Thursday, 9/24/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 10

p. 1

Lecture 10 | **EE 660** | **Sep 24, 2020**

Announcements

- Homework ~~4~~ 3 is due tomorrow
- Homework ~~5~~ 4 will be posted

Today's Lecture

- Error measures and target types
- Approximation-generalization tradeoff
  - Bias-variance decomposition
  - Learning curves

Error Measures and Target Types

Previously we had: $E_{\mathcal{D}}(h) = \frac{1}{N} \sum_{n=1}^{N} [\![h(\underline{x}_n) \neq f(\underline{x}_n)]\!]$ = in-sample error

$E_{out}(h) = P[h(\underline{x}) \neq f(\underline{x})]$ = out-of-sample error

h, f are binary valued.

To extend to regression problems, we can instead use MSE:

$$E_{\mathcal{D}}(h) = \frac{1}{N} \sum_{n=1}^{N} [h(\underline{x}_n) - f(\underline{x}_n)]^2$$

$$E_{out}(h) = \mathbb{E}\{[h(\underline{x}) - f(\underline{x})]^2\}$$

($\mathbb{E}$ is w.r.t. $\underline{x}$, according to $p(\underline{x})$)

h, f can be real valued.

For $h \in \{0, 1\}$ and $f \in \{0, 1\}$,

are the 2 expressions for $E_{\mathcal{D}}(h)$ equivalent? Yes.

are the 2 expressions for $E_{out}(h)$ equivalent? Yes.

Approximation-Generalization Tradeoff

Bias and variance are a different way of assessing complexity and generalization.
→ Use MSE measures for $E_{\mathcal{D}}(h)$, $E_{out}(h)$.

Let $h_g = h_g^{(\mathcal{D})}$ = best hypothesis as chosen by $\mathcal{D}$.

$$E_{out}\left(h_g^{(\mathcal{D})}\right) = \mathbb{E}_{\underline{x}}\left\{\left[h_g^{(\mathcal{D})}(\underline{x}) - f(\underline{x})\right]^2\right\}$$

Does $f(\underline{x})$ depend on $\mathcal{D}$?
→ No.

($\mathbb{E}$ is w.r.t. $\underline{x}$, according to $p(\underline{x})$)

Take $\mathbb{E}_{\mathcal{D}}$, w.r.t. all datasets $\mathcal{D}$ of size N.
Each point in $\mathcal{D}$ is drawn from $\mathcal{X}$ according to $p(\underline{x})$.

$$\mathbb{E}_{\mathcal{D}}\left\{E_{out}\left(h_g^{(\mathcal{D})}\right)\right\} = \mathbb{E}_{\mathcal{D}}\left\{\mathbb{E}_{\underline{x}}\left\{\left[h_g^{(\mathcal{D})}(\underline{x}) - f(\underline{x})\right]^2\right\}\right\}$$

$$= \mathbb{E}_{\underline{x}}\left\{\mathbb{E}_{\mathcal{D}}\left\{\left[h_g^{(\mathcal{D})}(\underline{x}) - f(\underline{x})\right]^2\right\}\right\}$$

Let $\bar{h}_g(\underline{x}) \triangleq \mathbb{E}_{\mathcal{D}}\left\{h_g^{(\mathcal{D})}(\underline{x})\right\}$, so $\bar{h}_g^2(\underline{x}) = \left(\mathbb{E}_{\mathcal{D}}\left\{h_g^{(\mathcal{D})}(\underline{x})\right\}\right)^2$

and $\overline{h_g^2}(\underline{x}) \triangleq \mathbb{E}_{\mathcal{D}}\left\{h_g^{(\mathcal{D})2}(\underline{x})\right\}$

Drop $\underline{x}$ dependence (for now)

$$\mathbb{E}_{\mathcal{D}}\left\{E_{out}\left(h_g^{(\mathcal{D})}\right)\right\}$$

$$= \mathbb{E}_{\underline{x}}\left\{\mathbb{E}_{\mathcal{D}}\left\{h_g^{(\mathcal{D})2}\right\} - 2\,\mathbb{E}_{\mathcal{D}}\left\{h_g^{(\mathcal{D})} f\right\} + \underbrace{\mathbb{E}_{\mathcal{D}}}_{\text{Drop}}\left\{f^2\right\}\right\}$$

$$= \mathbb{E}_{\underline{x}}\left\{\underbrace{\overline{h_g^2} - \overline{h_g}^2}\ + \underbrace{\overline{h_g}^2 - 2\,\overline{h_g}\, f + f^2}\right\}$$

$$= \mathbb{E}_{\underline{x}}\left\{\underbrace{\mathbb{E}_{\mathcal{D}}\left\{\left(h_g - \overline{h_g}\right)^2\right\}} + \underbrace{\left[\overline{h_g} - f\right]^2}\right\}$$

$$\text{bias}(\underline{x}) \triangleq \left[\overline{h_g}(\underline{x}) - f(\underline{x})\right]^2 \ *$$

$$\text{var}(\underline{x}) \triangleq \mathbb{E}_{\mathcal{D}}\left\{\left[h_g^{(\mathcal{D})}(\underline{x}) - \overline{h_g}(\underline{x})\right]^2\right\}$$

$$\therefore\ \mathbb{E}_{\mathcal{D}}\left\{E_{out}\left(h_g^{(\mathcal{D})}\right)\right\} = \underbrace{\mathbb{E}_{\underline{x}}\left\{\text{bias}(\underline{x})\right\}}_{\triangleq\ \text{bias}\ *} + \underbrace{\mathbb{E}_{\underline{x}}\left\{\text{var}(\underline{x})\right\}}_{\triangleq\ \text{var}}$$

$$\boxed{\mathbb{E}_{\mathcal{D}}\left\{E_{out}\left(h_g^{(\mathcal{D})}\right)\right\} = \text{bias} + \text{var}.}$$

*(Some refs. call this $\text{bias}^2(\underline{x})$ and bias^2.)

Interpretation

bias: how much the average learned hypothesis $\bar{h}_g$ (over all $\mathcal{D}$ of size N) differs from the true target function f, averaged over $\underline{x}$ (as MSE).

var: variation in learned hypotheses across different training datasets $\mathcal{D}$ of size N.

Example (AML example 2.8)

↦ 1D problem (regression)
Target fcn:

$$f(x) = \sin(\pi x), \quad x \text{ is sampled uniformly on } [-1, +1].$$

Dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$ $(N=2)$.

2 hypothesis sets:

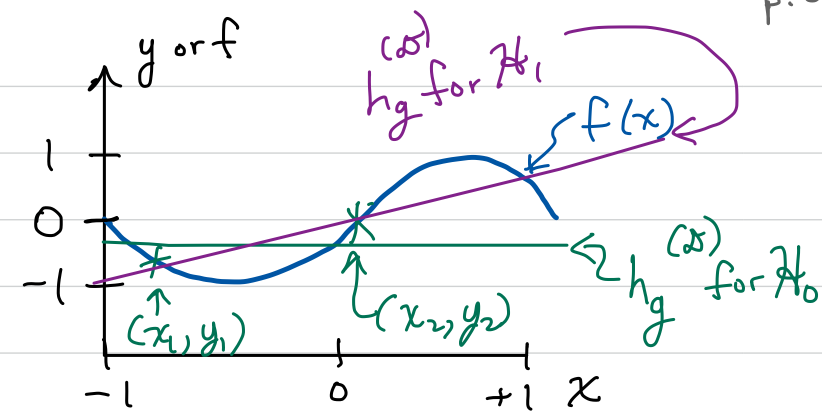
$\mathcal{H}_0$: set of all horizontal lines $\{h_b(x) = b\}$ $(M=\infty)$

$\mathcal{H}_1$: set of all lines $\{h_{a,b}(x) = ax + b\}$ $(M=\infty)$

Learning algorithm (minimizes MSE):

$\mathcal{H}_0$: $\quad b = \dfrac{y_1 + y_2}{2}$

$\mathcal{H}_1$: $\quad$ Line passing through (x_1, y_1) and (x_2, y_2)



1 draw of a dataset $\mathcal{D}$.

→ Figures on AML p. 65.

Learning Curves

Review:

Bias–Var. viewpoint

$$\mathbb{E}_{\mathcal{D}}\{E_{out}(h_g^{(\mathcal{D})})\} = \mathbb{E}_{\underline{x}}\{\text{bias}(\underline{x}) + \text{var}(\underline{x})\} = \text{bias} + \text{var}$$

$$\text{bias}(\underline{x}) = \left[\bar{h}_g(\underline{x}) - f(\underline{x})\right]^2 ; \quad \text{var}(\underline{x}) = \mathbb{E}_{\mathcal{D}}\left\{\left[h_g^{(\mathcal{D})}(\underline{x}) - \bar{h}_g(\underline{x})\right]^2\right\}$$

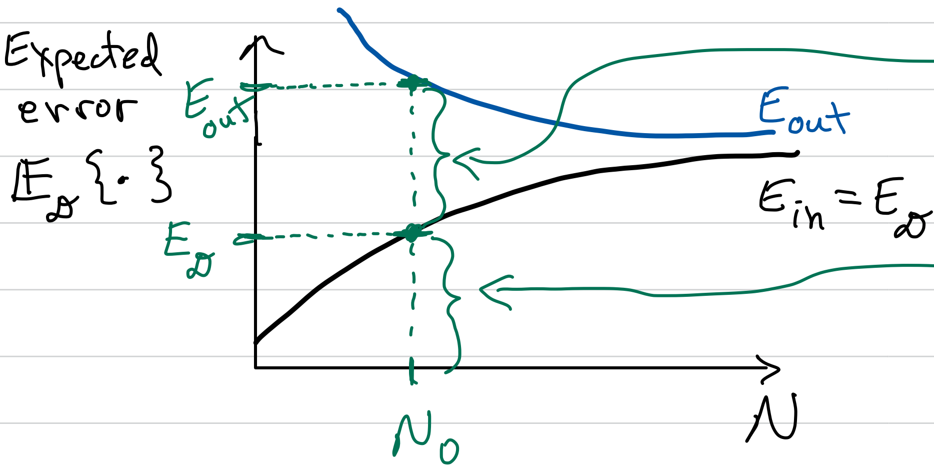
VC viewpoint

$$E_{out}(h_g^{(\mathcal{D})}) \leq E_{\mathcal{D}}(h_g^{(\mathcal{D})}) + \mathcal{E}(N, \mathcal{H}, \delta) \quad \text{with probability} \doteq 1-\delta.$$

$$\mathbb{E}_{\mathcal{D}}\{E_{out}(h_g^{(\mathcal{D})})\} \leq \mathbb{E}_{\mathcal{D}}\{E_{\mathcal{D}}(h_g^{(\mathcal{D})})\} + \mathcal{E}(N, \mathcal{H}, \delta)$$

$$\mathcal{E}(N, \mathcal{H}, \delta) \leq \sqrt{\frac{8}{N} \ln \frac{4\left[(2N)^{d_{vc}} + 1\right]}{\delta}} = \mathcal{E}_{vc}.$$

VC perspective



$$\mathbb{E}_{\mathcal{D}}\left\{E_{out}\left(h_g^{(\mathcal{D})}\right) - E_{\mathcal{D}}\left(h_g^{(\mathcal{D})}\right)\right\} \le \epsilon(N, \mathcal{H}, \delta)$$

$$\mathbb{E}_{\mathcal{D}}\left\{E_{\underset{\mathcal{D}}{\underset{\uparrow}{in}}}\left(h_g^{(\mathcal{D})}\right)\right\}$$