Tuesday, 10/13/2020

# EE 660

# MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

**Lecture 15**

| Lecture 15 | EE 660 | Oct 13, 2020 |
|---|---|---|

## Announcements

- Project Assignment is posted on D2L (Week 8)
  - Detailed description of the project
  - Tips and suggestions for Type 1 and Type 2 projects
- => Work on creating a project topic and choosing/designing dataset(s)
- Project proposal instructions and forms will be posted Friday 10/16
  - Project proposals (Homework 6) will be due Friday 10/23
- Discussions 7 and 8 cover some aspects of the project

## Today's Lecture

- Validation and test

v1.1: [3] and [4] added to p.6.

# Validation and Test [AML 4.3]

In comparing models, we always (ideally) want $E_{out}$, or an estimate or bound for it.

Ex1: Finding best value $\lambda$ for a regularizer.

$$f_{obj}(\underline{w}, \lambda) = E_{in}(\underline{w}) + \lambda \|\underline{w}\|_b$$

Typically:

- Learning algorithm – finds $\hat{\underline{w}}$ from $\mathcal{D}_{Tr}$ for a given $\lambda$.
- Model selection — finds best $\lambda = \lambda^*$.

Ex2: $\hat{f}(\underline{x}) = d^{\underline{th}}$ order polynomial

- Learning algorithm — finds $\hat{\underline{w}}$ given $d$
- Model selection — finds best $d = d^*$.

## Key eqns. / bounds

$$
(i)\begin{cases}
\left.\begin{aligned}
E_{out}(h_g) &\le E_{\mathcal{D}_0}(h_g) + \varepsilon_{eff} \\
E_{out}(h_g) &\le E_{\mathcal{D}_0}(h_g) + \varepsilon_{VC}
\end{aligned}\right\} \text{ with probability } \ge 1-\delta \\
\varepsilon_{eff} = \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}} \le \varepsilon_{VC} = \sqrt{\frac{8}{N} \ln \frac{4\left[(2N)^{d_{VC}}+1\right]}{\delta}} \\
N = |\mathcal{D}_0|
\end{cases}
$$

e.g.: $\mathcal{D}_0 = \mathcal{D}_{Tr}$, $\mathcal{H} = \mathcal{H}_{Tr}$, $N = |\mathcal{D}_{Tr}|$, $d_{VC} = d_{VC}(\mathcal{H}_{Tr})$

$$
(ii)\begin{cases}
E_{out}(h_g) \le E_{\mathcal{D}_a}(h_g) + \varepsilon_M, \quad \varepsilon_M = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}, \quad \text{with probability } \ge 1-\delta. \\
M = |\mathcal{H}|, \quad N = |\mathcal{D}_a|
\end{cases}
$$

e.g.: if $\mathcal{D}_a = \mathcal{D}_{Test}$, then usually $M=1$.

We have done:

$$\begin{tikzcd}[column sep=small]
 & & \mathcal{D} \arrow[dl] \arrow[dr] & \\
 & \mathcal{D}'' \arrow[dl] \arrow[dr] & & \mathcal{D}_{Test} \\
\mathcal{D}_{PT} & & \mathcal{D}' &
\end{tikzcd}$$

$\mathcal{D}_{PT}$ (optional) $\mathcal{D}'$ (use for training, model selection, validation, ...)

$\mathcal{D}_{PT}, \mathcal{D}', \mathcal{D}_{Test}$ disjoint

Let's take iris classification ex (using Ex 1 above)

Suppose:

$$\hat{f}(\underline{x}) = \text{sign}\{\underline{w}^T \underline{x} + w_0\}$$

$$\Rightarrow \mathcal{H}': \left\{ h_{\underline{w}, w_0}(\underline{x}) = \hat{f}(\underline{x}) \;\middle|\; \underline{w} \in \mathbb{R}^D,\ w_0 \in \mathbb{R} \right\}$$

$$f_{obj}(\underline{w}, \lambda) = E_{in}(\underline{w}, w_0) + \lambda \|\underline{w}\|_1$$

Models: $(\mathcal{H}', \lambda_1), (\mathcal{H}', \lambda_2), \dots, (\mathcal{H}', \lambda_{M'})$

with $\lambda_m = m\,\delta_\lambda$, $m = 1, 2, \dots, M'$.

$\leftarrow$ increment in $\lambda$ (e.g., $\delta_\lambda = 0.01$).

For model selection, can use a validation set $\mathcal{D}_{Val}$:

$$\mathcal{D}' = \{\mathcal{D}_{Tr}, \mathcal{D}_{Val}\}, \quad \text{with } \mathcal{D}_{Tr} \cap \mathcal{D}_{Val} = \emptyset.$$

## Procedure

1. $$\mathcal{D} \to \mathcal{D}_{Test}, \quad \mathcal{D} \to \mathcal{D}_{PT} \text{ (optional)}, \quad \mathcal{D} \to \mathcal{D}'$$

2. Optionally use $\mathcal{D}_{PT}$.

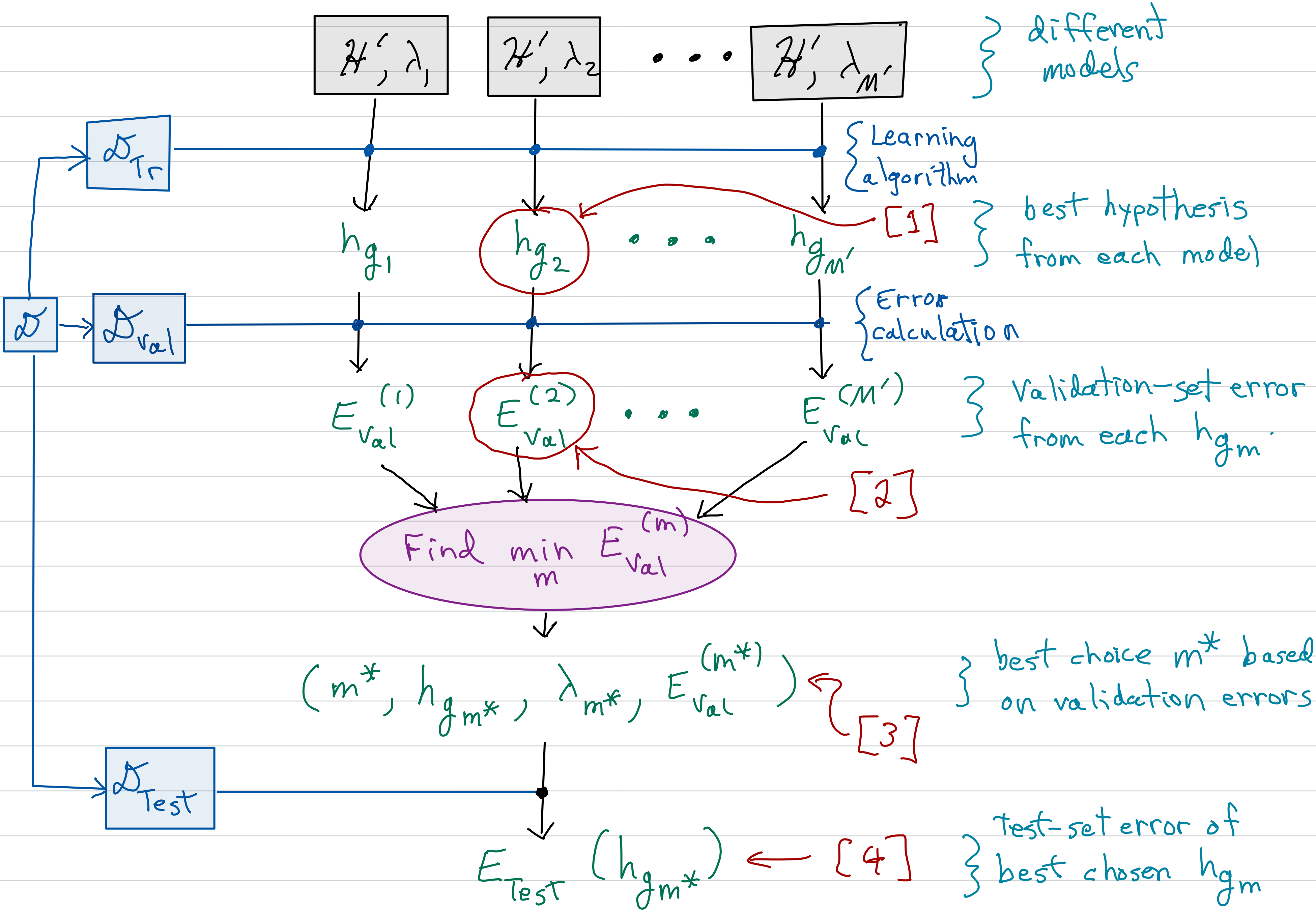
3. Set up $\mathcal{H}'$ and models $(\mathcal{H}', \lambda_m)$, $m = 1, 2, \cdots, M'$
   → Each model is defined by hyp. set $\mathcal{H}'$ and choice of $\lambda = \lambda_m$.

4. $$\mathcal{D}' \to \mathcal{D}_{Tr}, \quad \mathcal{D}' \to \mathcal{D}_{Val}$$ $\Big\}$ disjoint

   → Learning algorithm finds best hypothesis $h_{g_m}$ from $\mathcal{H}'$, by minimizing the objective fcn. with $\lambda = \lambda_m$.
   → Results in best hypothesis $h_{g_m}$ for each $m$.

   ⋮ (See diagram below).

For model selection:



[1] What hypothesis set did $h_{g_2}$ come from? $\mathcal{H}'$, with $d = d_2$.

Can we get a theoretical bound on $E_{out}(h_{g_2})$, using $\mathcal{D}_{Tr}$?

$\rightarrow$ Use (i) with $\mathcal{D}_o = \mathcal{D}_{Tr}$, $N = |\mathcal{D}_{Tr}|$

$$E_{out}\{h_{g_2}\} \le E_{\mathcal{D}_{Tr}}\{h_{g_2}\} + \sqrt{\frac{8}{N_{Tr}} \ln \frac{4\, m_{\mathcal{H}'}(2N_{Tr})}{\delta}}$$

[2] What hypothesis set did $E_{val}^{(2)}$ come from?

$\mathcal{H} = \{h_{g_2}\}$. $\Rightarrow$ Use (ii) with $\mathcal{D}_a = \mathcal{D}_{val}$, $M = 1$,

$N = |\mathcal{D}_{val}|$.