Learning From Data
Lecture 13
Validation and Model Selection

The Validation Set
Model Selection
Cross Validation

M. Magdon-Ismail
CSCI 4100/6100
RECAP: Regularization

Regularization combats the effects of noise by putting a leash on the algorithm.

\[ E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N} \Omega(h) \]

\( \Omega(h) \rightarrow \text{smooth, simple } h \)
— noise is rough, complex.

Different regularizers give different results
— can choose \( \lambda \), the amount of regularization.

Overfitting \( \rightarrow \) \( \rightarrow \) Underfitting

Optimal \( \lambda \) balances approximation and generalization, bias and variance.
Validation: A Sneak Peek at $E_{\text{out}}$

\[ E_{\text{out}}(g) = E_{\text{in}}(g) + \text{overfit penalty} \]

VC bounds this using a complexity error bar for $\mathcal{H}$
regularization estimates this through a heuristic complexity penalty for $g$

Validation goes directly for the jugular:

\[ E_{\text{out}}(g) = E_{\text{in}}(g) + \text{overfit penalty} \]

validation estimates this directly

In-sample estimate of $E_{\text{out}}$ is the Holy Grail of learning from data.
The Test Set

$\mathcal{D}$ (N data points) $\xrightarrow{\text{ } } \mathcal{D}_{\text{test}}$ (K test points)

$g \xrightarrow{\text{ } } e_k = e(g(x_k), y_k) \xrightarrow{\text{ } } e_1, e_2, \ldots, e_K \xrightarrow{\text{ } } g_{\text{test}} = \frac{1}{K} \sum_{k=1}^{K} e_k \xrightarrow{\text{ } } E_{\text{out}}(g)$

$E_{\text{test}}$ is an estimate for $E_{\text{out}}(g)$

$\mathbb{E}_{\mathcal{D}_{\text{test}}} [e_k] = E_{\text{out}}(g)$

$\mathbb{E}[E_{\text{test}}] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[e_k] = \frac{1}{K} \sum_{k=1}^{K} E_{\text{out}}(g) = E_{\text{out}}(g)$

$e_1, \ldots, e_K$ are independent

$\text{Var}[E_{\text{test}}] = \frac{1}{K^2} \sum_{k=1}^{K} \text{Var}[e_k] = \frac{1}{K} \text{Var}[e]$

decreases like $\frac{1}{K}$

bigger $K \implies$ more reliable $E_{\text{test}}$. 
The Validation Set

1. Remove $K$ points from $\mathcal{D}$

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}.$$ 

$\mathcal{D}$ (N data points) → $\mathcal{D}_{\text{train}}$ (N – $K$ training points) → $\mathcal{D}_{\text{val}}$ ($K$ validation points) → $g = (g(x_k), y_k)$ → $e_1, e_2, \ldots, e_K$ → $E_{\text{val}} = \frac{1}{K} \sum_{k=1}^{K} e_k$ → $E_{\text{out}}(g)$
The Validation Set

- \( \mathcal{D} \) (\( N \) data points)

\[ \mathcal{D} \xrightarrow{\text{Remove } K \text{ points}} \mathcal{D} \]

\[ \mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}. \]

1. Remove \( K \) points from \( \mathcal{D} \)

2. Learn using \( \mathcal{D}_{\text{train}} \rightarrow g \).

3. Test \( g \) on \( \mathcal{D}_{\text{val}} \rightarrow E_{\text{val}}. \)

4. Use error \( E_{\text{val}} \) to estimate \( E_{\text{out}}(g) \).

\[ E_{\text{val}} = \frac{1}{K} \sum_{k=1}^{K} e_k \]
The Validation Set

\[ D \]  
(N data points)

\[ D_{\text{train}} \]  
(N − K training points)

\[ D_{\text{val}} \]  
(K validation points)

\[ g \]  
\[ e_k = e(g^{-}(x_k), y_k) \rightarrow e_1, e_2, \ldots, e_K \]

\[ \mathbb{E}[E_{\text{test}}] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[e_k] = \frac{1}{K} \sum_{k=1}^{K} E_{\text{out}}(g^{-}) = E_{\text{out}}(g^{-}) \]

\[ e_1, \ldots, e_K \text{ are independent} \]

\[ Var[E_{\text{val}}] = \frac{1}{K^2} \sum_{k=1}^{K} Var[e_k] \]

\[ = \frac{1}{K} \text{Var}[e(g^{-})] \]

\[ \text{decreases like } \frac{1}{K} \]

\[ \text{depends on } g^{-}, \text{not } H \]

\[ \text{bigger } K \implies \text{more reliable } E_{\text{val}}? \]
Rule of thumb: $K^* = \frac{N}{5}$.
Restoring $\mathcal{D}$

**Primary goal:** output best hypothesis.

$g$ was trained on all the data.

**Secondary goal:** estimate $E_{out}(g)$.

$g^{-}$ is behind closed doors.

\[ E_{out}(g) \quad E_{out}(g^{-}) \]
\[ \downarrow \quad \downarrow \]
\[ E_{in}(g) \quad E_{val}(g^{-}) \]

which should we use?

CUSTOMER
$E_{\text{val}} \text{ Versus } E_{\text{in}}$

\[ E_{\text{out}}(g) \leq E_{\text{in}}(g) + O\left(\sqrt{\frac{d_{\text{vc}}}{N}} \log N\right) \]

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

$E_{\text{val}}(g)$ usually wins as an estimate for $E_{\text{out}}(g)$, especially when the learning curve is not steep.

Biased error bar depends on $\mathcal{H}$.

Unbiased error bar depends on $g^{-}$.

Learning curve is decreasing (a practical truth, not a theorem)
The most important use of validation

\[ \mathcal{H}_1 \rightarrow g_1 \]
\[ \mathcal{H}_2 \rightarrow g_2 \]
\[ \mathcal{H}_3 \rightarrow g_3 \]
\[ \cdots \]
\[ \mathcal{H}_M \rightarrow g_M \]
The most important use of validation

Validation Estimate for $\mathcal{H}_1, g_1$
The most important use of validation

\[
\mathcal{H}_1 \quad \overset{\mathcal{D}_{\text{train}}}{\longrightarrow} \quad \overset{\mathcal{G}_1}{\downarrow} \quad \overset{\mathcal{D}_{\text{val}}}{\longrightarrow} \quad E_1
\]
Compute Validation Estimates for All Models

The most important use of validation

\[ \mathcal{H}_1 \rightarrow \mathcal{H}_2 \rightarrow \mathcal{H}_3 \rightarrow \cdots \rightarrow \mathcal{H}_M \]

\[ \mathcal{D}_{\text{train}} \rightarrow \mathcal{G}_1 \rightarrow \mathcal{G}_2 \rightarrow \mathcal{G}_3 \rightarrow \cdots \rightarrow \mathcal{G}_M \]

\[ \mathcal{D}_{\text{val}} \rightarrow \mathcal{E}_1 \rightarrow \mathcal{E}_2 \rightarrow \mathcal{E}_3 \rightarrow \cdots \rightarrow \mathcal{E}_M \]
Pick The Best Model According to Validation Error

The most important use of validation

\[ \mathcal{H}_1 \xrightarrow{D_{\text{train}}} \mathcal{G}_1 \xrightarrow{} E_1 \]
\[ \mathcal{H}_2 \xrightarrow{} \mathcal{G}_2 \xrightarrow{} E_2 \]
\[ \mathcal{H}_3 \xrightarrow{} \mathcal{G}_3 \xrightarrow{} E_3 \]
\[ \cdots \]
\[ \mathcal{H}_M \xrightarrow{} \mathcal{G}_M \xrightarrow{} E_M \]
\( E_{\text{val}}(g_{m^*}) \) is not Unbiased For \( E_{\text{out}}(g_{m^*}) \)

\[ E_{\text{out}}(g_{m^*}) \leq E_{\text{val}}(g_{m^*}) + O\left(\sqrt{\frac{\ln M}{K}}\right) \]

\( \uparrow \)

VC error bar for selecting a hypothesis from \( M \) using a data set of size \( K \).

\[ \text{...because we choose one of the } M \text{ finalists.} \]
Restoring $\mathcal{D}$

$\mathcal{H}_1$  \hspace{1cm} $\mathcal{H}_2$  \hspace{1cm} $\mathcal{H}_3$  \hspace{1cm} $\cdots$  \hspace{1cm} $\mathcal{H}_M$

$\downarrow$  \hspace{1cm} $\downarrow$  \hspace{1cm} $\downarrow$  \hspace{1cm} $\cdots$  \hspace{1cm} $\downarrow$

$g_1$  \hspace{1cm} $g_2$  \hspace{1cm} $g_3$  \hspace{1cm} $\cdots$  \hspace{1cm} $g_M$

Model with best $g$ also has best $\mathfrak{g}$

We can find model with best $\mathfrak{g}$ using validation

$\leftarrow$ leap of faith

$\leftarrow$ true modulo $E_{\text{val}}$ error bar
Comparing $E_{in}$ and $E_{val}$ for Model Selection

Validation Set Size, $K$

Expected $E_{out}$

- Validation: $g_{m^*}$
- In-sample: $g_{\hat{m}}$
- Optimal $g_{m^*}$

$D_{train}$ $\rightarrow$ $H_1$ $\rightarrow$ $g_1$ $\rightarrow$ $E_1$ $\rightarrow$ $H_{m^*}$

$D_{val}$ $\rightarrow$ $H_2$ $\rightarrow$ $g_2$ $\rightarrow$ $E_2$ $\rightarrow$ $H_M$

$D_{val}$ $\rightarrow$ $H_M$ $\rightarrow$ $g_M$ $\rightarrow$ $E_M$

Pick the best $(H_{m^*}, E_{m^*})$

$g_{m^*}$
Application to Selecting $\lambda$

Which regularization parameter to use?

$$\lambda_1, \lambda_2, \ldots, \lambda_M.$$ 

This is a special case of \textit{model selection} over $M$ models,

$$\begin{align*}
(\mathcal{H}, \lambda_1) &\rightarrow g_1 \\
(\mathcal{H}, \lambda_2) &\rightarrow g_2 \\
(\mathcal{H}, \lambda_3) &\rightarrow g_3 \\
&\cdots \\
(\mathcal{H}, \lambda_M) &\rightarrow g_M
\end{align*}$$

Picking a model amounts to choosing the optimal $\lambda$
The Dilemma When Choosing $K$

Validation relies on the following chain of reasoning,

$$E_{\text{out}}(g) \approx E_{\text{out}}(g^-) \approx E_{\text{val}}(g^-)$$

(small $K$) \hspace{1cm} (large $K$)
Can we get away with $K = 1$?

Yes, almost!
The Leave One Out Error \((K = 1)\)

\[
E[e_1] = E_{out}(\mathcal{g}_1)
\]

...but it is a **wild** estimate
The Leave One Out Errors

\[ E_{cv} = \frac{1}{N} \sum_{n=1}^{N} e_n \]
Cross Validation is Unbiased

**Theorem.** $E_{cv}$ is an unbiased estimate of $E_{out}(N - 1)$.
Reliability of $E_{cv}$

$e_n$ and $e_m$ are not independent.

- $e_n$ depends on $g_n$ which was trained on $(x_m, y_m)$.
- $e_m$ is evaluated on $(x_m, y_m)$.

Effective number of fresh examples giving a comparable estimate of $E_{out}$
Cross Validation is Computationally Intensive

\( N \) epochs of learning each on a data set of size \( N - 1 \).

- Analytic approaches, for example linear regression with weight decay

\[
\mathbf{w}_{\text{reg}} = (Z^T Z + \lambda I)^{-1} Z^T \mathbf{y}
\]

\[
E_{\text{cv}} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{\hat{y}_n - y_n}{1 - H_{nn}(\lambda)} \right)^2
\]

\[
H(\lambda) = Z(Z^T Z + \lambda I)^{-1} Z^T.
\]

- 10-fold cross validation

\[ \begin{array}{cccccccccc}
D_1 & D_2 & D_3 & D_4 & D_5 & D_6 & D_7 & D_8 & D_9 & D_{10} \\
\text{train} & \text{validate} & \text{train} & & & & & & & \\
\end{array} \]
Restoring $\mathcal{D}$

$$E_{\text{out}}(g^{(N)}) \leq \bar{E}_{\text{out}}(N - 1) \leq E_{\text{cv}} + O\left(\frac{1}{\sqrt{N}}\right).$$

$E_{\text{cv}}$ can be used for model selection just as $E_{\text{val}}$, for example to choose $\lambda$. 

CUSTOMER
Digits Problem: ‘1’ Versus ‘Not 1’

Average Intensity

\[ x = (1, x_1, x_2) \]

\[ z = (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3, \ldots, x_1^5, x_1^4 x_2, x_1^3 x_2^2, x_1^2 x_2^3, x_1 x_2^4, x_2^5) \]

5th order polynomial transform \( \rightarrow \) 20 dimensional non linear feature space
Validation Wins In the Real World

Average Intensity

no validation (20 features)

\[ E_{in} = 0\% \]

\[ E_{out} = 2.5\% \]

cross validation (6 features)

\[ E_{in} = 0.8\% \]

\[ E_{out} = 1.5\% \]