1 VC Dimension

Definition 1.0.1. Let X be a space and $C \subset \mathcal{P}(X)$ a set of subsets. Call these *concepts* and C a *concept class*. Then for $S \subset X$ we say that C *shatters* S if,

$$C \cap S := \{c \cap S \mid c \in C\} = \mathcal{P}(S)$$

meaning that every subset of S is of the form $c \cap S$ for some $c \in C$.

Definition 1.0.2. If C is a concept class. Then *Vapnik–Chervonenkis* is,

$$\text{VC}(C) = \sup\{|S| \mid C \text{ shatters } S\}$$

Definition 1.0.3. For each $d \in \mathbb{N}$ we define the polynomial, Φ_d as follows,

$$\begin{aligned} \Phi_d(0) &= 1 \\ \Phi_d(m) &= \Phi_d(m-1) + \Phi_{d-1}(m-1) \end{aligned}$$

Definition 1.0.4. Let C be a concept class. Then,

$$\Pi_C(m) = \sup\{|S \cap C| \mid |S| = m\}$$

Remark. By definition, $\Pi_C(m) \le 2^m$ and

$$\Pi_C(m) = 2^m \iff m \le \text{VC}(C)$$

We use the notation $C|_S := C \cap S$ and $c|_S := C \cap S$.

Proposition 1.0.5. Let C be a concept class with $\text{VC}(C) = d$ then $\Pi_C(m) \le \Phi_d(m)$.

Proof. We prove this by complete induciton on m, d. First, for the case $m = 0$ we have,

$$\Phi_C(m) \le 2^0 = 1 = \Phi_d(0)$$

For the case $d = 0$, no nontrivial set can be shattered meaning that C must contain only one concept so $\Phi_C(m) = 1$ and also $\Phi_0(m) = 1$.

Now we assume the theorem for all $d' \le d$ and $m' \le m$ with at least one strict. Let S be a set with $|S| = m$ achieving $\Pi_C(m)$. Choose $x \in S$,

$$|(S \backslash \{x\}) \cap C| \le \Pi_C(m-1) \le \Phi_d(m-1)$$

We define a new concept class on S,

$$C' = \{c \in C|_S \mid x \notin c \text{ and } \exists \tilde{c} \in C : x \in \tilde{c} \text{ and } c|_{S \backslash \{x\}} = \tilde{c}|_{S \backslash \{x\}}\}$$

which is the class of concepts who have a modification on x with the same restriction to $S \backslash \{x\}$. Then,

$$|C \cap S| = |C \cap (S \backslash \{x\})| + |C' \cap (S \backslash \{x\})|$$

I claim $\text{VC}(C') \le d - 1$. Indeed, for any $S' \subset S$ shattered by C' notice that $x \notin S'$ since $x \notin c$ for all $c \in C'$. Thus $S' \subset S \backslash \{x\}$. Then, by definition $S' \sup\{x\}$ is shattered by C so $|S'| \le d - 1$. Therefore,

$$|C \cap S| \le \Phi_d(m-1) + \Phi_{d-1}(m-1) = \Phi_d(m)$$

proving the result by induction. □

2 PAC Learning

We will work with learning contexts where we are trying to learn binary classification on space X equipped with a measure μ representing the distribution of data. Explicitly, we are trying to learn one of a given set of hypotheses for where in X the data is clustered according to μ. This means we are given a set of concepts C consisting of subsets of C and a loss function $\ell(x, c)$ measuring how poorly our hypothesis about the data

Usually we think of learning in terms of trying to learn a function $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is a space of observables and $\mathcal{Y}$ is a space of labels we are trying to predict best with a concept class of possible functions C. Then there is a measure μ on $\mathcal{X} \times \mathcal{Y}$ which represents how the real data and labels are distributed. Note that we can think of the functions C as subsets of $X := \mathcal{X} \times \mathcal{Y}$ and the loss of $h : \mathcal{X} \to \mathcal{Y}$ on a particular example (x, y) means that we have a loss function on $X \times C$. Therefore, we have reduced to the binary classification problem.

Definition 2.0.1. A *learning context* is a triple (X, C, ℓ) where X is a measurable space, C a concept class on X, and $\ell : X \times C \to [0, 1]$ is a loss function. For a probability measure μ on X define the expected loss of $c \in C$,

$$L_\mu(c) := \mathbb{E}_\mu[\ell(-, c)]$$

Remark. Having loss bounded will be important in the proofs of our main results so for convenience we restrict the loss to $[0, 1]$.

Definition 2.0.2. A learning context (X, C, ℓ) is *(efficiently) PAC learnable* if there is a (Laurent polynomial) $m_C : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm A such that for all $\epsilon, \delta \in (0, 1)$ and all probability measures μ over X let S be a set of $m \geq m_C(\epsilon, \delta)$ i.i.d. samples from (X, μ) then $A(S) = c$ such that,

$$\mathbb{P}(L_\mu(c) \leq \min_{c' \in C} L_\mu(c') + \epsilon) \geq 1 - \delta$$

Example 2.0.3. Let $\ell : X \times C \to [0, 1]$ be the "zero-one" loss,

$$\ell(x, c) = \begin{cases} 0 & x \in c \\ 1 & x \notin c \end{cases}$$

and suppose that μ is supported on some $c \in C$. Then the context (X, C, ℓ) is PAC learnable for any algorithm chosing the concept that fits a sample perfectly exactly if there is some $m_C : (0, 1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0, 1)$ and all measures μ on X for S for a set of $m \geq m_C(\epsilon, \delta)$ iid samples from (X, μ) then $A(S) = c$ so that

$$\mathbb{P}_{x \sim \mu}(x \in c) \geq 1 - \epsilon$$

with probability at least $1 - \delta$.

2.1 No Free Lunch

Theorem 2.1.1. Let X be finite, $C = \mathcal{P}(X)$, and $\ell : X \times C \to [0, 1]$ be the zero-one loss. Then there exist universal fixed $\epsilon, \delta \in (0, 1)$ such that if A is any learning algorithm for (X, C, ℓ) and $m < |X|/2$ then there is a distribution μ on X such that,

$$\mathbb{P}($$

2.2 How to Prove an Algorithm is PAC

Remark. We want to show under what conditions empirical risk minimization (ERM) is an (efficient) PAC learning algorithm. If the data is good enough to provide an accurate estimate of the expected loss then ERM will work.

Definition 2.2.1. Data $D \subset X$ is called an ϵ-representatvie for the learning context (X, C, ℓ) and probability measure μ if,

$$\forall c \in C : |L_D(c) - L_\mu(c)| \le \epsilon$$

Lemma 2.2.2. If D is an $\epsilon/2$-representative. Then any output c_D of ERM meaning $c_D \in \operatorname{argmin}_{c \in C} L_D(c)$ satisfies,

$$L_\mu(c_D) \le \min_{c \in C} L_\mu(c) + \epsilon$$

Proof. For any $c \in C$, we have $L_D(c_D) \le L_D(c)$ and,

$$|L_D(c) - L_\mu(c)| \le \epsilon$$

therefore,

$$L_\mu(c_D) \le L_D(c_D) + \epsilon/2 \le L_D(c) + \epsilon/2 \le L_\mu(c) + \epsilon/2 + \epsilon/2 = L_\mu(c) + \epsilon$$

□

Definition 2.2.3. A learning context (X, C, ℓ) has the *(efficient) uniform convergence property* if there is a (Laurent polynomial) $m_C^{\text{UC}} : (0,1)^2 \to \mathbb{N}$ such that for any probabilty measure μ on X let S be a set of $m \ge m_C^{\text{UC}}(\epsilon, \delta)$ i.i.d. samples from (X, μ) then S is an ϵ-representative with probability at least $1 - \delta$.

Corollary 2.2.4. If (X, C, ℓ) has the (efficient) uniform convergence property then ERM is a PAC learning algorithim for (X, C, ℓ) with $m(\epsilon, \delta) = m^{\text{UC}}(\epsilon/2, \delta)$.

2.3 Proving Uniform Convergence

Proposition 2.3.1. Consider the function,

$$f_C(n) = \frac{1 + \sqrt{\log \Pi_C(2n)}}{\sqrt{2n}}$$

and for a set $S \subset X$,

$$\Delta_{\mu,S}(c) = |L_\mu(c) - L_S(c)|$$

For any probability distribution μ and $\delta \in (0,1)$ and $S \sim \mu^n$,

$$\mathbb{P}[\exists c \in C : \Delta_S(c) \ge \delta^{-1} f(\delta, n)] \le \delta$$

Proof. Consider,

$$\mathbb{E}[\sup_{c \in C} \Delta_{\mu,S}(c)] = \mathbb{E}\left[\sup_{c \in C} \frac{1}{n} \left| \sum_{i=1}^{n} (\ell(s_i, c) - L_\mu(c)) \right|\right]$$

The idea is to bound the expectation in terms of an expectation over samples and then find a uniform bound regardless of the samples so that we only need to maximize over the finite set of concepts over the samples. Since the empirical loss is unbiased for $S' \sim \mu^n$,

$$L_\mu(c) = \mathbb{E}_{S'}[L_{S'}(c)]$$

we can introduce an independent sample to write,

$$\begin{aligned}\mathbb{E}[\sup_{c\in C} \Delta_{\mu,S}(c)] = \mathbb{E}\left[\sup_{c\in C}\left|\mathbb{E}_{S'}[L_S(c) - L_{S'}(c)]\right|\right] &\le \mathbb{E}_{S,S'}\left[\sup_{c\in C}\left|L_S(c) - L_{S'}(c)\right|\right] \\ &= \mathbb{E}_{S,S'}\left[\sup_{c\in C}\frac{1}{n}\left|\sum_{i=1}^n [\ell(s_i, c) - \ell(s_i', c)]\right|\right]\end{aligned}$$

Since $S \sim \mu^n$ and $S' \sim \mu^n$ are identically distributed, there is no difference in swapping each. So if we introduce V_i uniformly distributed on $\{\pm 1\}$ then,

$$\mathbb{E}[\sup_{c\in C} \Delta_{\mu,S}(c)] \le \mathbb{E}_{S,S'}\mathbb{E}_{V\sim U^n}\left[\sup_{c\in C}\frac{1}{n}\left|\sum_{i=1}^n V_i[\ell(s_i, c) - \ell(s_i', c)]\right|\right]$$

The point of introducing these new variables is to apply Hoeffding's inequality for fixed data S, S'. Fixing the S, S', the random variables,

$$W_{c,i} = V_i[\ell(s_i, c) - \ell(s_i', c)]$$

are i.i.d and supported on $[-1, 1]$ (using $\operatorname{im}\ell \subset [0, 1]$) so by Hoeffding's inequality, for any $\rho \ge 0$,

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^n W_{c,i}\right| \ge \rho\right] \le 2\exp\left(-2n\rho^2\right)$$

Therefore, let $Q = S \cup S'$ be the set of samples (rather that the list),

$$\mathbb{P}\left[\sup_{c\in C}\frac{1}{n}\left|\sum_{i=1}^n W_{c,i}\right| \ge \rho\right] = \mathbb{P}\left[\sup_{c\in C|_Q}\frac{1}{n}\left|\sum_{i=1}^n W_{c,i}\right| \ge \rho\right] \le 2|C|_Q|\exp\left(-2n\rho^2\right)$$

Then by Lemma 2.3,

$$\mathbb{E}\left[\sup_{c\in C}\frac{1}{n}\left|\sum_{i=1}^n W_{c,i}\right|\right] = f_C(n)$$

and therefore,

$$\mathbb{E}[\sup_{c\in C} \Delta_{\mu,S}(c)] \le \mathbb{E}_{S,S'}\mathbb{E}_{V\sim U^n}\left[\sup_{c\in C}\frac{1}{n}\left|\sum_{i=1}^n V_i[\ell(s_i, c) - \ell(s_i', c)]\right|\right] \le f_C(n)$$

Now, by Markov's inequality,

$$\mathbb{P}[\sup_{c\in C} \Delta_{\mu,S}(c) \ge \delta^{-1} f_C(\delta, n)] \le \delta$$

□

Lemma 2.3.2. Let X be a random variable, $a, b \in \mathbb{R}$ with $a > 0$ and $b \geq e$ such that for all $t \geq 0$ we have $\mathbb{P}[|X| > t] \leq 2be^{-t^2/a^2}$. Then,

$$\mathbb{E}[|X|] \leq a(1 + \sqrt{\log b})$$

Proof. Recall,

$$\mathbb{E}[|X|] = \int |X| \mathrm{d}\mu = \int_0^\infty \mathbb{P}[|X| > t] \mathrm{d}t \leq 2b \int_0^\infty e^{-t^2/a^2} \mathrm{d}t$$

Since the probability is bounded above by 1, for $t < a\sqrt{\log b}$ we can replace our bound by 1 to get,

$$\mathbb{E}[|X|] \leq a\sqrt{\log b} + 2b \int_{a\sqrt{\log b}}^\infty e^{-t^2/a^2} \mathrm{d}t$$

Since $b \geq e$, in the integral,

$$t \geq a\sqrt{\log b} \geq a$$

and hence,

$$\int_{a\sqrt{\log b}}^\infty e^{-t^2/a^2} \mathrm{d}t \leq \int_{a\sqrt{\log b}}^\infty \frac{t}{a} e^{-t^2/a^2} \mathrm{d}t = a \int_{\sqrt{\log b}}^\infty u e^{-u^2} \mathrm{d}u = a \left[-\tfrac{1}{2} e^{-u^2} \right]_{\sqrt{\log b}}^\infty = \frac{a}{2b}$$

Therefore,

$$\mathbb{E}[|X|] \leq a(1 + \sqrt{\log b})$$

□

Corollary 2.3.3. If $\mathrm{VC}(C) = d$ then

2.4 The Main Theorem

Theorem 2.4.1. Let (X, C, ℓ) be a learning context. Then the following are equivalent,

(a) C has the uniform convergence property

(b) any ERM rule is a PAC learning algorithm for (X, C, ℓ)

(c) (X, C, ℓ) is PAC learable

(d) (X, C, ℓ) is efficiently PAC learnable

(e) any ERM rule is an efficient PAC learning algorithm for (X, C, ℓ)

(f) $\mathrm{VC}(C) < \infty$.

Proof. □