Lecture 16: Three Learning Principles

Hsuan-Tien Lin (林軒田)
htlin@csie.ntu.edu.tw

Department of Computer Science & Information Engineering
National Taiwan University (國立台灣大學資訊工程系)
Roadmap

1. When Can Machines Learn?
2. Why Can Machines Learn?
3. How Can Machines Learn?
4. How Can Machines Learn Better?

Lecture 15: Validation
(crossly) reserve validation data to simulate testing procedure for model selection

Lecture 16: Three Learning Principles
- Occam’s Razor
- Sampling Bias
- Data Snooping
- Power of Three
Occam’s Razor

An explanation of the data should be made as simple as possible, but no simpler.—Albert Einstein? (1879-1955)

entia non sunt multiplicanda praeter necessitatem
(entities must not be multiplied beyond necessity)
—William of Occam (1287-1347)

‘Occam’s razor’ for trimming down unnecessary explanation

figure by Fred the Oyster (Own work) [CC-BY-SA-3.0], via Wikimedia Commons
Occam’s Razor for Learning

The simplest model that fits the data is also the most plausible.

which one do you prefer? :-)

two questions:

1. What does it mean for a model to be simple?
2. How do we know that simpler is better?
Three Learning Principles

**Occam’s Razor**

**Simple Model**

**simple hypothesis** $h$

- small $\Omega(h) = \text{‘looks’ simple}$
- specified by **few** parameters

**simple model** $\mathcal{H}$

- small $\Omega(\mathcal{H}) = \text{not many}$
- contains **small number of hypotheses**

**connection**

$h$ specified by $\ell$ bits $\iff |\mathcal{H}|$ of size $2^\ell$

small $\Omega(h)$ $\iff$ small $\Omega(\mathcal{H})$

**simple:** small hypothesis/model complexity
in addition to math proof that you have seen, philosophically:

simple $\mathcal{H}$
\[ \implies \text{smaller } m_{\mathcal{H}}(N) \]
\[ \implies \text{less ‘likely’ to fit data perfectly } \frac{m_{\mathcal{H}}(N)}{2^N} \]
\[ \implies \text{more significant when fit happens} \]

direct action: linear first;
always ask whether data over-modeled
Fun Time
Three Learning Principles

Sampling Bias

Presidential Story

- 1948 US President election: Truman versus Dewey
- A newspaper phone-poll of how people voted, and set the title ‘Dewey Defeats Truman’ based on polling.

who is this? :-)
Sampling Bias

The Big Smile Came from . . .

Truman, and yes he won

suspect of the mistake:
- editorial bug?—no
- bad luck of polling ($\delta$)?—no

hint: phones were expensive :-)}
Sampling Bias

If the data is sampled in a biased way, learning will produce a similarly biased outcome.

- Technical explanation:
  data from $P_1(x, y)$ but test under $P_2 \neq P_1$: VC fails

- Philosophical explanation:
  study Math hard but test English: no strong test guarantee

‘minor’ VC assumption:
  data and testing both iid from $P$
Three Learning Principles

Sampling Bias

Sampling Bias in Learning

A True Personal Story

- Netflix competition for movie recommender system:
  \(10\% \text{ improvement} = 1\text{M US dollars}\)
- formed \(D_{val}\), in my first shot, \(E_{val}(g)\) showed 13\% improvement
- why am I still teaching here? :-)

validation: random examples within \(D\);
  test: ‘last’ user records ‘after’ \(D\)

Hsuan-Tien Lin  (NTU CSIE)  Machine Learning Foundations  10/25
Sampling Bias

Dealing with Sampling Bias

If the data is sampled in a biased way, learning will produce a similarly biased outcome.

- practical rule of thumb:
  **match test scenario as much as possible**
- e.g. if test: ‘last’ user records ‘after’ $D$
  - training: emphasize later examples (KDDCup 2011)
  - validation: use ‘late’ user records

last puzzle:

danger when learning ‘credit card approval’ with existing bank records?
Three Learning Principles

Data Snooping

Visual Data Snooping

Visualize $\mathcal{X} = \mathbb{R}^2$

- full $\Phi_2$: $z = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$, $d_{VC} = 6$
- or $z = (1, x_1^2, x_2^2)$, $d_{VC} = 3$, after visualizing?
- or better $z = (1, x_1^2 + x_2^2)$, $d_{VC} = 2$?
- or even better $z = (\text{sign}(0.6 - x_1^2 - x_2^2))$?

—careful about your brain’s ‘model complexity’

for VC-safety, $\Phi$ shall be decided without ‘snooping’ data
If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.

- 8 years of currency trading data
- first 6 years for training, last two 2 years for testing
- $x =$ previous 20 days, $y =$ 21th day
- snooping versus no snooping: superior profit possible

- snooping: shift-scale all values by training + testing
- no snooping: shift-scale all values by training only
Three Learning Principles

Data Snooping

Data Snooping by Data Reusing

Research Scenario

benchmark data $\mathcal{D}$

- paper 1: propose $\mathcal{H}_1$ that works well on $\mathcal{D}$
- paper 2: find room for improvement, propose $\mathcal{H}_2$ —and publish only if better than $\mathcal{H}_1$ on $\mathcal{D}$
- paper 3: find room for improvement, propose $\mathcal{H}_3$ —and publish only if better than $\mathcal{H}_2$ on $\mathcal{D}$
- ... 

- if all papers from the same author in one big paper:
  bad generalization due to $d_{VC}(\bigcup_m \mathcal{H}_m)$
- step-wise: later author snooped data by reading earlier papers, bad generalization worsen by publish only if better

if you torture the data long enough, it will confess :-)
Dealing with Data Snooping

- truth—very hard to avoid, unless being extremely honest
- extremely honest: lock your test data in safe
- less honest: reserve validation and use cautiously
- be blind: avoid making modeling decision by data
- be suspicious: interpret research results (including your own) by proper feeling of contamination

one secret to winning KDDCups:

careful balance between
data-driven modeling (snooping) and
validation (no-snooping)
Fun Time
Three Learning Principles

Power of Three

Three Related Fields

Data Mining
- use (huge) data to find property that is interesting
- difficult to distinguish ML and DM in reality

Artificial Intelligence
- compute something that shows intelligent behavior
- ML is one possible route to realize AI

Statistics
- use data to make inference about an unknown process
- statistics contains many useful tools for ML
### Three Theoretical Bounds

<table>
<thead>
<tr>
<th>Principle</th>
<th>Formula</th>
<th>Hypothesis</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hoeffding</strong></td>
<td>$P[\text{BAD}] \leq 2 \exp(-2\epsilon^2 N)$</td>
<td>one</td>
<td>useful for verifying/testing</td>
</tr>
<tr>
<td><strong>Multi-Bin Hoeffding</strong></td>
<td>$P[\text{BAD}] \leq 2M \exp(-2\epsilon^2 N)$</td>
<td>$M$</td>
<td>useful for validation</td>
</tr>
<tr>
<td><strong>VC</strong></td>
<td>$P[\text{BAD}] \leq 4m_{\mathcal{H}}(2N) \exp(\ldots)$</td>
<td>all $\mathcal{H}$</td>
<td>useful for training</td>
</tr>
</tbody>
</table>
Three Linear Models

PLA/pocket

$h(x) = \text{sign}(s)$

plausible $err = 0/1$
(small flipping noise)
minimize specially

linear regression

$h(x) = s$

friendly $err = \text{squared}$
(easy to minimize)
minimize analytically

logistic regression

$h(x) = \theta(s)$

plausible $err = \text{CE}$
(maximum likelihood)
minimize iteratively

Hsuan-Tien Lin (NTU CSIE)
Three Key Tools

Feature Transform

\[ E_{\text{in}}(w) \rightarrow E_{\text{in}}(\tilde{w}) \]
\[ d_{\text{VC}}(\mathcal{H}) \rightarrow d_{\text{VC}}(\mathcal{H}_\Phi) \]

- by using more complicated \( \Phi \)
- lower \( E_{\text{in}} \)
- higher \( d_{\text{VC}} \)

Regularization

\[ E_{\text{in}}(w) \rightarrow E_{\text{in}}(w_{\text{REG}}) \]
\[ d_{\text{VC}}(\mathcal{H}) \rightarrow d_{\text{EFF}}(\mathcal{H}, \mathcal{A}) \]

- by augmenting regularizer \( \Omega \)
- lower \( d_{\text{EFF}} \)
- higher \( E_{\text{in}} \)

Validation

\[ E_{\text{in}}(h) \rightarrow E_{\text{val}}(h) \]
\[ \mathcal{H} \rightarrow \{g_1^-, \ldots, g_M^-\} \]

- by reserving \( K \) examples as \( \mathcal{D}_{\text{val}} \)
- fewer choices
- fewer examples
Three Learning Principles

Power of Three

- Occam's Razer: simple is good
- Sampling Bias: class matches exam
- Data Snooping: honesty is best policy
Three Future Directions

- More Transform
- More Regularization
- Less Label

Ready for the jungle!
Fun Time
Summary

1. When Can Machines Learn?
2. Why Can Machines Learn?
3. How Can Machines Learn?
4. How Can Machines Learn Better?

Lecture 15: Validation
Lecture 16: Three Learning Principles

- Occam’s Razor
  simple, simple, simple!
- Sampling Bias
  match test scenario as much as possible
- Data Snooping
  any use of data is ‘contamination’
- Power of Three
  relatives, bounds, models, tools, principles

• next: ready for jungle!