

# Week 16: Three Learning Principles

## Table of Contents

1. [Occam's Razor](#)
2. [Sampling Bias](#)
3. [Data Snooping](#)
4. [Power of Three \(Course Summary\)](#)

## Occam's Razor

1. Occam's razor for learning: The **simplest** model that fits the data is also the **most plausible**
2. What does "simple" mean in model context?
  - **Simple hypothesis  $h$** 
    - Small  $\Omega(h) \rightarrow$  "Looks" simple
    - Specified by **few parameters**
  - **Simple model  $\mathcal{H}$** 
    - Small  $\Omega(\mathcal{H}) \rightarrow$  Small number
    - Hypothesis set (given by the model) contains small number of hypothesis
  - Connection between simple hypothesis and simple model
    - Assume a hypothesis set  $\|\mathcal{H}\|$  of size  $2^l$
    - Requires **at most**  $l$  bits to specify any hypothesis  $h$  in  $\mathcal{H}$
    - Small  $\Omega(\mathcal{H}) \Rightarrow$  Small  $\Omega(h)$
  - **Simple**, in modeling context, means **small hypothesis set, or low model complexity**
3. Why simpler is better?

Simple  $\mathcal{H}$

$\Rightarrow$  Smaller growth function  $m_{\mathcal{H}}(N)$

$\Rightarrow$  Less likely to produce perfect fit for input data. Likelihood of perfect fit =  $\frac{m_{\mathcal{H}}(N)}{2^N}$

$\Rightarrow$  Each parameter is more significant when fit does happen

- Always try **linear models first**
- Always ask whether **data is over-modeled**

## Sampling Bias

1. If the data is sampled in a biased way, learning will produce a similarly biased outcome
  - If training data is drawn from  $P_1(x, y)$ , but test is done under  $P_2 \neq P_1 \Rightarrow$  **VC guarantee fails**

- VC assumes training and testing data are **i.i.d** from the same  $P$
- 2. Note that sampling bias can occur **unintentionally**
  - e.g. Inherent sequence in the population means that if training/validation set are formed with naive 'first-70%, leftover-30%' split, the resulting samples will be biased
- 3. Practical rule of thumb: **Match test scenario as much as possible**
  - e.g. If test uses **last** records *after* sample set  $D$ , such as performing video recommendation based on user's watch history
    - Training: Emphasize *later* examples (assuming there exists sequence identifier)
    - Validation: Use "late" user records

## Data Snooping

1. For VC-safety, any transformation  $\phi$  should be decided **without snooping** data
  - Snooping: Constructing/choosing models **after** looking into the data to find patterns
2. If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.
  - Because patterns in the data set has already been integrated (fully or in parts) into the models learned
3. Snooping vs. training
  - When data intended for testing is used instead as part of training set, while the exact type/dimensions of the selected model might not be affected, the individual parameter weights can still be affected
  - **Snooping**: Shift-scale all values by **training + testing** data
  - **No snooping**: Shift-scale by **training** data only
4. Data snooping can happen unintendedly
  - e.g. Research paper using the same public data set for benchmarking
  - Every paper after the original one gets published **only if the model described out-performs the original model**
  - Later models unintentionally incorporate information from the benchmark dataset that were already learned by pervious models
5. Dealing with data snooping
  - Data snooping is **very hard to avoid**, unless being extremely honest
    - "Lock test data in a safe and use only for final testing"
  - The more practical approach (with some risk of data snooping)
    - Reserve validation set
    - Use validation set with caution. The more a given validation set is used, the less effective it will be for model selection, because the set has been "containminated", or snooped by some of the models that have seen this set.
  - Rules of thumb
    - **Be Blind**: Avoid making modeling decision by data
    - **Be suspicious**: Interpret research results with proper **feeling of data contamination**

## Power of Three (Course Summary)

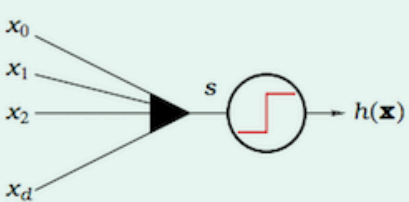
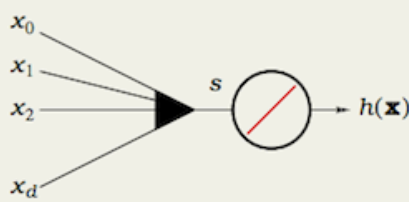
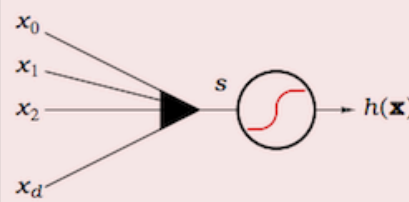
1. Three related fields

Data Mining	Artificial Intelligence	Statistics
<ul style="list-style-type: none"> <li>use <b>(huge)</b> data to <b>find property</b> that is interesting</li> <li>difficult to distinguish ML and DM in reality</li> </ul>	<ul style="list-style-type: none"> <li>compute something that shows <b>intelligent behavior</b></li> <li>ML is one possible route to realize AI</li> </ul>	<ul style="list-style-type: none"> <li>use data to <b>make inference</b> about an unknown process</li> <li>statistics contains many useful tools for ML</li> </ul>

2. Three theoretical bounds

Hoeffding	Multi-Bin Hoeffding	VC
$P[\text{BAD}] \leq 2 \exp(-2\epsilon^2 N)$ <ul style="list-style-type: none"> <li><b>one</b> hypothesis</li> <li>useful for <b>verifying/testing</b></li> </ul>	$P[\text{BAD}] \leq 2M \exp(-2\epsilon^2 N)$ <ul style="list-style-type: none"> <li><b>M</b> hypotheses</li> <li>useful for <b>validation</b></li> </ul>	$P[\text{BAD}] \leq 4m_{\mathcal{H}}(2N) \exp(\dots)$ <ul style="list-style-type: none"> <li>all <math>\mathcal{H}</math></li> <li>useful for <b>training</b></li> </ul>

3. Three linear models

PLA/pocket	linear regression	logistic regression
$h(\mathbf{x}) = \text{sign}(s)$  <p>plausible err = 0/1 (small flipping noise) minimize <b>specially</b></p>	$h(\mathbf{x}) = s$  <p>friendly err = squared (easy to minimize) minimize <b>analytically</b></p>	$h(\mathbf{x}) = \theta(s)$  <p>plausible err = CE (maximum likelihood) minimize <b>iteratively</b></p>

4. Three key tools

Feature Transform	Regularization	Validation
$E_{\text{in}}(\mathbf{w}) \rightarrow E_{\text{in}}(\tilde{\mathbf{w}})$ $d_{\text{VC}}(\mathcal{H}) \rightarrow d_{\text{VC}}(\mathcal{H}_{\Phi})$	$E_{\text{in}}(\mathbf{w}) \rightarrow E_{\text{in}}(\mathbf{w}_{\text{REG}})$ $d_{\text{VC}}(\mathcal{H}) \rightarrow d_{\text{EFF}}(\mathcal{H}, \mathcal{A})$	$E_{\text{in}}(h) \rightarrow E_{\text{val}}(h)$ $\mathcal{H} \rightarrow \{g_1^-, \dots, g_M^-\}$
<ul style="list-style-type: none"> <li>• by using <b>more complicated <math>\Phi</math></b></li> <li>• <b>lower <math>E_{\text{in}}</math></b></li> <li>• higher <math>d_{\text{VC}}</math></li> </ul>	<ul style="list-style-type: none"> <li>• by augmenting <b>regularizer <math>\Omega</math></b></li> <li>• <b>lower <math>d_{\text{EFF}}</math></b></li> <li>• higher <math>E_{\text{in}}</math></li> </ul>	<ul style="list-style-type: none"> <li>• by reserving <math>K</math> examples as <math>\mathcal{D}_{\text{val}}</math></li> <li>• <b>fewer choices</b></li> <li>• fewer examples</li> </ul>

5. Three learning principles

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