Machine Learning
(機器學習)

Lecture 09: Wisdom on Using Machine Learning

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 09: Wisdom on Using Machine Learning

- 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) 不做不必要的事

'**Occam's razor**' for trimming down unnecessary explanation

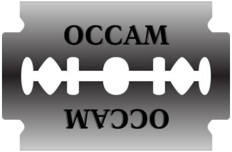
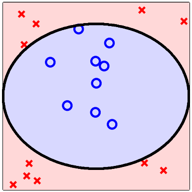


figure by Fred the Oyster (Own work) [CC-BY-SA-3.0], via Wikimedia Commons

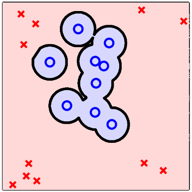
Occam's Razor for Learning

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



eaiser better

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?

Simple Model

simple hypothesis h

- small $\Omega(h)$ = 'looks' simple
- specified by **few parameters**

simple model $\mathcal{H}$ Hypothesis set 小

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

connection 用 ℓ bits 描述即可

h specified by ℓ bits $\Leftarrow$ $|\mathcal{H}|$ of size 2^ℓ

Hypothesis set

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

simple: **small hypothesis/model complexity**

Simple is Better

in addition to **math proof** that you have seen, philosophically:

simple $\mathcal{H}$

用簡单模型來預測，如果Ein大：代表資料沒有規律

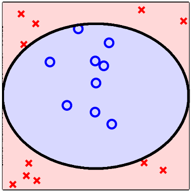
$\Longrightarrow$ smaller $m_{\mathcal{H}}(N)$

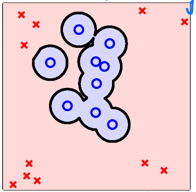
Ein小：代表資料有規律

$\Longrightarrow$ less 'likely' to fit data perfectly $\dfrac{m_{\mathcal{H}}(N)}{2^N}$

$\Longrightarrow$ more significant when fit happens

fit發生時代表一定資料有規律





先試簡单的模型

direct action: **linear first**;
always ask whether **data over-modeled**

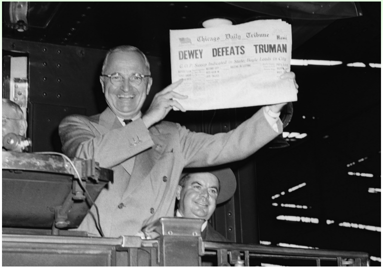
Questions?

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? :-)

The Big Smile Came from . . .



Truman, and **yes he won**

suspect of the mistake:

- editorial bug?—**no**
- bad luck of polling (δ)?—**no**

hint: phones were **expensive :-)**

Sampling Bias 抽樣有偏差

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

traing data 与 testing data 来自不同的地方

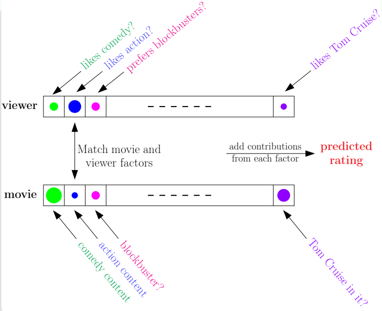
- technical explanation:
  data from $P_1(\mathbf{x}, y)$ but test under $P_2 \neq P_1$: **VC fails**
- philosophical explanation:
  study Math hard but test English: **no strong test guarantee**

'minor' VC assumption:
data and testing **both iid from *P***

Sampling Bias in Learning

A True Personal Story

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



validation: random examples within $\mathcal{D}$;
test: 'last' user records 'after' $\mathcal{D}$

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' $\mathcal{D}$
  - training: emphasize later examples (KDDCup 2011)
  - validation: use 'late' user records

盡可能跟 testing 愈接近愈好.

experience: hard to **avoid** sampling bias,
but can **sysematically correct some**

Questions?

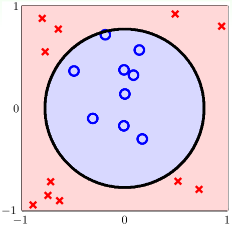
Visual Data Snooping 偷看

2维data,我就用2次多项式?

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

- full Φ_2: $\mathbf{z} = (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2)$, $d_{\mathrm{VC}} = 6$
- or $\mathbf{z} = (1, x_1^2, x_2^2)$, $d_{\mathrm{VC}} = 3$, **after visualizing**?
- or better $\mathbf{z} = (1, x_1^2 + x_2^2)$, $d_{\mathrm{VC}} = 2$?
- or even better $\mathbf{z} = \left(\mathrm{sign}(0.6 - x_1^2 - x_2^2)\right)$?

—careful about **your brain's 'model complexity'**



for VC-safety, Φ shall be
decided **without 'snooping'** data

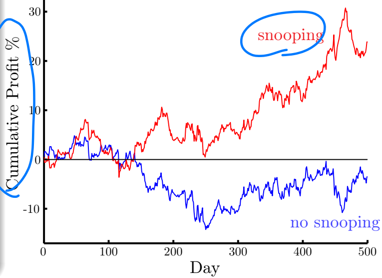
Data Snooping by Mere Shifting-Scaling

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

投報率

拿了8年的data.



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

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 $d_{\text{VC}}(\cup_m \mathcal{H}_m)$ ← d_{vc}是所有paper加起來
- step-wise: later author **snooped** data by reading earlier papers,
  bad generalization worsen by **publish only if better** 不要文獻回顧也

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:

偷看也小心一点

careful balance between
data-driven modeling (snooping) and
validation (no-snooping)

Questions?

Three Related Fields

Power of Three

3個領域

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

Power of Three

3个理論

Hoeffding

$$P[\text{BAD}] \leq 2\exp(-2\epsilon^2 N)$$

- **one** hypothesis
- useful for **verifying/testing**

Multi-Bin Hoeffding

有限选择

$$P[\text{BAD}] \leq 2M\exp(-2\epsilon^2 N)$$

- M hypotheses
- useful for **validation**

VC

無限选擇

$$P[\text{BAD}] \leq 4m_{\mathcal{H}}(2N)\exp(\ldots)$$

- all $\mathcal{H}$
- useful for **training**

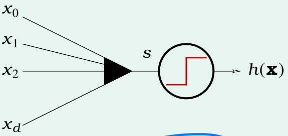
Three Linear Models

Power of Three

3个 model

PLA

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



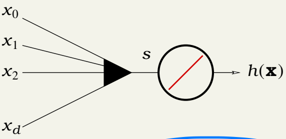
plausible err = 0/1

(small flipping noise)

minimize **specially**

linear regression

$$h(\mathbf{x}) = s$$



friendly err = squared

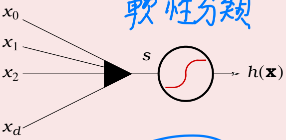
(easy to minimize)

minimize **analytically**

logistic regression

$$h(\mathbf{x}) = \theta(s)$$

軟性分類



plausible err = CE

(maximum likelihood)

minimize **iteratively**

Three Key Tools

Power of Three

3个工具

Feature Transform

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

- by using **more complicated** Φ
- **lower** E_{in}
- higher d_{VC}

Regularization

$$E_{\text{in}}(\mathbf{w}) \rightarrow E_{\text{in}}(\mathbf{w}_{\text{REG}})$$
$$d_{\text{VC}}(\mathcal{H}) \rightarrow d_{\text{EFF}}(\mathcal{H}, \mathcal{A})$$

限制 d_{vc}.

- by augmenting **regularizer** Ω
- **lower** d_{EFF}
- higher E_{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

更接近 E_{out}

Three Learning Principles

Power of Three

3个妙計

Occam's Razer	Sampling Bias	Data Snooping
simple is good	class matches exam	honesty is best policy

Questions?

Summary

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

Lecture 09: Wisdom on Using Machine Learning

- 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