Review of Lecture 12

- Regularization
  
  constrained $\rightarrow$ unconstrained

  $E_{\text{in}} = \text{const.}$

  $w_{\text{lin}}$

  normal

  $w^T w = C$

  Minimize $E_{\text{aug}}(w) = E_{\text{in}}(w) + \frac{\lambda}{N} w^T w$

- Choosing a regularizer

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

  $\Omega(h)$: heuristic $\rightarrow$ smooth, simple $h$

  most used: weight decay

  $\lambda$: principled; validation

  $\lambda = 0.0001$  $\lambda = 1.0$
Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 13: Validation
Outline

• The validation set

• Model selection

• Cross validation
Validation versus regularization

In one form or another, \( E_{\text{out}}(h) = E_{\text{in}}(h) + \text{overfit penalty} \)

**Regularization:**
\[
E_{\text{out}}(h) = E_{\text{in}}(h) + \cancel{\text{overfit penalty}}
\]
regularization estimates this quantity

**Validation:**
\[
\hat{E}_{\text{out}}(h) = E_{\text{in}}(h) + \text{overfit penalty}
\]
validation estimates this quantity
Analyzing the estimate

On out-of-sample point \((\mathbf{x}, y)\), the error is \(e(h(\mathbf{x}), y)\)

Squared error: \((h(\mathbf{x}) - y)^2\)

Binary error: \([h(\mathbf{x}) \neq y]\)

\(\mathbb{E}[e(h(\mathbf{x}), y)] = E_{out}(h)\)

\(\text{var}[e(h(\mathbf{x}), y)] = \sigma^2\)
From a point to a set

On a validation set \((x_1, y_1), \ldots, (x_K, y_K)\), the error is \(E_{\text{val}}(h) = \frac{1}{K} \sum_{k=1}^{K} e(h(x_k), y_k)\)

\[
E \left[ E_{\text{val}}(h) \right] = \frac{1}{K} \sum_{k=1}^{K} E \left[ e(h(x_k), y_k) \right] = E_{\text{out}}(h)
\]

\[
\text{var} \left[ E_{\text{val}}(h) \right] = \frac{1}{K^2} \sum_{k=1}^{K} \text{var} \left[ e(h(x_k), y_k) \right] = \frac{\sigma^2}{K}
\]

\(E_{\text{val}}(h) = E_{\text{out}}(h) \pm O \left(\frac{1}{\sqrt{K}}\right)\)
$K$ is taken out of $N$

Given the data set $\mathcal{D} = (x_1, y_1), \cdots, (x_N, y_N)$

$K$ points $\to$ validation $\quad N - K$ points $\to$ training

$O\left(\frac{1}{\sqrt{K}}\right)$: Small $K \implies$ bad estimate

Large $K \implies$ ?

![Graph showing expected error vs. number of data points](image)
$K$ is put back into $N$

$$
\mathcal{D} \quad \rightarrow \quad \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \\
\downarrow \quad \downarrow \quad \downarrow \\
\mathcal{N} \quad \mathcal{N} - K \quad K
$$

$$
\mathcal{D} \quad \Rightarrow \quad g \\
\mathcal{D}_{\text{train}} \quad \Rightarrow \quad g^-
$$

$$
E_{\text{val}} = E_{\text{val}}(g^-) \quad \text{Large } K \quad \Rightarrow \quad \text{bad estimate!}
$$

Rule of Thumb:

$$
K = \frac{N}{5}
$$
Why ‘validation’

$D_{\text{val}}$ is used to make learning choices.

If an estimate of $E_{\text{out}}$ affects learning:

the set is no longer a test set!

It becomes a validation set.
What’s the difference?

Test set is unbiased; validation set has optimistic bias

Two hypotheses \( h_1 \) and \( h_2 \) with \( E_{\text{out}}(h_1) = E_{\text{out}}(h_2) = 0.5 \)

Error estimates \( e_1 \) and \( e_2 \) uniform on \([0, 1]\)

Pick \( h \in \{h_1, h_2\} \) with \( e = \min(e_1, e_2) \)

\( \mathbb{E}(e) < 0.5 \) optimistic bias
Outline

- The validation set
- Model selection
- Cross validation
Using $\mathcal{D}_{\text{val}}$ more than once

$M$ models $\mathcal{H}_1, \ldots, \mathcal{H}_M$

Use $\mathcal{D}_{\text{train}}$ to learn $g_m^{-}$ for each model

Evaluate $g_m^{-}$ using $\mathcal{D}_{\text{val}}$:  

$$E_m = E_{\text{val}}(g_m^{-}); \quad m = 1, \ldots, M$$

Pick model $m = m^*$ with smallest $E_m$
The bias

We selected the model $\mathcal{H}_{m^*}$ using $D_{\text{val}}$

$E_{\text{val}}(g_{m^*})$ is a biased estimate of $E_{\text{out}}(g_{m^*})$

Illustration: selecting between 2 models

![Graph showing expected error vs validation set size for $E_{\text{out}}(g_{m^*})$ and $E_{\text{val}}(g_{m^*})$.]
How much bias

For $M$ models: $\mathcal{H}_1, \ldots, \mathcal{H}_M$ $\mathcal{D}_{\text{val}}$ is used for "training" on the finalists model:

$$\mathcal{H}_{\text{val}} = \{ g_1^-, g_2^-, \ldots, g_M^- \}$$

Back to Hoeffding and VC!

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

regularization $\lambda$ early-stopping $T$
Data contamination

Error estimates: $E_{\text{in}}, E_{\text{test}}, E_{\text{val}}$

Contamination: Optimistic (deceptive) bias in estimating $E_{\text{out}}$

Training set: totally contaminated

Validation set: slightly contaminated

Test set: totally ‘clean’
Outline

- The validation set
- Model selection
- Cross validation
The dilemma about $K$

The following chain of reasoning:

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

(small $K$) \hspace{1cm} (large $K$)

highlights the dilemma in selecting $K$:

Can we have $K$ both small and large? 😊
Leave one out

$N - 1$ points for training, and \textbf{1 point} for validation!

$$\mathcal{D}_n = (x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n), (x_{n+1}, y_{n+1}), \ldots, (x_N, y_N)$$

Final hypothesis learned from $\mathcal{D}_n$ is $g^-_n$

$$e_n = E_{\text{val}}(g^-_n) = e(g^-_n(x_n), y_n)$$

cross validation error: \[ E_{cv} = \frac{1}{N} \sum_{n=1}^{N} e_n \]
Illustration of cross validation

\[ E_{cv} = \frac{1}{3} (e_1 + e_2 + e_3) \]
Model selection using CV

**Linear:**

**Constant:**
Cross validation in action

Digits classification task

Different errors

\[(1, x_1, x_2) \rightarrow (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, \ldots, x_1^5, x_1^4 x_2, x_1^3 x_2^2, x_1^2 x_2^3, x_1 x_2^4, x_2^5)\]
The result

without validation

\[ E_{\text{in}} = 0\% \quad E_{\text{out}} = 2.5\% \]

with validation

\[ E_{\text{in}} = 0.8\% \quad E_{\text{out}} = 1.5\% \]
Leave more than one out

Leave one out: \( N \) training sessions on \( N - 1 \) points each

More points for validation?

\[
\frac{N}{K} \text{ training sessions on } N - K \text{ points each}
\]

10-fold cross validation: \( K = \frac{N}{10} \)