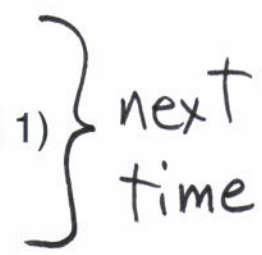


Announcements

- Homework 6 has been posted
 - Midterm exam is 4 weeks from today (Tues., Oct. 23)
-

Today's Lecture

- Midterm ground rules
 - Implications in dataset methodology (part 2)
 - Other error measures and target types
 - Approximation-generalization tradeoff (part 1)
 - Bias Variance decomposition
- 

MIDTERM GROUND RULES

12 FULLY CLOSED BOOK

29 CLOSED BOOK WITH 1 FORMULA SHEET

1 OPEN CLASS NOTES AND NOTES IN YOUR OWN
HANDWRITING

~~OPEN " " , HWS~~

0 OPEN ALL CLASS MATERIALS EXCEPT TEXTBOOKS.

WATCH FOR EMAIL W/ LINK TO A POLL.

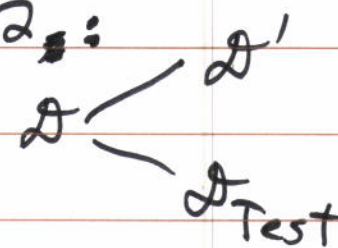
IMPLICATIONS IN DATASET METHODOLOGY (part 2)

POSSIBLE ALTERNATE PROCEDURES:

from last time -

II. SAME AS I, EXCEPT:

→ BEFORE STEP 2:


1.5 DIVIDE \mathcal{D} 

SET $\mathcal{D}_{\text{Test}}$ ASIDE (NO SNOOPING!)

→ DELETE 7(a).

7. BEFORE COMPUTING $E_{\text{Test}}(h_g)$, APPLY SAME PREPROC., FEAT. EXTR., AS WAS DONE FOR FINAL SYSTEM ON \mathcal{D}' . (USING ONLY INFORMATION FROM \mathcal{D}' .)

POSSIBLE PITFALL: $(E_{\text{Val}} \text{ or } E_{\text{Tr}})$

\mathcal{D}_{Val} (or \mathcal{D}_{Tr})  MIGHT NOT GENERALIZE TO E_{out} .

- INCREASE N IF POSSIBLE

- MORE (CORRECTLY DONE) CROSS VAL.

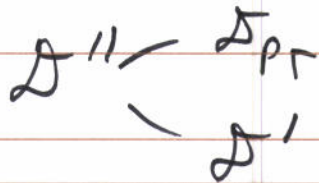
4

(LIMIT)
- WATCH THE COMPLEXITY OF ALL
OF USED.

III. ~~(1)(a) DRAW~~

1.5 MAY WANT TO \mathcal{D} $\begin{matrix} \nearrow \mathcal{D}'' \\ \searrow \mathcal{D}_{\text{Test}} \end{matrix}$ AS IN II.

1.6 (a) DRAW \mathcal{D}_{PT} (PRE-TRAINING)
(WITHOUT REPLACEMENT) FROM \mathcal{D}' :



(b) USE \mathcal{D}_{PT} TO LOOK AT DATA,
CONDUCT INITIAL TESTS, ETC.
(E.G., STEPS (2)-(5)).

(c) DISCARD \mathcal{D}_{PT} AFTER I.

(6) ~~(5)~~ PROCEED WITH (6). (Already
removed $\mathcal{D}_{\text{Test}}$ though).

~~6.5 DISCARD \mathcal{D}_{PT} .~~

(7) (a) - IS VALID.

IV. (1) USE PRIOR KNOWLEDGE OF PROBLEM TO:

— DECIDE ON FEATURES, PRE-PROC, ETC

— CONSTRUCT \mathcal{H} (CONSIDERING $d_{vc}(\mathcal{H})$,
 N_{Tr} , N_{Val} , ETC.)

(2) DRAW \mathcal{D}' , \mathcal{D}_{Test}

(3) SET ASIDE \mathcal{D}_{Test}

(4) USE \mathcal{D}' FOR MODEL SELECTION, TRAINING,
 CHOOSING h_g .

(5) ONCE h_g IS FINAL, 7(a) AND 7(b)
 CAN BE USED.

V. CONSTRUCT YOUR OWN ~~NEW~~ METHOD, AS LONG AS
 YOU DON'T VIOLATE ASSUMPTIONS FOR
 GENERALIZATION ERROR ~~THE~~ THAT YOU
 COMPUTE AND RELY ON.

THE KEY ASSUMPTIONS:

$$(i) \quad E_{out}(h_g) \leq E_{\mathcal{D}_0}^{(h_g)} + \underbrace{\sqrt{\frac{8}{N} \ln \frac{4[C_2 N]^{d_{VC}} + 1}{8}}}_{\text{BASED ON } \mathcal{H}_0}$$

ASSUMES:

BASED ON \mathcal{H}_0 .

1. \mathcal{D}_0 AND \mathcal{H}_0 MUST BE CONSISTENT.

2. INFO. IN \mathcal{D}_0 CAN'T BE USED TO CONSTRUCT \mathcal{H}_0 .

3. $N = N_{\mathcal{D}_0}$.

$$(ii) \quad E_{out}(h_g) \leq E_{Test}(h_g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{8}}$$

$$M = 1.$$

ASSUME:

1. INFO. IN \mathcal{D}_{Test} CAN'T INFLUENCE CHOICE OF h_g .

2. $N = N_{Test}$.

(MORE ON \mathcal{D}_{Val} , MODEL SELECTION, DATASET USAGE, LATER.).