ILLC Project Course in Statistical Learning Theory

Mathias Winther Madsen
mathias.winther@gmail.com

Institute for Logic, Language, and Computation
University of Amsterdam

January 2015

Membership Processes

Problem

I select one of four sets:

1. $A_1 = \mathbb{N}$;
2. $A_2 = \{1, 3, 5, \ldots\}$;
3. $A_3 = \{2, 4, 6, \ldots\}$;
4. $A_4 = \emptyset$.

Asking only yes/no questions, how quickly can you determine which set I chose?

Problem

I select a set $A \subseteq \mathbb{N}$ containing two or fewer elements, and you ask me whether $x \in A$ for $x = 1, 2, 3, 4$.

How many ways can I potentially answer those four questions?

Membership Processes

Definition

A **conditional membership process** is a binary process defined in terms of the following parameters:

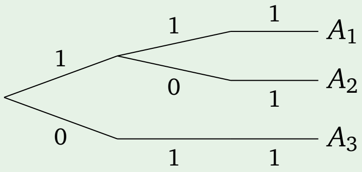
1. a sample space Ω;
2. a set S of subsets $A \subseteq \Omega$;
3. a sequence $x = x_1, x_2, x_3, \ldots$ of elements of Ω.

This process admits the sequences $y = y_1, y_2, y_3, \ldots$ for which there is an $A \in S$ such that

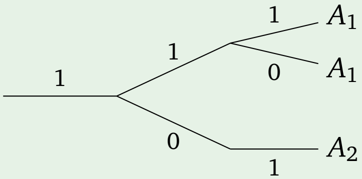
$$y_i = \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{if } x_i \notin A \end{cases}$$

Membership Processes

$$A_1 = \{1, 2, 3\}; \quad A_2 = \{1, 3\}; \quad A_3 = \{2, 3\}.$$



$x_1 = 1 \quad x_2 = 2 \quad x_3 = 3$



$x_1 = 3 \quad x_2 = 2 \quad x_3 = 1$

$$N(t \,|\, X = x) = 2, 3, 3$$

$$N(t \,|\, X = x) = 1, 2, 3$$

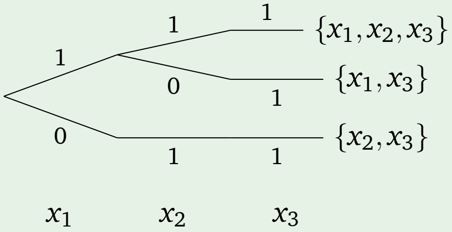
Membership Processes

Definition

The **projection** of a portfolio S onto a sample x is

$$S \downarrow x \;=\; \{A \cap \{x_1, x_2, \ldots, x_t\} \,|\, A \in S\}.$$

Example



Membership processes

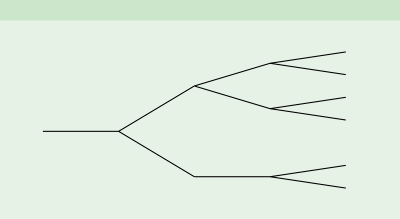
Definition

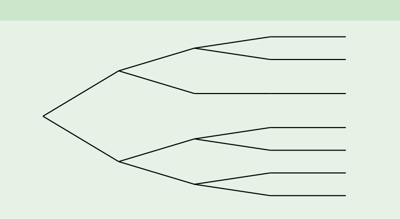
A **membership process** is a family of conditional membership processes, one for each sequence $x = x_1, x_2, x_3, \ldots$

Definition

The **growth function** of a membership processes is

$$N(t) \;=\; \max_x N(t \,|\, X = x).$$





Membership processes

Problem

Let $\Omega = \mathbb{N}$ and $S = \{\mathbb{N}, \text{odds}, \text{evens}, \emptyset\}$. What is the growth function and entropy rate of the corresponding membership process?

Problem

Let $\Omega = \mathbb{R}$ and $S = \{\{r \leq \theta\} \mid \theta \in \mathbb{R}\}$. What is $N(t)$ and H?

Problem

Let S consist of all sets $A \subseteq \Omega$ with $|A| \leq 2$. What is $N(t)$ and H?

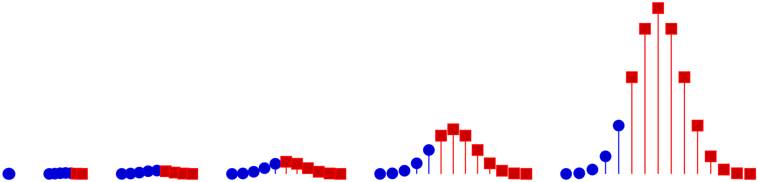
The VC Bound

$$\Phi(k,t) = \binom{t}{0} + \binom{t}{1} + \binom{t}{2} + \cdots + \binom{t}{k}$$

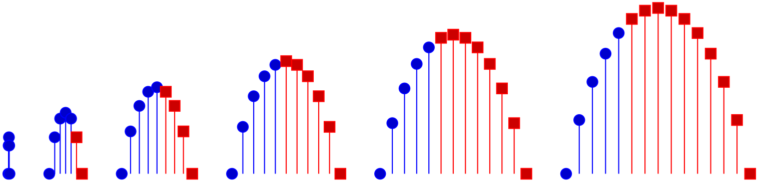
	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$t=0$	1	1	1	1	1	1
$t=1$	1	2	2	2	2	2
$t=2$	1	3	4	4	4	4
$t=3$	1	4	7	8	8	8
$t=4$	1	5	11	15	16	16
$t=5$	1	6	16	26	31	32

The VC Bound

For $k = 4$ and $t = 4, 6, 8, 10, 12, 14$:



Same, logarithmic plot:



The VC Bound

						1						
					1		1					
				1		2		1				
			1		3		3		1			
		1		4		6		4		1		
	1		5		10		10		5		1	
1		6		15		20		15		6		1

$$\binom{t}{k} = \binom{t-1}{k-1} + \binom{t-1}{k}$$

$$\Phi(k,t) = \Phi(k-1,t-1) + \Phi(k,t-1)$$

The VC Bound

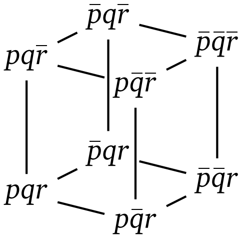
Theorem

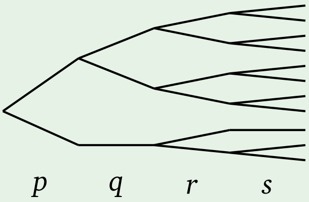
Suppose there is a k such that

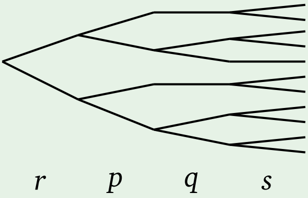
$$N(t \mid x) \;\geq\; \Phi(k, t).$$

Then $x = x_1, x_2, \ldots, x_t$ *has a subsequence* $z = z_1, z_2, \ldots, z_k$ *for which*

$$N(k \mid z) \;=\; 2^k.$$







The VC Bound

Theorem

Suppose that for any subsequence z *of* x*,*

$$N(k \mid z) \;<\; 2^k.$$

Then

$$N(t \mid x) \;<\; \Phi(k, t).$$

Proof.

By induction on t for a fixed but arbitrary k. $\square$