ECE595 / STAT598: Machine Learning I

Lecture 32.3: Validation - Validation in Regularization

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

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

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

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

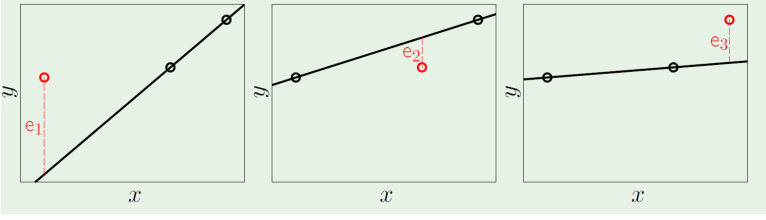
Cross Validation

- This will give you

$$\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_N$$

- Let's compute the average

$$E_{\mathrm{cv}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{e}_n.$$



- 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

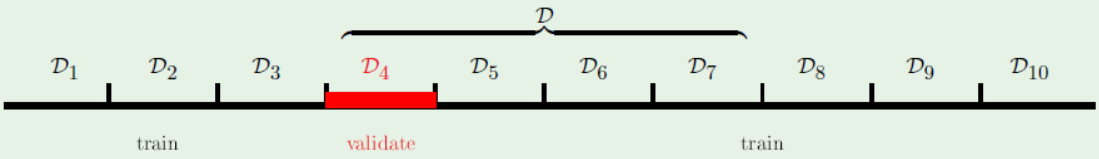
$$\boldsymbol{H}(\lambda) = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T$$
$$\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$$

- Compute the cross validation score:

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

- $H_{n,n}(\lambda) = \boldsymbol{x}_n^T(\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{x}_n$. (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

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

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

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

Unbiasedness of E_{cv}

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

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

- So,

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

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