# ECE 0402 - Pattern Recognition

Lecture 9 2/14/2022

**Reference reading:** Please also refer to Learning from Data 2.1.3 and 2.1.4 for further discussion of the topic discussed in this lecture note.

**Review:** We showed that for a given $\mathcal{H}$, if we know that $k$ is a **break point** (meaning that no data set of size k can be **shattered**), then the growth function is bounded by this sum:

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{k} \binom{n}{i} \implies \text{polynomial with leading term } n^k$$

Exponential growth vs Polynomial growth

- Positive rays ($k = 2$):

$$m_{\mathcal{H}}(n) = n + 1 \leq n + 1$$

- Positive intervals ($k = 3$):

$$m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1 \leq \frac{1}{2}n^2 + \frac{1}{2}n + 1$$

- Linear Classifiers in $\mathbb{R}^2$ (k=4):

$$m_{\mathcal{H}}(n) \leq \frac{1}{6}n^3 + \frac{5}{6}n + 1$$

Bottom line if is we have a set of classifier, all we need is a break point to exist–¿ growth function is a polynomial $n^{k-1}$

we can actually replace $| \mathcal{H} |$ with $m_{\mathcal{H}}(n)$ to obtain an inequality along the lines of

$$R(h^*) \leq \hat{R}_n(h^*) + \sqrt{\frac{1}{2n} \; log \; \frac{2m_{\mathcal{H}}(n)}{\delta}}$$

We won't be able to quite show this for technical reasons (which we will soon see). We will only be able to show that with probability $\geq 1 - \delta$

$$\boxed{R(h^*) \leq \hat{R}_n(h^*) + \sqrt{\frac{8}{n} \; log \; \frac{4m_{\mathcal{H}}(2n)}{\delta}}}$$

This is called **VC generalization bound**. VC stands for Vapnik and Chervonenkis who proved in 1971.

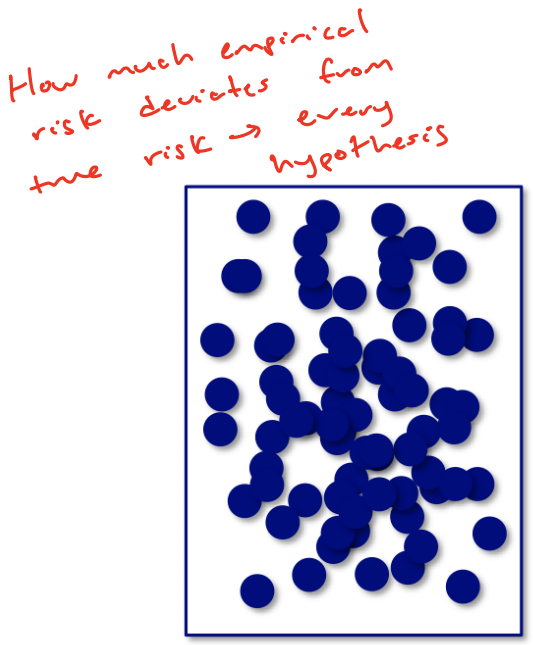



Figure 1: UB


Figure 2: VC

Mathematically, using Hoeffding's inequality together with a union bound, we were able to show that

$$\mathbb{P}[max_{h\in\mathcal{H}} \mid \hat{R}_n(h) - R(h) \mid > \epsilon] \leq \mid \mathcal{H} \mid \cdot 2e^{-2\epsilon^2 n}$$

What the VC bound gives us is a generalization of the form:

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - R(h) \mid > \epsilon\right] \leq 2 \cdot m_{\mathcal{H}}(2n) \cdot 2e^{-\frac{1}{8}\epsilon^2 n}$$

**Supremum**: The supremum of a set $\mathcal{S} \subset T$ is the least element of $T$ that is greater than of equal to all elements of $\mathcal{S}$. (This is sometimes called the least-upper-bound which probably makes a lot more sense). Here are several examples:

- $sup\{1, 2, 3\} = 3$
- $sup\{x : 0 \leq x \leq 1\} = 1$
- $sup\{x : 0 < x < q1\} = 1$
- $sup\{1 - 1/n : n > 0\} = 1$

The magic in the proof of the VC bound is to realize that we can relate the supremum over all $h \in \mathcal{H}$ to the maximum over a finite set of $h \in \mathcal{H}$ using a really cool trick!

### VC Bound

$$R(h) \leq \hat{R}_n(h) + \sqrt{\frac{8}{n}\, log\, \frac{4m_{\mathcal{H}}(2n)}{\delta}}$$

**Role of Growth Function**

We aim to get a bound on

$$\mathbb{P}\left[\mid \hat{R}_n(h) - R(h) \mid \ > \epsilon\right]$$

Don't want this to be a big deviation seen → unseen data

that hold for any $h \in \mathcal{H}$, i.e., a bound on

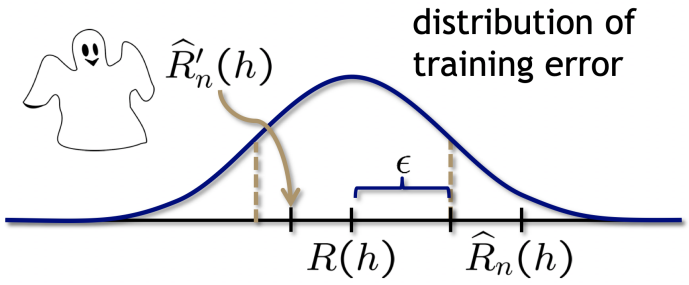
Maximum error hypothesis

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - R(h) \mid \ > \ \epsilon\right]$$

Perhaps it is not surprising that we can understand $\hat{R}_n(h)$ using growth function...

There may be infinitely many $h \in \mathcal{H}$, but $\mathcal{H}$ can only generate $m_{\mathcal{H}}(n)$ **unique dichotomies** for $n$ data points. Thus, empirical risk $\hat{R}_n(h)$ can only take finitely many – at most $m_{\mathcal{H}}(n)$ – different values. Unfortunately, $R(h)$ can still take infinitely many different values, and so there are infinitely many $\mid \hat{R}_n(h) - R(h) \mid$.

**Fundamental insight**: The key trick is to consider two different datasets!–this is for sake of analysis. We will imagine that in addition to our training data, we have access to a second independent dataset (of size $n$), which we call the **ghost sample**. Here the blue line is the




distribution of empirical risk, and $\hat{R'}_n(h)$ is the second estimate of $R(h)$.

Can we relate $\mathbb{P}\left[\mid \hat{R}_n(h) - R(h) \mid \ > \epsilon\right]$ to some $\mathbb{P}\left[\mid \hat{R'}_n(h) - \hat{R}_n(h) \mid \ > \epsilon\right]$?

Suppose (for the moment) that the empirical estimates $\hat{R}_n(h)$ and $\hat{R'}_n(h)$ are RVs that are drawn from a symmetric distribution with mean $R_n(h)$.

Consider the following events:

- A: the event that $\mid \hat{R}_n(h) - R(h) \mid \ > \epsilon$
- B: the event that $\mid \hat{R}_n(h) - \hat{R'}_n(h) \mid \ > \epsilon$

**Claim**: $\mathbb{P}\left[B|A\right] \geq \frac{1}{2}$

Because of this symmetry assumption: the probability $B|A$ bigger than a half. So why is that true?

$$\text{Thus } \mathbb{P}[B] \geq \mathbb{P}\left[B|A\right] \cdot \mathbb{P}[A] \geq \tfrac{1}{2}\mathbb{P}[A]$$

$$\implies \mathbb{P}\left[\mid \hat{R}_n(h) - R(h) \mid \ > \epsilon\right] \leq 2\ \mathbb{P}\left[\mid \hat{R}_n(h) - \hat{R'}_n(h) \mid \ > \epsilon\right]$$

Unfortunately the distribution of $\hat{R}_n(h)$ and $\hat{R'}_n(h)$ is not binomial (not symmetric) so this exact statement doesn't hold in general, but the intuition is valid.

Instead, we have the following bound:

**Lemma 1 (Ghost sample)**

$$\begin{aligned} &\mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - R(h) \mid \ > \ \epsilon\right] \\ \leq\ &2\mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - \hat{R'}_n(h) \mid \ > \ \frac{\epsilon}{2}\right] \end{aligned}$$

We wanna understand worse-case deviation between our empirical risk and the true risk. But rather than analyzing that directly, we are gonna instead upper-bound that by looking at the worst case deviation between pairs of two empirical estimates.

**Lemma 2 (Where the magic happens)**

$$\begin{aligned} &\mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - \hat{R'}_n(h) \mid \ > \ \frac{\epsilon}{2}\right] \\ \leq\ &m_{\mathbb{H}}(2n) \cdot \sup_{\mathcal{S}} \mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - \hat{R'}_n(h) \mid \ > \ \frac{\epsilon}{2}\middle|\mathcal{S}\right] \end{aligned}$$

**Lemma 3**

For any fixed classifier $h \in \mathcal{H}$ and **any** fixed set of $2n$ data points $\mathcal{S}$,

$$\mathbb{P}\left[\mid \hat{R}_n(h) - \hat{R'}_n(h) \mid \ > \ \frac{\epsilon}{2}|\mathcal{S}\right] \leq 2e^{-\frac{\epsilon^2 n}{8}}$$

where again the probability is w.r.t. a random partitioning of **any** $\mathcal{S}$ into 2 training sets of size $n$.

Putting all of this together, we get

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \mid \hat{R}_n(h) - R(h) \mid \ > \ \epsilon\right] \leq 2\ m_{\mathcal{H}}(2n)\ e^{-\frac{\epsilon^2 n}{8}}$$

This was the result we were after. We can also state this as confidence bounds. For any $h \in \mathcal{H}$, we have that with probability $\geq 1 - \delta$

$$R(h) \leq \hat{R}_n(h) + \sqrt{\frac{8}{n}\ log\frac{4m_{\mathcal{H}}(2n)}{\delta}}$$

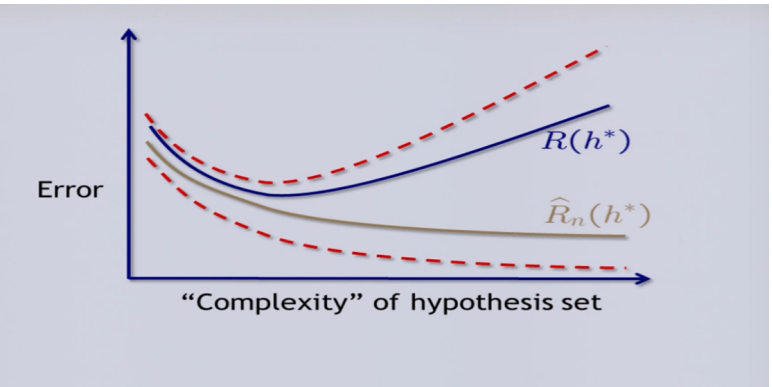



Figure 3: VC Bound

We showed that if $k$ is a break point for $\mathcal{H}$ then $m_{\mathcal{H}}(n) \leq \sum_{i=0}^{k-1} \binom{n}{i} \leq n^{k-1} + 1$.

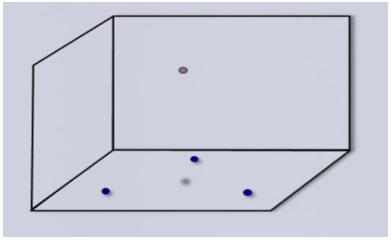
$$\implies R(h) \leq \hat{R}_n(h) + \sqrt{\frac{8}{n} \log \frac{4((2n)^{k-1}+1)}{\delta}}$$
$$\lessapprox \hat{R}_n(h) + \sqrt{\frac{8(k-1)}{n} \log \frac{8n}{\delta}}$$

**Definition:** The **VC dimension** of $\mathcal{H}$ is the largest $n$ for which $m_{\mathcal{H}}(n) = 2^n$. The denotation for the VC dimension of a hypothesis set $\mathcal{H}$ is, $d_{VC}(\mathcal{H})$. In other words, $d_{VC}(\mathcal{H})$ is the maximum number of points that our hypothesis shatters. In other other words, $d_{VC}(\mathcal{H})$ is 1 less than the smallest break point.

$$R(h) \lessapprox \hat{R}_n(h) + \sqrt{\frac{8 d_{VC}}{n} \log \frac{8n}{\delta}}$$

**Examples:**

- Positive rays: $d_{VC} = 1$.
- Positive intervals: $d_{VC} = 2$.
- Convex sets: $d_{VC} = \infty$.
- Linear classifiers in $\mathbb{R}^2$ : $d_{VC} = 3$.
- How about linear classifier in $\mathbb{R}^3$?



So the fact that the linear classifiers had a break point of $k = 4$ was only due to the fact that we were only looking at $\mathbb{R}^2$. In the higher dimensions it changes.

In general $d_{VC} = d + 1$ for linear classifiers. You can prove that by showing $d_{VC} \leq d + 1$ and $d_{VC} \geq d + 1$, usual tricks! If you haven't seen, this is very often and the easiest ways to proof things of same sort.

how many parameters does a linear classifier in $\mathbb{R}^d$ have?

$$\begin{aligned} w &\in \mathbb{R}^d \\ b &\in \mathbb{R} \implies d + 1 \text{ parameters} \end{aligned}$$

Is this a coincidence? Let;s look at our other examples:

- Positive rays:
  - $d_{VC} = 1$
  - 1 parameter
- Positive intervals:
  - $d_{VC} = 2$
  - 2 parameters
- Convex sets:
  - $d_{VC} = \infty$
  - as many as you want

So VC dimension is basically the effective number of parameters, meaning that additional parameters do not always contribute additional degrees of freedom. You can introduce parameters that do nothing, that have no actual effect on the VC dimension. As an example of this: take the output of a linear classifier, and then feed this into another linear classifier

$$y_i = sign(\omega'(sign(\theta^T x_i) + b')$$

This is adding no additional degrees of freedom.

**How big does our training set need to be?**

$$R(h) \lessapprox \hat{R}_n(h) + \sqrt{\frac{8 d_{VC}}{n} \log \frac{8n}{\delta}}$$

Just to see how this bound behaves, we can ignore the constants and look at:

$$\epsilon \sim \sqrt{\frac{d_{VC}}{n} \log n}$$

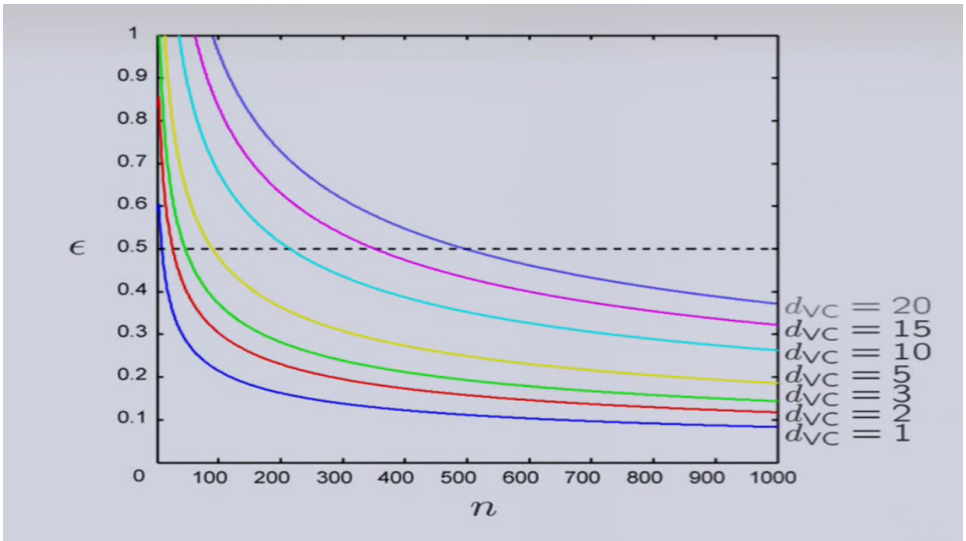



Figure 4: VC tightness versus data size

**Rule of thumb:** $n \geq 10 d_{VC}$