unbisedness of } E_{\mathrm{val}}} \\
&= E_{\mathrm{out}}(N-1). && \color{red}{\text{expectation of } \mathcal{D}_n}
\end{aligned}
$$

- So,

$$
\mathbb{E}_{\mathcal{D}}[E_{\mathrm{cv}}] = E_{\mathrm{out}}(N-1).
$$

- That means: E_{cv} is an unbiased estimate of $E_{\mathrm{out}}(N-1)$
- Remark: This gives us the mean of E_{cv}. The variance is a lot harder because $\mathcal{D}_m$ and $\mathcal{D}_n$ overlaps.