CSE522, Winter 2011, Learning Theory	Lecture 8 - 01/27/2011

Vapnik-Chervonenkis Theory

Lecturer: Ofer Dekel — *Scribe: Amol Kapila*

1 Recap

1. With probability at least $1-\delta$, if $\ell \in [0,c]$, then $\forall h \in H$, $\ell(h;\mathcal{D}) \le \ell(h;S) + R_m(\ell \circ H) + c\sqrt{\frac{\log(1/\delta)}{2m}}$. A bound like this immediately implies a bound on the excess risk of the empirical risk minimizer. We prove this by proving a stronger, uniform bound on the excess risk across all $h \in H$.

2. With high probability, $\widehat{R}_m(\ell \circ H, S) \approx R_m(\ell \circ H)$, where

$$R_m(\ell \circ H) = \frac{2}{m}\mathbb{E}_S\mathbb{E}_\sigma \sup_{h\in H}\sum_{i=1}^m \sigma_i \ell(h;(x_i,y_i)).$$

   The empirical Rademacher complexity

$$\widehat{R}_m = \frac{2}{m}\mathbb{E}_\sigma \sup_{h\in H}\sum_{i=1}^m \sigma_i \ell(h;(x_i,y_i))$$

   is the same thing without the expectation over S.

3. In the case of binary classification ($\mathcal{Y} = \{1,-1\}$, ℓ = error indicator),

$$\widehat{R}_m(\ell \circ H, S) = 1 - 2\min_{h\in H}\ell(h;S'),$$

   where $S' = \{(x_i,\sigma_i)\}_{i=1}^m$ and $\sigma_i = \pm 1$ with probability 1/2 each.

4. If $h : X \to \mathbb{R}$, $\ell = \ell(yh(x))$ or $\ell(h(x)-y)$, and ℓ is λ-Lipschitz in $h(x)$, then $R_m(\ell \circ H) \le \lambda R_m(H)$. The same property holds for the empirical Rademacher average: $\widehat{R}_m(\ell \circ H, S) \le \lambda \widehat{R}_m(H,S)$.

5. Class of linear hypotheses with norm $\le B$: $H = \{h_w = \langle w,x\rangle \mid \|w\|_2 \le B\}$. In this case,

$$\widehat{R}_m(H,S) = \frac{2B}{m}\sqrt{\sum_{i=1}^m \|x_i\|_2^2}.$$

   If $\mathcal{D}$ is such that $\|x\| \le X$, then $R_m(H) \le 2BX/\sqrt{m}$.

6. If $\overline{H}$ is the convex hull of H, then $R_m(\overline{H}) = R_m(H)$. (Homework problem).

2 VC Theory

Binary Classification: $\mathcal{Y} = \{1,-1\}$, ℓ is the 0-1 loss (a.k.a., error indicator loss).
VC Theory is a combinatorial theory, based on discrete math.

Observation 1. *We only need to worry about $R_m(H)$, not $R_m(\ell \circ H)$, if we have 0-1 loss.*

Observation 2. *If S is a sample of m examples, then there are at most 2^m vectors of the form $(h(x_1), h(x_2), \ldots, h(x_m))$. We will explore how many ways can we label a concrete dataset.*

Fact $(e^{\alpha}+e^{-\alpha})/2 \le e^{\alpha^2/2}$. Proof by Taylor expansion of the exponential function.

Theorem 3. *(Massart's Finite Class Lemma) Suppose $A \subseteq \mathbb{R}^m$, $|A| < \infty$, and $\forall a \in A$, $\|a\|_2 \le \rho$. Then,*

$$\widehat{R}_m(H,S) = \frac{2}{m}\mathbb{E}_\sigma \max_{a\in A}\sum_{i=1}^m \sigma_i a_i \le \frac{2}{m}\rho\sqrt{2\log|A|}.$$

Here, each $a \in A$ is a vector of the form $a = (h(x_1), h(x_2), \ldots, h(x_m))$. So, if H can label our set in only a finite number of ways, then the empirical Rademacher average is bounded by the expression on the right-hand side of the inequality.

Proof. For each $s > 0$,

$$\begin{aligned}
\exp\left(s\mathbb{E}_\sigma \max_{a\in A}\sum_{i=1}^m \sigma_i a_i\right) &\le \text{[Jensen's inequality and the convexity of } \exp(\cdot)\text{]}\\
&\le \mathbb{E}\left(\exp\left(s\max_{a\in A}\sum_{i=1}^m \sigma_i a_i\right)\right)\\
&= \text{[monotonicity of } \exp(\cdot)\text{]}\\
&= \mathbb{E}_\sigma \max_{a\in A}\exp\left(s\sum_{i=1}^m \sigma_i a_i\right)\\
&= \mathbb{E}_\sigma \max_{a\in A}\prod_{i=1}^m \exp\left(s a_i \sigma_i\right)\\
&\le \mathbb{E}_\sigma \sum_{a\in A}\prod_{i=1}^m \exp(s a_i \sigma_i)\\
&= \text{[independence of } \sigma_i\text{'s]}\\
&= \sum_{a\in A}\prod_{i=1}^m \mathbb{E}_{\sigma_i}\exp\left(s a_i \sigma_i\right)\\
&= \sum_{a\in A}\prod_{i=1}^m \frac{e^{s a_i}+e^{-s a_i}}{2}\\
&\le \text{[fact stated above]}\\
&\le \sum_{a\in A}\prod_{i=1}^m \exp\left(\frac{(s a_i)^2}{2}\right)\\
&= \sum_{a\in A}\exp\left(\frac{s^2}{2}\|a\|^2\right)\\
&\le |A|\exp\left(\frac{s^2\rho^2}{2}\right).
\end{aligned}$$

Hence, we can conclude that

$$\mathbb{E}_\sigma \max_{a\in A}\sum_{i=1}^m \sigma_i a_i \le \frac{1}{s}\log\left(|A|\exp\left(\frac{s^2\rho^2}{2}\right)\right) = \frac{\log|A|}{s} + \frac{s\rho^2}{2}.$$

Plug in $s = \sqrt{2\log|A|}/\rho$ to get

$$\frac{2}{m}\mathbb{E}_\sigma \max_{a\in A}\sum_{i=1}^m \sigma_i a_i \le \frac{2}{m}\rho\sqrt{2\log|A|}.$$

$\square$

Observation 4. *So, we now have a bound on the empirical Rademacher average. Basically, to bound the empirical Rademacher average, we want to limit the size of* $|A|$.

Definition 5. *The* growth function of H *is defined as* $g_H(m) = \max_S |\{(h(x_1), \ldots, h(x_m))\}_{h \in H}|$. *Because we have a set, labelings do not get counted twice. Note that* $g_H(m) \leq 2^m$.

Fact 6. *We can restate the result in Theorem 3 in terms of the growth function as follows: If* H *is a hypothesis space of binary classifiers, then*

$$R(H) \leq \frac{2}{m}\sqrt{2 \log g_H(m)}\sqrt{m} = \frac{2}{\sqrt{m}}\sqrt{2 \log g_H(m)}.$$

So, for all S,

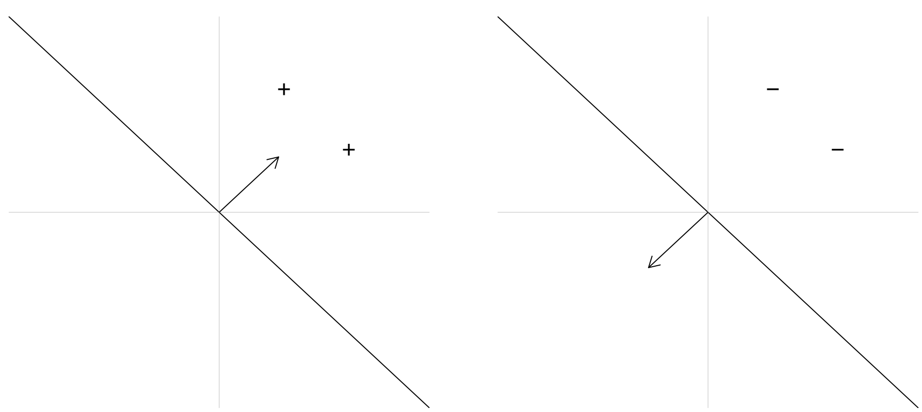
$$\widehat{R}(H, S) \leq 2\sqrt{\frac{2 \log g_H(m)}{m}}.$$

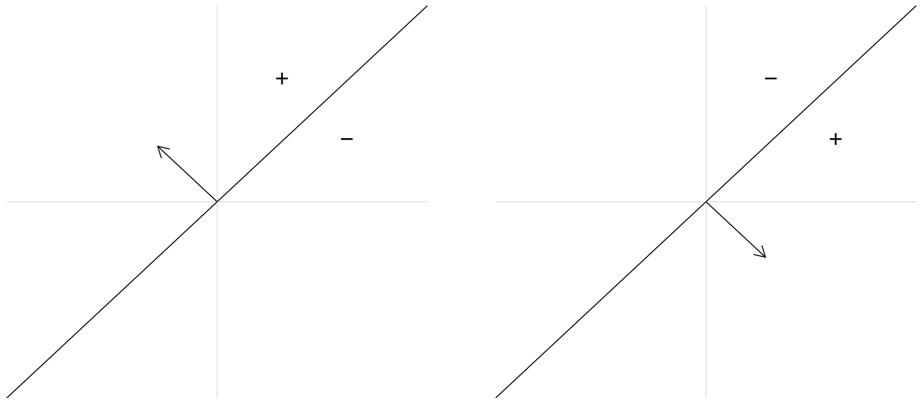
Observation 7. *If* $g_H(m) = 2^m$, *the bound is a constant, not diminishing as* $O(1/\sqrt{m})$.

3 Examples

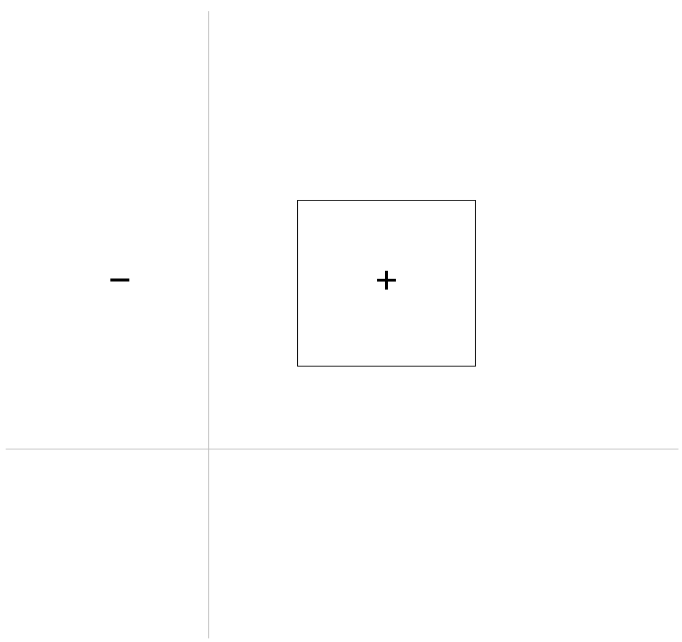
If H is a hypothesis class of binary classifiers, in how many different ways can H label S? This is moving from linear algebra to combinatorics.

Example 1 H = linear classifiers in $\mathbb{R}^2$. If $m = 2$, then $g_H(m) = 4 = 2^m$. The figures below provide the justification for this.



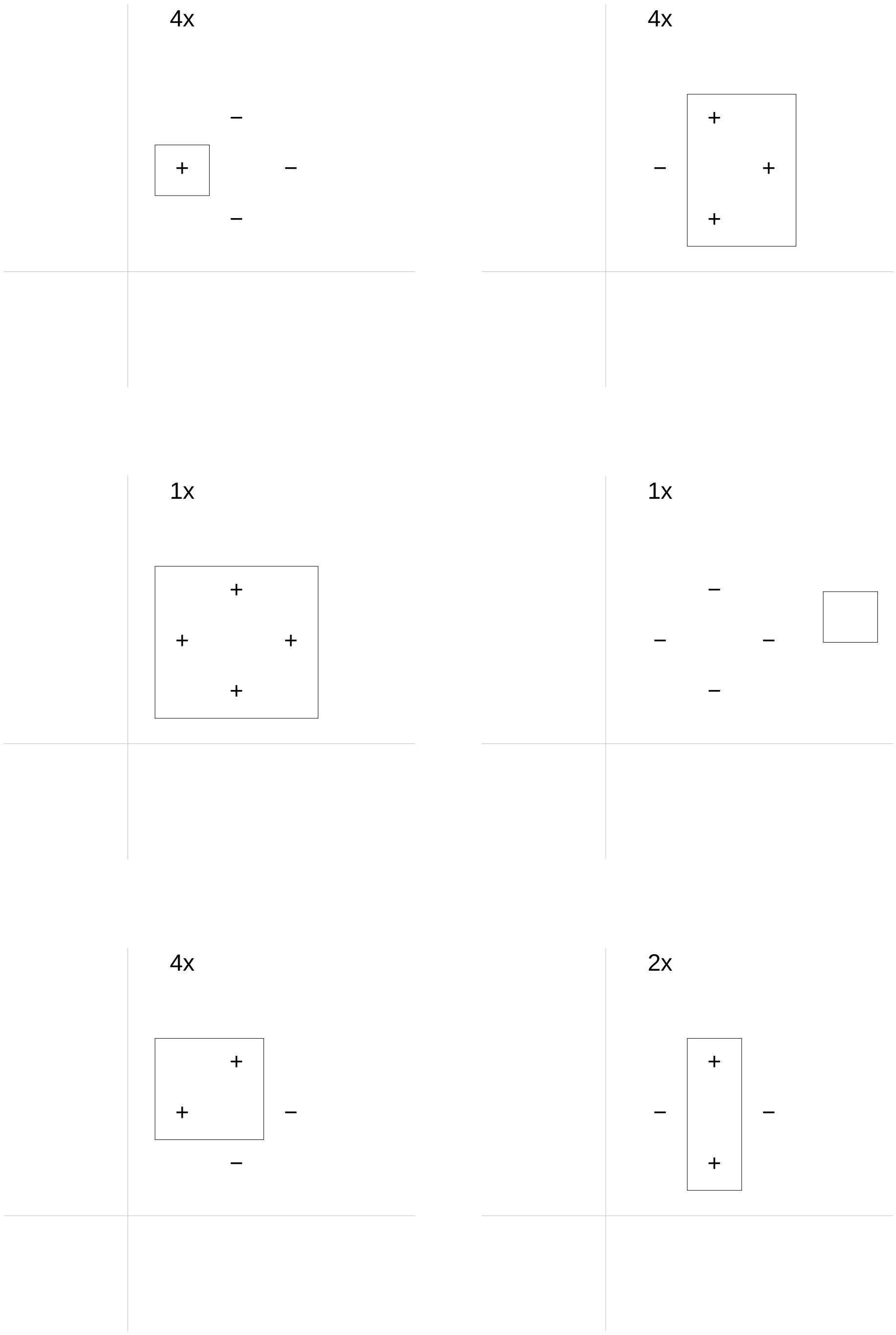


Example 2 H = axis-parallel boxes in $\mathbb{R}^2$.



An example of an axis-parallel box. Points inside the box are labeled positive, and points outside the box are labeled negative.

If $m = 1$, then clearly $g_H(m) = 2 = 2^m$. If $m = 4$, then $g_H(m) = 16 = 2^m$, as show using the figures below. Each figure abstractly represents one or more possible labelings (the multiplicity is shown as kx, where k is the multiplicity).



One can also show that $g_H(5) = 31 < 2^5$.

Definition 8. *If H can label S in all 2^m ways ($m = |S|$), then we say that H* shatters *S. So, we say that axis-parallel boxes shatter 4 points, but not 5.*

Definition 9. *The* VC Dimension *of a class H is $VCdim(H) = \max\{|S| \mid H \text{ shatters } S\}$.*

Example 1 H = intervals in $\mathbb{R}$. $g_H(1) = 2 = 2^1$. $g_H(2) = 4 = 2^2$. $g_H(3) < 2^3$, so H cannot shatter 3 points, as the example below shows.

A labeling of three points in $\mathbb{R}$ that cannot be generated by intervals in $\mathbb{R}$.

4 Useful Lemmas

Lemma 10. *(Sauer) Let H be a hypothesis class of binary classifiers with $VCdim(H) = d$. Then,*

$$g_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \Phi_d(m).$$

Lemma 11. *(Stirling)*

$$\Phi_d(m) \leq \left(\frac{em}{d}\right)^d.$$

Proof.

$$\begin{aligned} \left(\frac{d}{m}\right)^d \Phi_d(m) &= \left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \\ &\leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \\ &\leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} \\ &= \text{[Binomial Theorem]} \\ &= \left(1 + \frac{d}{m}\right)^m \\ &\leq e^d. \end{aligned}$$

Hence,

$$\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d.$$

□