Machine Learning - B
Lectures 4
Lecture 4: VC Analysis

Amartya Sanyal

University of Copenhagen

Generalization Bounds: Finite Hypothesis Class

- **Lower bound:** If $|\mathcal{H}| = 2^{2n}$, then for some distribution $p(x, y)$, we have
$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \frac{1}{8}\right) \geq \frac{1}{8}.$$
- **Upper bound:** If $|\mathcal{H}| = M$, then with probability at least $1 - \delta$,
$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\log(M/\delta)}{2n}}\right) \leq \delta.$$
- The $\log M$ term is the **price of selection**; we require $M \ll e^n$.

Generalization Bounds: Countable Hypothesis Class

- Let $\mathcal{H}$ be countable.
- Fix a prior $\pi(h)$ over $\mathcal{H}$ such that $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$.
- Then with probability at least $1 - \delta$, for all $h \in \mathcal{H}$:

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\log(1/(\pi(h)\delta))}{2n}}.$$

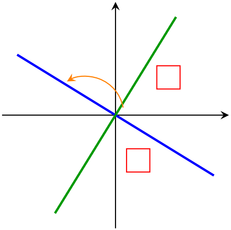
- This is a non-uniform bound – the **price of selection** now depends on $\log(1/\pi(h))$.

Motivation for VC Analysis

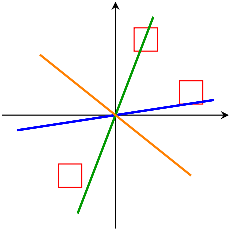
- What if $\mathcal{H}$ is uncountably infinite?
- Any selection from a finite sample S can only consider finitely many distinct labelings.
- VC theory bounds the **effective number of labelings** possible on a dataset.
- This leads to generalization guarantees without requiring a finite hypothesis class.

Learning by selection from uncountably infinite $\mathcal{H}$: VC analysis

- Given a finite S, we can only make finite selection based on S. The remaining (uncountably infinite) choices introduce no further bias.
- Example: homogeneous linear separators in $\mathbb{R}^2$.



$M(2) = 4 = 2^2$



$M(3) = 6 < 2^3$

Dichotomies

- Suppose we fix a finite sequence of instances $x_1, \ldots, x_n \in \mathcal{X}$.
- For a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, define the set of labelings realized by $\mathcal{H}$ on this sequence:

$$\Pi_{\mathcal{H}}(x_1, \ldots, x_n) = \{\, (h(x_1), \ldots, h(x_n)) : h \in \mathcal{H} \,\}.$$

- This is the set of dichotomies that $\mathcal{H}$ can realize on $x_1, \ldots, x_n$.
- The number of distinct labelings is denoted by

$$|\Pi_{\mathcal{H}}(x_1, \ldots, x_n)|.$$

Growth Function

Definition

The *growth function* of a hypothesis class $\mathcal{H}$ is defined as

$$m_{\mathcal{H}}(n) := \max_{x_1,\ldots,x_n \in \mathcal{X}} |\Pi_{\mathcal{H}}(x_1, \ldots, x_n)| .$$

- $m_{\mathcal{H}}(n)$ is the maximum number of labelings $\mathcal{H}$ can realize on any n-point subset of $\mathcal{X}$.
- Clearly, $m_{\mathcal{H}}(n) \leq 2^n$.
- If $m_{\mathcal{H}}(n) = 2^n$, then $\mathcal{H}$ **shatters** some set of n points.

VC Dimension

Definition

The *VC dimension* of $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ is the largest integer d such that

$$m_{\mathcal{H}}(d) = 2^d.$$

That is, there exists a set of d points shattered by $\mathcal{H}$, but no set of size $d+1$ is shattered.

- To show $\mathrm{VCdim}(\mathcal{H}) \geq d$: find a set of d points where $\mathcal{H}$ realizes all 2^d labelings.
- To show $\mathrm{VCdim}(\mathcal{H}) < d+1$: prove no set of size $d+1$ can be shattered by $\mathcal{H}$.
- Both directions are required!

Sauer's Lemma

Theorem (Sauer's Lemma)

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a class with VC dimension d. Then for all $n \geq d$,

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^{d}.$$

- The number of realizable dichotomies grows **polynomially** in n, not exponentially, once $n > d$.
- Hence, a finite VC dimension implies the hypothesis class is not "too rich" to generalize from data.

The VC Generalization Bound (Clean Notation)

Theorem (VC Generalization Bound)

For any $\mathcal{H}$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:

$$L(h) \;\le\; \widehat{L}(h,S) + \sqrt{\frac{8\ln(2m_{\mathcal{H}}(2n)) + \ln(1/\delta)}{n}}.$$

- The bound is interesting when $m_{\mathcal{H}}(2n) \ll 2^{2n}$.
- **Idea:** Introduce a "ghost sample" $S' = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$
- **Symmetrization:** Bound the deviation betweentwo empirical errors:

$$\sup_{h\in\mathcal{H}} \left(L(h) - \widehat{L}(h,S)\right) \;\le\; \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \left(\widehat{L}(h,S') - \widehat{L}(h,S)\right)\right].$$

- **Projection argument:** For a fixed pair of datasets S, S', the set

$$\left\{\widehat{L}(h,S') - \widehat{L}(h,S) : h \in \mathcal{H}\right\}$$

  is finite if $\mathcal{H}$ induces finitely many labelings on $2n$ points.
- **Conclusion:** Apply a union bound over $m_{\mathcal{H}}(2n)$ labelings and use Hoeffding's inequality to get the final VC bound.

Proof Idea: Symmetrization

- We want to bound the event:

$$\exists h \in \mathcal{H} \text{ such that } L(h) - \hat{L}(h, S) > \varepsilon$$

- Introduce a **ghost sample** $S' \sim D^n$ independent of S.
- Then, via symmetrization:

Lemma (Symmetrisation)

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) - \hat{L}(h, S) > \varepsilon\right) \leq 2 \cdot \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) > \frac{\varepsilon}{2}\right)$$

- Reduces generalization gap to deviation between two empirical averages.

Symmetrization Lemma

Lemma (Symmetrisation)

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) > \varepsilon\right) \leq 2 \cdot \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) > \frac{\varepsilon}{2}\right)$$

$$\begin{aligned}
&\mathbb{P}\left(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\right) \\
\geq\ &\mathbb{P}\left(\left\{\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\right\} \cap \underbrace{\left\{\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\right\}}_{\text{event A}}\right) \\
=\ &\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\right) \cdot \\
&\quad \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2} \,\middle|\, \mathbf{A}\right)
\end{aligned}$$

Symmetrization Lemma

Lemma (Symmetrisation)

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) - \hat{L}(h, S) > \varepsilon\right) \leq 2 \cdot \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) > \frac{\varepsilon}{2}\right)$$

- **Proof:**

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2}\right)$$

$$\geq \mathbb{P}\left(\left\{\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2}\right\} \cap \underbrace{\left\{\exists h \in \mathcal{H} : L(h) - \widehat{L}(h, S) \geq \varepsilon\right\}}_{\text{event A}}\right)$$

$$= \mathbb{P}\left(\exists h \in \mathcal{H} : L(h) - \widehat{L}(h, S) \geq \varepsilon\right) \cdot \mathbb{P}\left(\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2} \,\middle|\, \mathbf{A}\right)$$

Proof – continued

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2} \;\middle|\; \exists h \in \mathcal{H} : L(h) - \hat{L}(h, S) \geq \varepsilon\right)$$

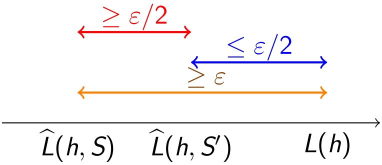
$$\geq \mathbb{P}\left(\hat{L}(h^*, S') - \hat{L}(h^*, S) \geq \frac{\varepsilon}{2} \;\middle|\; L(h^*) - \hat{L}(h^*, S) \geq \varepsilon\right) \qquad \text{(fix such } h^*\text{)}$$

$$\geq \mathbb{P}\left(L(h^*) - \hat{L}(h^*, S') \leq \frac{\varepsilon}{2} \;\middle|\; L(h^*) - \hat{L}(h^*, S) \geq \varepsilon\right) \qquad \text{(see figure)}$$

$$= \mathbb{P}\left(L(h^*) - \hat{L}(h^*, S') \leq \frac{\varepsilon}{2}\right) \qquad \text{(independence of } S, S'\text{)}$$

$$\geq 1 - \mathbb{P}\left(L(h^*) - \hat{L}(h^*, S') > \frac{\varepsilon}{2}\right) \geq 1 - \exp\left(-2n\left(\frac{\varepsilon}{2}\right)^2\right) \quad \text{(Hoeffding's inequality)}$$

$$\geq \frac{1}{2} \qquad \text{(by assumption for large } n\text{: } e^{-n\varepsilon^2/2} \leq \tfrac{1}{2}\text{)}$$



Projection Lemma

Lemma

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2}\right) \leq m_{\mathcal{H}}(2n) \cdot e^{-\frac{n\varepsilon^2}{8}}$$

Two ways to generate S, S':

- (1) Sample $S \sim \mathcal{D}^n$, then $S' \sim \mathcal{D}^n$
- (2) Sample $S \cup S' \sim \mathcal{D}^{2n}$, then split randomly into S and S'

$$\begin{aligned}
&\mathbb{P}\left(\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2}\right) \\
&= \sum_{S \cup S'} \underbrace{\mathbb{P}(S \cup S')}_{\text{sampling}} \cdot \underbrace{\mathbb{P}_{\text{split}}\left(\exists h : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2} \,\middle|\, S \cup S'\right)}_{\text{random split}} \\
&\leq \sup_{S \cup S'} \mathbb{P}_{\text{split}}\left(\exists h : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2} \,\middle|\, S \cup S'\right) \\
&\leq m_{\mathcal{H}}(2n) \cdot \sup_{S \cup S', h} \mathbb{P}_{\text{split}}\left(\widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2} \,\middle|\, S \cup S'\right).
\end{aligned}$$

Projection Tree

Projection – continued

$$\mathbb{P}_{\text{split}}\left(\widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2} \;\middle|\; S \cup S'\right)$$

$$= \mathbb{P}_{\text{split}}\left(\underbrace{\widehat{L}(h, S') - \frac{\widehat{L}(h, S) + \widehat{L}(h, S')}{2}}_{\text{centered around expectation}} \geq \frac{\varepsilon}{4} \;\middle|\; S \cup S'\right)$$

$$\leq e^{-2n(\frac{\varepsilon}{4})^2} = e^{-n\varepsilon^2/8}$$

Hoeffding's inequality for sampling without replacement

Putting It All Together

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \widehat{L}(h, S) + \varepsilon\right) \leq 2 \cdot \mathbb{P}\left(\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2}\right)$$

$$\leq 2 \cdot m_{\mathcal{H}}(2n) \cdot e^{-\frac{n\varepsilon^2}{8}}$$ (Sauer's + Hoeffding for sampling without replacement)

$$\text{Set this equal to } \delta \implies \varepsilon = \sqrt{\frac{8 \ln\left(\frac{2m_{\mathcal{H}}(2n)}{\delta}\right)}{n}}$$

- The bound is useful if $m_{\mathcal{H}}(2n) \ll 2^{2n}$
- Finite VC dimension implies polynomial growth of $m_{\mathcal{H}}(n)$

Generalization Bound via VC Dimension

From the previous slide:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \widehat{L}(h,S) + \varepsilon\right) \leq 2 \cdot m_{\mathcal{H}}(2n) \cdot e^{-\frac{n\varepsilon^2}{8}}$$

If $\mathcal{H}$ has VC dimension d, then by Sauer's lemma: $m_{\mathcal{H}}(2n) \leq \left(\frac{2en}{d}\right)^d$
Plugging this in:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \widehat{L}(h,S) + \varepsilon\right) \leq 2\left(\frac{2en}{d}\right)^d \cdot e^{-\frac{n\varepsilon^2}{8}}$$

Solving for ε such that the RHS is $\leq \delta$, we get:

$$\varepsilon = \sqrt{\frac{8}{n}\left(d \log\left(\frac{2en}{d}\right) + \log\left(\frac{2}{\delta}\right)\right)}$$

Conclusion: With probability at least $1-\delta$, the following holds for *all* $h \in \mathcal{H}$:

$$L(h) \leq \widehat{L}(h,S) + \sqrt{\frac{8}{n}\left(d \log\left(\frac{2en}{d}\right) + \log\left(\frac{2}{\delta}\right)\right)}$$

This uniform convergence result underpins what we next formalise as 'PAC'.