Thursday, 10/15/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 16

Announcements

- Project proposal instructions and forms will be posted tomorrow
  - Project proposals (Homework 6) will be due Friday 10/23
- Project FAQs will also be posted
- Discussions 7 and 8 cover some aspects of the project

Today's Lecture

- Finish validation and test
- AML material - concluding remarks
  - Occam's Razor
  - Axiom of Non-Falsifiability
  - Data Snooping
  - Sampling Bias

deferred

* One omission and one error have been fixed (post lecture) on p. 3.

Continue Validation and Test

[3] To estimate E_{out} from $E_{val}^{(\mathcal{D}_{val})}(h_{g_{m^*}})$, what is $\mathcal{H}$?

$\rightarrow \mathcal{H} = \{h_{g_1}, h_{g_2}, \cdots, h_{g_{M'}}\} \triangleq \mathcal{H}''$

We can use (ii), with $M = |\mathcal{H}''| = M'$; $\mathcal{D}_a = \mathcal{D}_{val}$, $N = |\mathcal{D}_{val}|$

Or we can use (i), with $d_{vc}(\mathcal{H}'')$ or $m_{\mathcal{H}''}(2N)$; $\mathcal{D}_b = \mathcal{D}_{val}$, $N = |\mathcal{D}_{val}|$

[4] For $E_{out}(h_{g_{m^*}})$ based on $E_{Test}(h_{g_{m^*}})$:

$\mathcal{H} = \{h_{g_{m^*}}\}$

Use (ii) with $M = 1$, $\mathcal{D}_a = \mathcal{D}_{Test}$, $N = |\mathcal{D}_{Test}|$

Comment: Flow charts for cross validation are described in AML 4.3.3.

Occam's Razor [AML 5.1]

∃ many versions.

"When choosing among multiple theories* that are otherwise equivalent, the simplest one is best."

* or: models, hypotheses, or hypothesis sets.

In ML:

$$E_{out}(h) \le \underbrace{E_{\underset{\substack{\uparrow \\ (\mathcal{D})}}{in}}(h)}_{\text{e.g.: } \mathcal{D} = \mathcal{D}_{Test}} + \underbrace{\varepsilon(N, \mathcal{H}, \delta)}_{\text{Then smaller is better}}$$

If same or similar among multiple h

Then smaller is better (lower M, lower d_{vc}, lower $m_{\mathcal{H}}(2N)$ $\Rightarrow$ smaller ε and simpler $\mathcal{H}$.)

In terms of regularization:

$$E_{aug}(h, \lambda) = E_{in}(h) + \underbrace{\frac{\lambda}{N}\,\Omega(h)}, \qquad \lambda \ge 0$$

larger λ prefer smaller set of hypotheses within $\mathcal{H}$, resulting in a simpler h_g.

Axiom of Non-Falsifiability [AML 5.1]

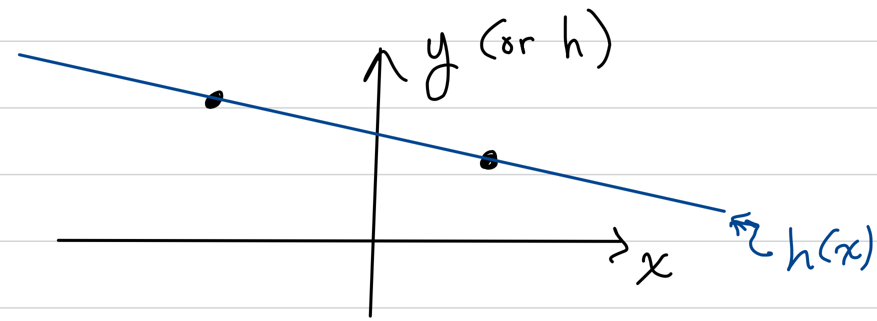
For data to provide evidence of a hypothesis, it must be possible for the data to falsify the hypothesis

Ex 1 Hypothesis (or "conjecture"):

The target function $f(x)$ is a linear fcn. of x.

Let $\mathcal{H}: \{h \mid h(x) = ax+b, a \in \mathbb{R}, b \in \mathbb{R}\}$,

$\mathcal{D}$ { $N=2$.



Given: data is noiseless.

$\mathcal{H}$ can fit the data with $MSE = 0. = E_{\mathcal{D}}(h_g)$.

Can we conclude target fcn. $f(x)$ actually is linear?

→ No, because: it's not possible for $\mathcal{D}$ with $N=2$ to contradict (or falsify) $\mathcal{H}$ or our conjecture.

→ In ML terms, complexity of $\mathcal{H}$ (VCdim) is too high for the given N, D.

Axiom of Non-Falsifiability — Example 2 [based on AML Example 5.2]

Financial firm, hiring a stock trader.
Want to choose the best candidate out of all applicants.
See how well they can predict the stock market.

1. Ask each candidate to predict the stock market (up or down) for the next day, and do this each day for 5 days.
2. The number of candidates $M >> 2^5$.
3. After 5 days of predictions, all 2^5 possible 5-day predictions are represented.
4. For each possible set of 5-day predictions, pick the candidate who otherwise has the best qualifications.
5. Of the remaining 2^5 candidates, hire the one who got all 5 predictions correct.

Is the chosen candidate the best one?

What if all the candidates made their predictions by flipping coins?

→ There would still likely be one or more that got all 5 predictions correct.

→ A "best" trader will be identified, even if none of them are any better than the others.

→ Can't falsify the conjecture that some of the candidates are better at predicting the stock market than others.

ML interpretation:

This is a model selection procedure.
Each candidate uses their own model for making each prediction.

What is $\mathcal{H}$?

$\mathcal{H}$ is a set of 5-tuples; each 5-tuple is 5 next-day predictions.

$$\mathcal{H} = \{ h_m, m = 1, 2, 3, \cdots, 32 \}.$$

For multiple candidates that give the same 5-day predictions, other information is used to pick the best ($\longleftrightarrow$ learning algorithm)

The model selection picks the best model h_{m^*} out of $\mathcal{H}$.

What is $\mathcal{D}_{val}$?

$\rightarrow$ The set of 5 next-day predictions: x = day i, $\hat{y}$ = prediction for day $i+1$.

$\rightarrow$ N_{val} = # predictions made by each model $= 5$

$\mathcal{H}$: $M = ?$ $\quad d_{vc} = ?$

$M = 32$ $\quad d_{vc} = 5$