Shattering and VC Dimension

Agnostic Learnability of Function Classes

- Consider function class $\mathcal{F} : \mathcal{X} \rightarrow \{0, 1\}$, and Risk function *R*

$$R(f) = E(\ell(X, Y, f(X))$$

$$R_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i)).$$

$$f_n := \operatorname*{argmin}_{f \in \mathcal{F}} \mathrm{R}_{\text{emp}}(f).$$

Consistency of ERM

$$f_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ R(f).$$

$$P(R(f_n) - R(f_{\mathcal{F}}) > \varepsilon) \to 0 \quad as\ n \to \infty.$$

Uniform Convergence

Theorem 3 (Vapnik and Chervonenkis) *Uniform convergence*

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \to 0 \text{ as } n \to \infty, \tag{12}$$

for all $\epsilon > 0$, *is a* necessary and sufficient *condition for consistency of empirical risk minimization with respect to* $\mathcal{F}$.

Uniform Convergence for Finite Hypothesis Class

$$P(|R_{\text{emp}}(f) - R(f)| \geq \epsilon) \leq 2\exp(-2n\epsilon^2).$$

$$\begin{aligned}
&P(\sup_{f\in\mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon) \\
&= P\Big(|R(f_1) - R_{\text{emp}}(f_1)| \geq \varepsilon \ \text{ or } \ |R(f_2) - R_{\text{emp}}(f_2)| \geq \varepsilon \ \text{ or } \ \ldots \ \text{ or } \ |R(f_m) - R_{\text{emp}}(f_m)| \geq \varepsilon \Big) \\
&\leq \sum_{i=1}^{m} P(|R(f_i) - R_{\text{emp}}(f_i)| \geq \varepsilon) \\
&\leq 2m\exp(-2n\varepsilon^2) \qquad (15)
\end{aligned}$$

Applying the union bound

- *m* is the size of the function class

Capacity of Infinite Hypothesis Classes

- Infinite hypothesis classes can also be PAC learnable

Restrictions of Function Classes to a Finite Sample

$$Z_n := ((X_1, Y_1), \ldots, (X_n, Y_n))$$

$\mathcal{F}_{Zn}$ - Restriction to Z_n

$|\mathcal{F}_{Z_n}|$ be the cardinality of $\mathcal{F}$ when restricted to $\{X_1, \ldots, X_n\}$

Even if the original function class is infinite the restriction is finite.
The maximum cardinality the restriction can have is 2^n

Shattering Coefficient

$$\mathcal{N}(\mathcal{F}, n) = \max\{|\mathcal{F}_{Z_n}| \mid X_1, ..., X_n \in \mathcal{X}\}.$$

Number of ways that the function space can separate the patterns into two classes

$$\mathcal{N}(\mathcal{F}, n) = 2^n, \qquad Z_n \text{ is shattered}$$

- Shattering means that there exists a sample of *n* patterns which can be separated in all possible ways

- it does not mean that this applies to all possible samples of *n* patterns.

- Measure of capacity of function class

Example

Example 1: $\mathcal{X} = \mathbb{R}$, $\mathcal{F}$ as below (positive class = right half-space)

Generalization bound with Shattering Coeff.

Let $\mathcal{F}$ be any arbitrary function class. Then for all $0 < \varepsilon < 1$,

$$\Pr(\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| > \varepsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\varepsilon^2/4).$$

The other way round: With probability at least $1 - \delta$, all functions $f \in \mathcal{F}$ satisfy

$$R(f) \leq R_n(f) + 2\sqrt{\frac{\log(\mathcal{N}(\mathcal{F}, 2n)) - \log(\delta)}{n}}.$$

Symmetrization Lemma

- By R_n we denote the risk on our given sample of n points.
- By R'_n we denote the risk that we get on a second, independent sample of n points, called the "ghost sample".

Proposition 40 (Symmetrization lemma)

$$\begin{aligned}
&\Pr(\sup_{f\in\mathcal{F}} |R(f) - R_n(f)| > \varepsilon) \\
&\leq \quad 2\Pr(\sup_{f\in\mathcal{F}} |R_n(f) - R'_n(f)| > \varepsilon/2).
\end{aligned}$$

Proof of Generalization Bound

$$
\begin{aligned}
& P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \\
& \quad \textit{(due to symmetrization)} \\
& \leq 2P(\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \epsilon/2) \\
& \quad \textit{(only functions in } \mathcal{F}_{Z_{2n}} \textit{ are important)} \\
& = 2P(\sup_{f \in \mathcal{F}_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f))| > \epsilon/2) \\
& \quad \textit{(} \mathcal{F}_{Z_{2n}} \textit{ contains at most } \mathcal{N}(\mathcal{F}, 2n) \textit{ functions, independently of } Z_{2n}\textit{)} \\
& \quad \textit{(use union bound argument and Chernoff)} \\
& \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\varepsilon^2/4)
\end{aligned}
$$

Generalization Bound

$$R(f) \leq R_{\mathrm{emp}}(f) + \sqrt{\frac{4}{n}\big(\log(2\mathcal{N}(\mathcal{F}, n)) - \log(\delta)\big)}.$$

VC Dimension

We say that a sample Z_n of size n is *shattered by function class* $\mathcal{F}$ if the function class can realize any labeling on the given sample, that is $|\mathcal{F}_{Z_n}| = 2^n$. The *VC dimension of* $\mathcal{F}$, denoted by $\mathrm{VC}(\mathcal{F})$, is now defined as the largest number n such that there exists a sample of size n which is shattered by $\mathcal{F}$. Formally,

$$\mathrm{VC}(\mathcal{F}) = \max\{n \in \mathbb{N} \mid |\mathcal{F}_{Z_n}| = 2^n \text{ for some } Z_n\}.$$

VC Dimension Examples

- Half planes
- Axis parallel rectangles
- Convex d cornered polygons
- Sine waves