Lecture 18 announcements

- HW 8 is due Friday (only 24-hour 2-point penalty period)
- Solutions to HW 8 will be posted on Saturday
- Sample midterm exam has been posted
- Midterm exam is Tuesday, 10/23, 3:30 - 4:50 PM.
  - On campus exam will be in THH 201.

Today's topics

- Wrap-up of recent material:
  - Occam's Razor
  - Axiom of non-falsifiability
  - ~~Data snooping~~
- Review for midterm exam

* ANNOTATIONS ADDED TO PG. 2.

* - [REVIEW] (LECTURES 1-17) -

* LECTURE 18

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.

WHAT DOES THIS MEAN IN TERMS OF:

$$E_{out}(h) \leq E_{in}(h) + \mathcal{E}$$

IF SAME OR SIMILAR — THEN CHOOSE SMALLER OR SMALLEST (M, $m_{\mathcal{H}}(N)$, $d_{vc}(N)$)

*

OR + REGULARIZATION:

$$E_{aug}(h, \lambda) = E_{in}(h) + \frac{\lambda}{N}\Omega(h), \quad \lambda \geq 0.$$

LARGER λ RESTRICTS $\mathcal{H}$

MORE, OR, PREFERS SIMPLER (SMOOTHER) h.

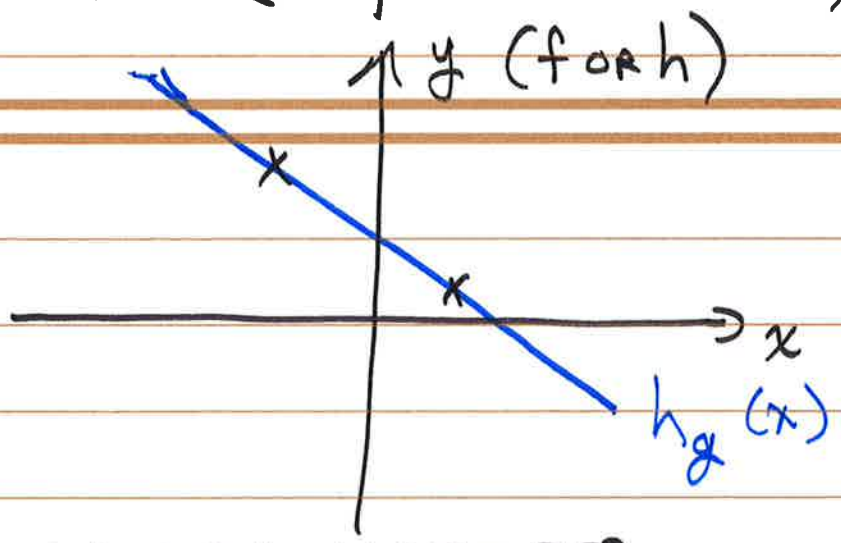
Axiom on Non-Falsifiability

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

Ex: Hypothesis: the target function $f(x)$ is a linear function 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.

$\exists\, h(x)$ that gives $E_{in}(h_g) = 0$.

Can we conclude $f(x)$ is linear?

→ No. It's not possible for $\mathcal{D}$ with $N=2$ to contradict (falsify) $\mathcal{H}$.

EXAMPLE [BASED ON AML EXMPL. 5.2]

FINANCIAL FIRM, HIRING A STOCK TRADER.

- EA. CANDIDATE — PREDICT MARKET EACH DAY FOR NEXT 5 DAYS. (UP OR DOWN)
- FOR DUPLICATE PREDICTIONS, PICK BEST ONE (other info.)
- ASSUME ALL REMAINING POOL OF CAND.'S INCLUDE ALL 2^5 PREDICTIONS.
- WAIT 5 DAYS.

- THEN PICK CANDIDATE W/ BEST PREDICTION.

→ 1 CANDIDATE WILL GET 100% CORRECT.

IS THIS THE BEST CANDIDATE?

No.

E.G., EA-CANDIDATE DOES A RANDOM COIN TOSS

⇒ $\mathcal{H}$ IS NON-FALSIFIABLE

⇒ $\mathcal{H}$ IS TOO COMPLEX FOR THE DATASET.