

Tuesday, 9/29/2020

EE 660

MACHINE LEARNING
FROM SIGNALS:
FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture II

Lecture 11**EE 660****Sep 29, 2020**

Announcements

- Homework 4 (Week 5) has been posted

Today's Lecture

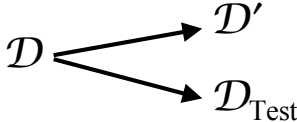
- Dataset methodology and generalization bounds (part 2)
- Overfitting

Implications of VC Generalization Bound on Dataset Methodology

Consider some possible ML scenarios

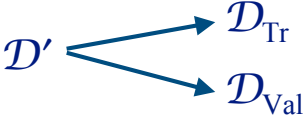
Scenario I: Don't think ahead

1. Collect data and construct dataset \mathcal{D}
2. Do preliminary data analysis
 - Plot the data
 - Look over the data, learn about its structure
3. Preprocessing
 - Standardize
 - Noise filtering
4. Feature extraction and selection
 - Extract a set of features (by design or automatically)
 - Feature selection process to reduce their number
5. Do some preliminary trials on \mathcal{D}
 - Try a few hypothesis sets and learning algorithms
 - Decide on \mathcal{H}

6. Divide dataset \mathcal{D}  such that $\mathcal{D}' \cap \mathcal{D}_{\text{Test}} = \emptyset$

7. Set up and run model selection and learning algorithms

- Run cross-validation on \mathcal{D}' to choose parameters, train, and find h_g

- Using \mathcal{D}'  in rotation

8. Evaluate performance of h_g

(a) Based on training-set or validation-set error

- Calculate $E_{\mathcal{D}_0}(h_g)$ using $\mathcal{D}_0 = \mathcal{D}_{\text{Tr}}$ or $\mathcal{D}_0 = \mathcal{D}_{\text{Val}}$
- Calculate ϵ_{VC} using $d_{VC}(\mathcal{H})$ (also using N, δ)
- Get “error bar” on $E_{\text{out}}(h_g)$

(b) Based on test-set error

- Calculate $E_{\text{Test}}(h_g)$ using $\mathcal{D}_{\text{Test}}$
- Calculate ϵ_M using $M = 1$ (also using N, δ)

Refer to Lecture 9,
page 8 for summary
of formulas and
assumptions

=> Are these $E_{\text{out}}(h_g)$ error bars valid?

(a) No

Used \mathcal{D}' in Steps 2-5, before setting up \mathcal{H} .
Hoeffding Inequality and VC bound don't apply.

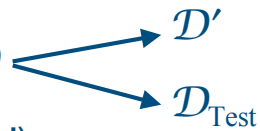
(b) No

Used \mathcal{D} (and therefore $\mathcal{D}_{\text{Test}}$) before deciding on h_g .
Can't use $M = 1$.

Scenario II: Set test-set aside at beginning

- Revisions to Scenario I:

- Before Step 2:

- Step 1.5: Divide dataset \mathcal{D}  such that $\mathcal{D}' \cap \mathcal{D}_{\text{Test}} = \emptyset$
- Set $\mathcal{D}_{\text{Test}}$ aside (no snooping!)

- Delete Step 6 ($\mathcal{D}_{\text{Test}}$ was already extracted)

- Evaluate performance of h_g :

- Is 8(a) valid ? (calculate $E_{\text{out}}(h_g)$ based on training-set or validation-set error, using ϵ_{VC} and $d_{VC}(\mathcal{H})$)

- No.

Validation set was used to construct \mathcal{H} .

- Is 8(b) valid ? (calculate $E_{\text{out}}(h_g)$ based on test-set error, using ϵ_M and $M=1$)

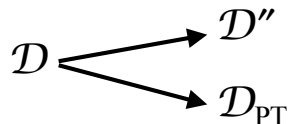
- Yes!

h_g and \mathcal{H} did not depend on the test set. (Equivalent to “drawing” test set after h_g was chosen, so $M=1$.)

- A common scenario in ML.

- The test set error can be generalized to $E_{\text{out}}(h_g)$ using VC generalization theorem
- But the validation-set and training-set error cannot be used in the VC generalization theorem

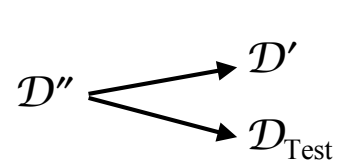
Scenario III: Use a pre-training set

1. Draw pre-training set \mathcal{D}_{PT} (without replacement) from \mathcal{D} : 
 - Such that $\mathcal{D}'' \cap \mathcal{D}_{PT} = \emptyset$

2. Use \mathcal{D}_{PT} to
 - Look at data, conduct initial trials, etc.

3. Construct \mathcal{H} (considering \mathcal{D}_{PT} results, $d_{VC}(\mathcal{H})$, N , etc.)

4. Discard \mathcal{D}_{PT}

5. Draw \mathcal{D}' and \mathcal{D}_{Test} from \mathcal{D}'' : 
 - $(\mathcal{D}_{Test} \cup \mathcal{D}' = \mathcal{D}'')$
 - $(\mathcal{D}' \cap \mathcal{D}_{Test} = \emptyset)$

6. Set aside \mathcal{D}_{Test}

7. Use \mathcal{D}' to run learning algorithms, model selection, and choose h_g

8. Evaluate performance of h_g

- Is 8(a) valid? (calculate $E_{out}(h_g)$ based on training-set or validation-set error, using ε_{VC} and $d_{VC}(\mathcal{H})$)

- Yes.

- Is 8(b) valid ? (calculate $E_{out}(h_g)$ based on test-set error, using ε_M and $M=1$)

- Yes.

Comments

- Other scenarios are possible
 - You can create your own!
- Later we will discuss
 - Using cross-validation in different ways, with VC generalization bound
 - A graphical way of keeping track of dataset usage, and what ε or M to use for calculating VC generalization bounds.

Overfitting [Am2 4-1]

Def: Overfit is "an analysis which corresponds too closely or exactly to a particular set of data." [Oxford Dictionary]

Common symptom of overfitting: picking a hypothesis with lower E_{in} results in a higher E_{out} .

Following Am2:

Consider an experiment:

Target function: $y = f(\underline{x}) + n$

\uparrow \nwarrow noise
 \underline{x} = a scalar = x

$f(x) = 10^{th}$ order polynomial in x . (target fcn.)

$N = 15$ points

Hypothesis sets: \mathcal{H}_2 : 2nd order polynomials
 \mathcal{H}_{10} : 10th " "

[Am2 figs & table, pp 120-121]

$\Rightarrow H_{10}$ includes the (true) target fcn.
 H_2 does not.
 H_{10} has better E_{in}
 H_2 has better E_{out} !

$\Rightarrow H_{10}$ can't distinguish between noise n and target $f(x)$.
Fits too much to noise.

\Rightarrow Best hypothesis set complexity depends on quality of data.