A Sharp Threshold Result for VC-Classes of Large-Margin Functions

Joel Ratsaby

Department of Computer Science

University College London

Plan

General theme: Combinatoric complexity of binary-function classes under different learning input

- Learning setting, learning input
- Effective hypothesis class
- Its cardinality (as combinatoric complexity)
- Three learning-input scenarios
- Compute the combinatoric complexity
- Results
- Conclusion

Setting

- Finite domain X, denoted as $[n] = \{1, \ldots, n\}$
- Hypothesis class $\mathcal{H}$ of functions $h : X \to Y$, $Y = \{0, 1\}$ (or $\{-1, 1\}$)
- Vapnik Chervonenkis dimension $VC(\mathcal{H}) \leq d$
- Effectively, $\mathcal{H}$ is finite, e.g., binary-output neural networks over binary-valued inputs
- Assume target t, but it need not be in $\mathcal{H}$ (possibly unrealizable)
- Learning input Z consists of an i.i.d. (unknown P) training sample $S_m \subseteq X \times Y$, $S_m = \{(x_i, y_i)\}_{i=1}^m$, $x_i \in X$, $y_i \in Y$, *plus* possibly other side information I_t dependent on t
- Sample S_m labeled by target, $y_i = t(x_i)$, $1 \leq i \leq m$, noiseless

Learning

- Empirical error $L_m(h) = (1/m)\sum_{i=1}^{m} \mathbb{I}(h(x_i) \neq y_i)$
- Generalization error $L(h) = \mathbf{P}(h(x) \neq y)$, $L^* = \min_{h\in\mathcal{H}} L(h)$
- Define $\mathcal{H}_I \equiv \{h \in \mathcal{H} : h \text{ satisfies } I_t\}$, e.g., all h s.t. $L(h) \leq L^* + \alpha$
- Instead of starting to learn from $\mathcal{H}$, the learner using Z starts from an effective hypothesis class $\mathcal{H}(Z)$ based on available learning input Z

$$\mathcal{H}(Z) \equiv \{h \in \mathcal{H}_I : L_m(h) = L_m^*(Z)\}$$

  where $L_m^*(Z) \equiv \min_{h\in\mathcal{H}_I} L_m(h)$
- Given a future set S_l of training examples
- **Aim**: Learn $\hat{h}_l$ from $\mathcal{H}(Z)$ in the ERM-sense, i.e., $\hat{h}_l = \text{argmin}_{h\in\mathcal{H}(Z)} L_l(h)$

Cardinality $|\mathcal{H}(Z)|$

- In ERM, learning generalization-error $L(\hat{h}_l)$ is basically controlled by cardinality of $\mathcal{H}(Z)$, i.e.,

$$\mathbf{P}\left(\exists S_l, \mathrm{L}(\hat{h}_l) > L^*(Z) + \epsilon\right) \leq |\mathcal{H}(Z)|e^{-2l\epsilon^2}$$

  where $L^*(Z) = \min_{h \in \mathcal{H}(Z)} L(h)$

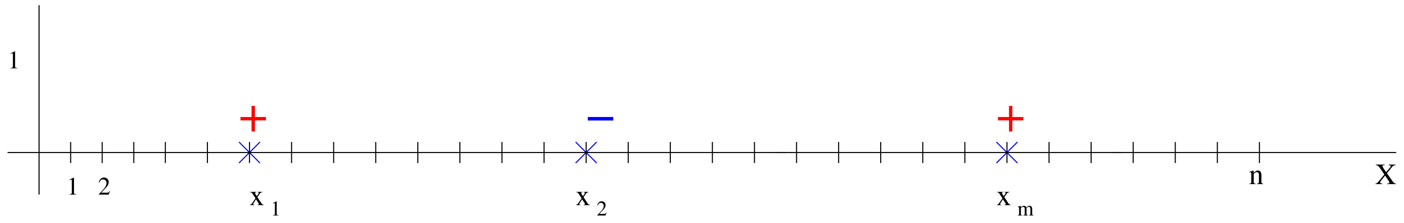
- Wish to understand how $|\mathcal{H}(Z)|$ varies for different learning inputs Z

Side information I_t

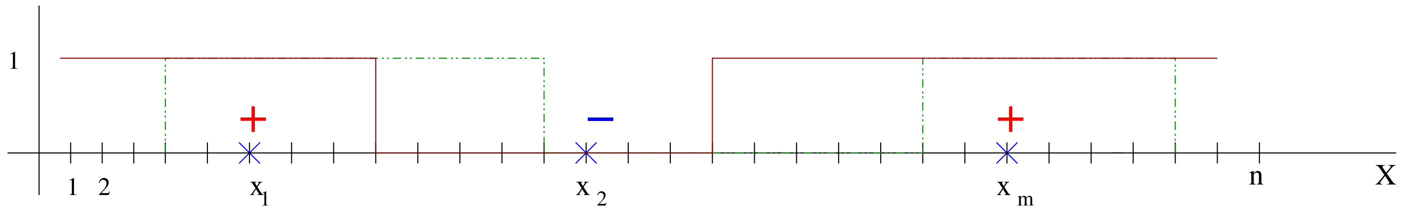
- In practice, side information I_t may come in various 'forms'
- For instance, with high probability $I_t =$ 'there is a maximal-margin (on S_m) hypothesis h_γ whose error $L(h_\gamma) \leq L^* + \epsilon_\gamma$'
- This information I_t could be the result of learning by maximizing-margin over $\mathcal{H}$ and finding a solution with a sample margin $> \gamma$ on S_m
- We may then consider to continue learning but restrict the effective class to hypotheses that have a large margin $> \gamma$

Case I: Basic learning-input

- Learning input Z: sample S_m



- Possible consistent hypotheses



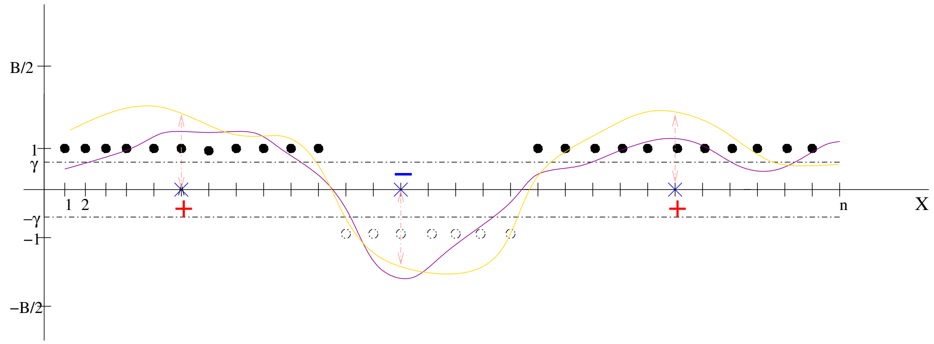
- Using Sauer's Lemma (Sauer '72, Vapnik & Chervonenkis '71),

$$|\mathcal{H}(Z)| \leq \sum_{k=0}^{d} \binom{n-m}{k} \leq \left(\frac{e(n-m)}{d}\right)^d \quad \text{(tight bound)} \qquad (1)$$

Case II: 'large-margin' learning-input

- Learning input Z: fixed parameter $\gamma > 0$, a sample S_m, information $I_t =$ 'there is an $\hat{h} \in \mathcal{H}$ with a margin on S_m no less than γ and $L(\hat{h}) \leq L^* + \epsilon_\gamma$'
- What is $\mathcal{H}(Z)$?
- Standard approach, consider class $\mathcal{F}$ of real-valued functions $f : X \to \mathbb{R}$, $Y = \{-1, +1\}$
- Margin $M_f(x_i)$ of f on x_i is $y_i f(x_i)$
- Sample margin of f, $M_{S_m}(f) = \min_{1 \leq i \leq m} M_f(x_i)$
- We can consider $\mathcal{F}_\gamma = \{f \in \mathcal{F} : M_{S_m}(f) \geq \gamma\}$

- Large redundancy: many f induce the same binary function h

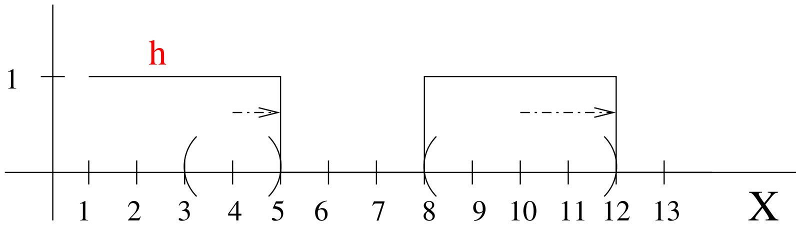


- No sense in taking these multiple f into an effective class
- We really need the class $\mathcal{H}_\gamma = \{h = \mathrm{sgn}(f) : M_{S_m}(f) \geq \gamma\}$
- or alternatively ...

A direct margin definition

Consider a margin directly defined for binary-valued functions.

Definition: For any binary-valued function $h \in \mathcal{H}$ the margin $\mu_h(x)$ on an element $x \in [n]$ is the largest non-negative integer a such that h has a constant value of either 0 or 1 on the interval set $I_a(x) = \{x - a, \ldots, x + a\}$ provided that $I_a(x) \subseteq [n]$.

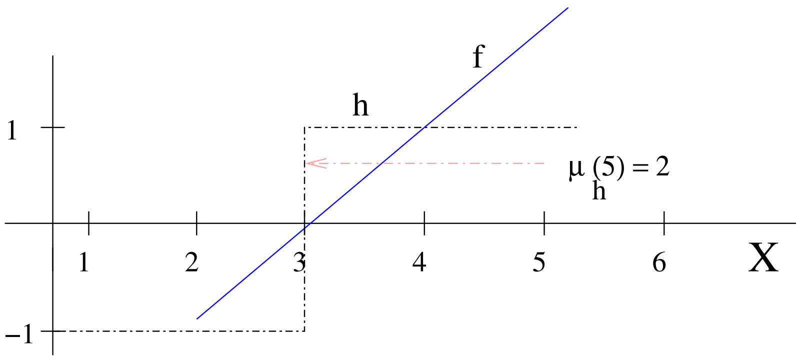


The margin $\mu_h(4) = 1$, $\mu_h(10) = 2$.

The sample margin $\mu_{S_m}(h) \equiv \min_{1 \le i \le m} \mu_h(x_i)$

Example

- $\mu_h(x)$ is natural extension of the real-value-based margin $M_f(x)$
- E.g., a linear function, $f = x - w$, on X gives $h_w(x) = \text{sgn}(f)$ which is a unit-step function



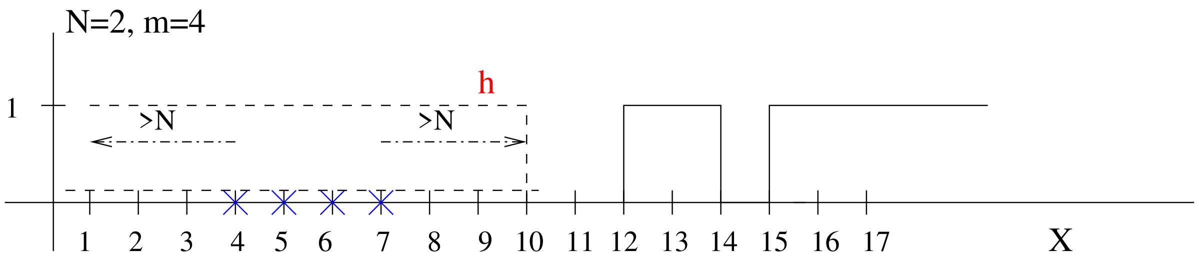
$$w = 3,\ M_f(5) = \mu_h(5) = 2.$$

- Both $M_f(x)$ and $\mu_h(x)$ correspond to the distance of x from the border.

Case II (revisited)

- Learning input Z: fixed parameter $N > 0$, sample S_m, information $I_t =$ 'there is an $\hat{h} \in \mathcal{H}$ with $\mu_{S_m}(\hat{h}) > N$ and $L(\hat{h}) \leq L^* + \epsilon_N$'
- Effective hypothesis class is $\mathcal{H}(Z) = \{h \in \mathcal{H} : \mu_{S_m}(h) > N\}$.
- For worst case sample S_m, we have

$$|\mathcal{H}(Z)| \leq \left(\frac{e(n - m - 2(N+1))}{d}\right)^d . \tag{2}$$



Comparison

- Comparing (1) and (2), $|\mathcal{H}(Z)|$ now decreases (polynomially) with N, i.e.,

$$|\mathcal{H}(Z)| \leq \sum_{k=0}^{d} \binom{n-m-2(N+1)}{k} \leq \left(\frac{e(n-m-2(N+1))}{d}\right)^{d}$$

versus

$$\sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{e(n-m)}{d}\right)^{d}.$$

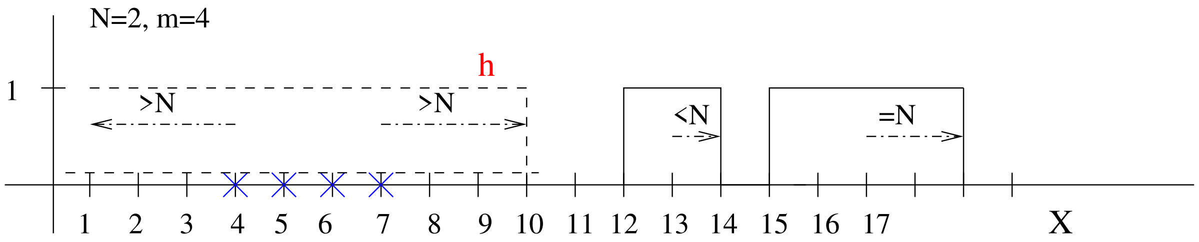
- Sauer's bound still applies in this case

Case III: 'maximal-sample' learning input

- Learning input Z: fixed parameter $N > 0$, sample S_m, information I_t = 'there is a $\hat{h}$ with $L(\hat{h}) \leq L^* + \epsilon_N$, and S_m is the *largest* sample such that $\mu_{S_m}(\hat{h}) > N$'

- Effective hypothesis class

$$\mathcal{H}(Z) = \{h \in \mathcal{H} : \mu_h(x) > N \text{ iff } x \in S_m\}. \tag{3}$$



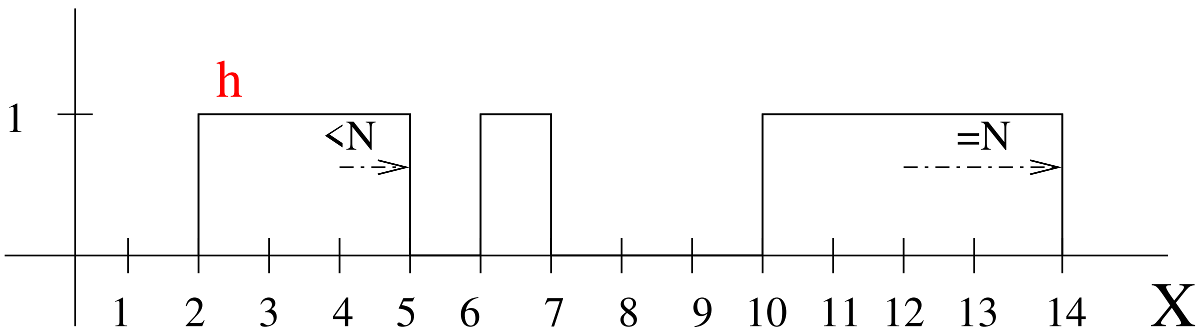
- $|\mathcal{H}(Z)| \leq$?

Need a ‘Sauer’-like lemma but with a margin-constraint on the functions

Counting binary-functions

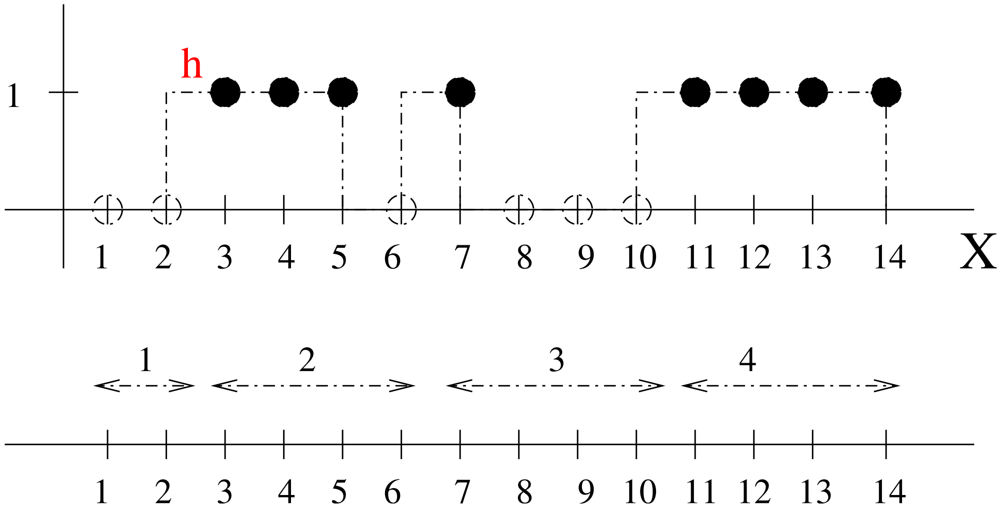
- Haussler & Long '95 extended Sauer's Lemma by generalizing the VC-dimension to other dimensions (see also Anthony & Bartlett '95, Ben-David,Cesa-Bianchi,Haussler & Long '95)
- They prove Sauer's result by counting binary-functions $h \in \mathcal{H}$ with at most k 1's, and summing over $1 \leq k \leq d$
- If we can count the number of h with k 1's and *also* satisfying $\mu_h(x) \leq N$, for all $x \in [n]$ then can extend their result

How do you count binary functions under margin constraint ?



2D ordered partitions of $[k, n-k]$

- Based on Ratsaby '04, view an h on $[n]$ that has k 1's as a 2D-ordered partition of $[k, n-k]$ into l pairs $[a_j, b_j]$, $1 \le j \le l$
- a_j is the number of 1's and b_j is the number of 0's in the j^{th} component



$l = 4$, where $[a_1, b_1] = [0, 2]$, $[a_2, b_2] = [3, 1]$, $[a_2, b_2] = [1, 3]$, $[a_4, b_4] = [4, 0]$

Main Result

Theorem Let $n, d, N \geq 1$. Let $\mathcal{H}$ be a class of binary-valued functions on $[n]$ satisfying $\mu_h(x) \leq N$ on all $x \in [n]$ and let $\mathrm{VC}(\mathcal{H}) \leq d$. Then

$$|\mathcal{H}| \leq \beta_d^{(N)}(n) \equiv \sum_{k=0}^{d} \sum_{r=1}^{n} c(k, n-k; r, N)$$

where

$$\begin{aligned} & c(k, n-k; r, N) \\ = \quad & \sum_{i,j=0}^{1} w_{r-i,2N}(k - r + 1 - i(2N+1)) \\ & \qquad w_{r-j,2N}(n - k - r + 1 - j(2N+1)) \end{aligned}$$

and

$$w_{r,N}(n) = \sum_{k=0}^{r} (-1)^k \binom{r}{k} \binom{n+r-1-k(N+1)}{r-1}$$

is the number of one-dimensional ordered partitions of n into r parts each $\leq N$.

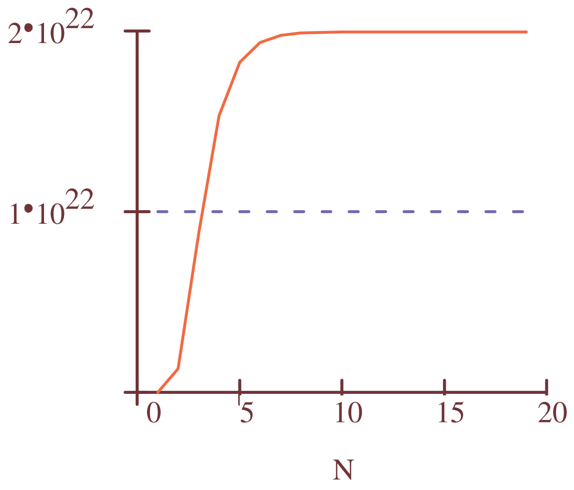
Note,

$$c(k, n-k; r, N) \approx \binom{k}{r-1} \binom{n-k}{r-1} \cdot \left(\left(1 - e^{-\frac{(r-1)(2N+1)}{k}}\right) \left(1 - e^{-\frac{(r-1)(2N+1)}{n-k}}\right) \right)^l.$$

Remark: For $N \geq d/(n-d)$, the upper bound is tight.

Properties of $\beta_d^{(N)}(n)$

- $\sum_{r=1}^{n} \binom{k}{r-1}\binom{n-k}{r-1} = \binom{n}{k}$
- $\beta_d^{(N)}(n) \leq \sum_{k=0}^{d} \binom{n}{k}$ ($\equiv$ Sauer's bound), converges to it with increasing N
- Non-linear form with respect to N
- E.g., for $d = 30$, $n = 80$



Threshold

- Define threshold $N^*(n,d)$ at the point at which $\beta_d^{(N)}(n)$ reaches half of its large-N limit value
- For $1 \le d \le n/2$,

$$N^*(n,d) \approx \left(\frac{n}{2d}\right) \ln \left(\frac{1}{\sqrt{1+\frac{2n\ln 2}{n+d(n-d)}}-1} \right) \le \left(\frac{n}{2d}\right) \ln \left(\frac{n}{8\ln 2}\right)$$

- 'Most' functions in $\mathcal{H}$ have a margin $N \ge N^*(n,d)$

$|\mathcal{H}(Z)|$ in case III

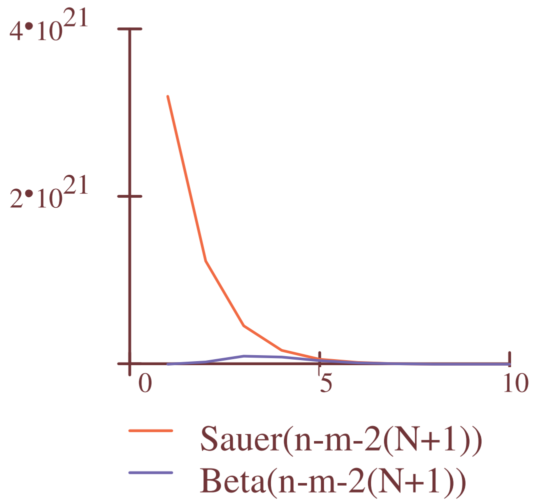
- We now have an estimate for the effective hypothesis class $\mathcal{H}(Z)$ in (3),

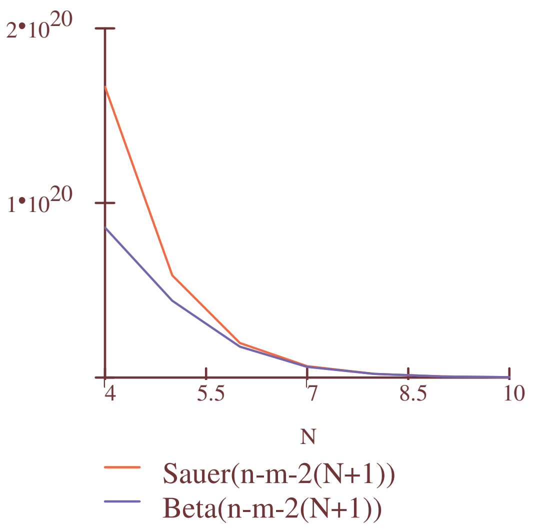
$$|\mathcal{H}(Z)| \leq \beta_d^{(N)}(n - m - 2(N+1)).$$

- For $N < N^*(n,d)$, it is significantly smaller than Sauer's bound (2) on $|\mathcal{H}(Z)|$, i.e.,

$$\beta_d^{(N)}(n - m - 2(N+1)) < \left(\frac{e(n - m - 2(N+1))}{d}\right)^d$$

while for $N \geq N^*(n,d)$ they are roughly the same, as n increases.





where $n = 90,\ d = 30,\ N^*(n, d) \approx 5.$

Interpretation

- Side-information I_t: for a fixed parameter $N > 0$, there is an ϵ-good $\hat{h}$ such that S_m is the largest sample $\mu_{S_m}(t) > N$
- Value of this Information: roughly, this I_t is significant compared to information of Case II only if $N < \left(\frac{n}{2d}\right) \ln\left(\frac{n}{8\ln 2}\right)$
- There are very few hypotheses which are 'complicated', i.e., not 'flat' on most of domain outside S_m
- So for instance if a maximal-margin learner happened to find a hypothesis $\hat{h}$ with a margin $N < N^*$ achieved only on the points of S_m then it is a 'rare' ϵ-good hypothesis
- Note: this result holds for any general class $\mathcal{H}$ with $VC(\mathcal{H}) \leq d$ on $[n]$

Further questions

- Is there a similar effect in the real-valued case ? (No ?)
- Consider other learning-inputs Z and compute $|\mathcal{H}(Z)|$
- Doing maximal-margin learning with the binary-margin function $\mu_S()$
- Compute the growth function on a sample S_m of the class $\mathcal{H}_N = \{h \in \mathcal{H} : \mu_{S_m}(h) > N\}$
- Replaces the covering number $\mathcal{N}(\mathcal{F}, \gamma, m)$ in the standard approach
- Obtain a bound on the generalization error in terms of N
- Efficient learning algorithm for finding a maximal margin hypothesis ?