10-716: Advanced Machine Learning **Spring 2019**

Lecture 9: February 14

Lecturer: Arun Sai Suggala *Scribes: Chirag Gupta, Aparna Joshi, Aniketh Reddy, Yao-Hung Tsai*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

In the previous class, we looked at the uniform convergence theorem. In this class, we start off by proving this theorem.

9.1 Uniform Convergence Theorem

We wish to bound the quantity $\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}\left[f(x)\right]\right|$.

Theorem 9.1 (Uniform convergence theorem) *Let $\mathcal{F}$ be b-uniformly bounded i.e., $\|f\|_\infty \leq b$ $\forall f \in \mathcal{F}$. Then $\forall n \geq 1$, $\forall \delta > 0$, we have with probability at least $1 - \exp\left(^{-n\delta^2}/_{2b^2}\right)$,*

$$\|\mathbb{P}_n - \mathbb{P}\|_2 \leq 2\mathcal{R}_n(\mathcal{F}) + \delta.$$

Proof: There are two key steps involved in the proof:

1. We need to show that the quantity $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ concentrates around the mean and,
2. Expectation of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is upper bounded.

Consider the following split:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] + \left|\|\mathbb{P}_n - \mathbb{P}\|_2 - \mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right]\right|$$

We first show the concentration around the mean by showing that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ satisfies the bounded difference inequality. Let $G(X_1^n) = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ where $X_1^n = (X_1, \ldots, X_n)$. Then let X_1^n and Y_1^n be two samples differing in j, i.e., $X_i = Y_i$ $\forall i \neq j$, and let $\bar{f}$ denote the centered random variable,

$$\bar{f}(X_i) = f(X_i) - \mathbb{E}\left[f(X)\right]$$

Then,

$$\begin{aligned}
G(X_1^n) - G(Y_1^n) &= \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \bar{f}(X_i) \right| - \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \bar{f}(Y_i) \right| \\
&\le \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \bar{f}(X_i) - \frac{1}{n}\sum_{i=1}^n \bar{f}(X_i) \right| && \text{(by property of supremum)} \\
&= \sup_{f\in\mathcal{F}} |f(X_j) - f(Y_j)| \\
&\le \frac{2b}{n}.
\end{aligned}$$

Thus, by bounded difference inequality, we can say that with probability at least $1 - \exp\left({-nt^2}/{2b^2}\right)$,

$$|G(X_1^n) - \mathbb{E}\left[G(X_1^n)\right]| \le t.$$

Next we show a bound on the expected value itself using the "symmetrization" trick. We introduce ghost samples $Y_1^n = (Y_1, \ldots, Y_n)$ such that they are iid and also independent of X_1^n with the same distribution.

$$\begin{aligned}
\mathbb{E}_X\left[\sup_f \left| \frac{1}{n}\sum_{i=1}^n \left(f(x_i) - \mathbb{E}_{y_i}\left[f(y_i)\right]\right) \right|\right] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \cdot \mathbb{E}_Y\left[\sum_{i=1}^n \left(f(x_i) - f(y_i)\right)\right] \right|\right] \\
&\overset{\xi_1}{\le} \mathbb{E}_X\left[\mathbb{E}_Y\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_i (f(x_i) - f(y_i)) \right|\right]\right] \\
&= \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_i (f(x_i) - f(y_i)) \right|\right] \\
&\overset{\xi_2}{=} \mathbb{E}_{X,Y,\epsilon_1^n}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_i \epsilon_i (f(x_i) - f(y_i)) \right|\right] \\
&\le \mathbb{E}_{X,Y,\epsilon_1^n}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_i \epsilon_i (f(x_i)) \right| + \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_i \epsilon_i (f(y_i)) \right|\right] \\
&= 2\mathcal{R}_N(\mathcal{F}).
\end{aligned}$$

Above, ξ_1 follows by applying Jensen's inequality on the convex function $\sup|\cdot|$. In ξ_2, we introduced iid Rademacher variables ϵ_1^n and then observe that $f(X_i) - f(Y_i) \overset{\text{distribution}}{=} f(Y_i) - f(X_i) \overset{\text{distribution}}{=} \epsilon_i(f(X_i) - f(Y_i))$. ■

This completes the proof but we still need to prove a bound on the Rademacher complexity. One technique to do this is via VC theory.

9.2 VC Theory

Consider a class of functions $\mathcal{F}$ where for each function $f \in \mathcal{F}$, $f : X \to \{0, 1\}$.

Definition 9.2 *In VC theory we are interested in talking about the quantity $\mathcal{T}(x_1^n)$ as defined below:*

- $\mathcal{T}_f = \{x : f(x) = 0\}$
- $\mathcal{T} = \{x : f(x) = 0\} : f \in \mathcal{F}\}$
- $\mathcal{T}(x_1^n) = \{\mathcal{T}_f \cap x_1^n : \mathcal{T}_f \in \mathcal{T}\}$

$F(x_1^n) = \{f(x_1), f(x_2), \cdots f(x_n)\}$.

Definition 9.3 *The class of sets $\mathcal{T}$ shatters $\{x_1, \cdots x_n\}$ if $|\mathcal{T}(x_1^n)| = 2^n$.*

Definition 9.4 *The VC-dimension of $\mathcal{T}$ is the largest n such that there exists a set $\{x_1, \cdots x_n\}$ of size n that can be shattered by $\mathcal{T}$.*

Example: Intervals on $\mathbb{R}$. Suppose $\mathcal{T} = \{\mathbb{I}_{(b,a]}, b < a\}$. First, it is easy to see that any set of two elements $\{x_1, x_2\}$ can be shattered. All the subsets $\{\{\}, \{x_1\}, \{x_2\}\{x_1, x_2\}\}$ are picked out by some interval in $\mathcal{T}$. However, if we consider a set of three elements $\{x_1, x_2, x_3\}$ then no interval in $\mathcal{T}$ can pick out $\{x_1, x_3\}$. Hence, $v(\mathcal{T}) = 2$.

We can also see that for any n, $|T(x_i^n)| \le (n+1)^2$. Take any $x_1 < x_2 < \cdots < x_n$. If $\mathcal{I}$ is an interval in $\mathcal{T}$, then,

$$\mathcal{I} \cap X_1^n = \{\{\}, \{x_i, x_{i+1}, \ldots, x_j\}\}, i < j$$

This is generalized by the famous VC theorem:

Theorem 9.5 *Let $\mathcal{T}$ be a class with finite VC dimension $v(\mathcal{T}) < \infty$. Then for any x_1^n with $n \ge v(\mathcal{T})$,*

$$|\mathcal{T}(x_1^n)| \le \sum_{i=0}^{v(\mathcal{T})} \binom{n}{i} \le (n+1)^{v(\mathcal{T})}.$$

Further, this bound can be used to show that

$$R_n(\mathcal{T}) \le \sqrt{\frac{4v(\mathcal{T}) \log(n+1)}{n}}.$$

Note: The log factors can be removed using some generic chaining arguments. Please refer to the course book by Martin Wainwright for more details.

9.2.1 Controlling VC dimension

We look at some basic operations on sets and how they affect the VC dimension.

Proposition: Let $\mathcal{T}_1, \mathcal{T}_2$ be two collections of sets with VC dimension $v(\mathcal{T}_1), v(\mathcal{T}_2)$ respectively. Then,

1. $v(\mathcal{T}_1^c) = v(\mathcal{T}_1)$.
2. $\mathcal{T}_1 \cup \mathcal{T}_2 := \{T_1 \cup T_2 : T_1 \in \mathcal{T}_1, T_2 \in \mathcal{T}_2\}$. Then, $v(\mathcal{T}_1 \cup \mathcal{T}_2) \leq v(\mathcal{T}_1) + v(\mathcal{T}_2)$.
3. $\mathcal{T}_1 \cap \mathcal{T}_2 := \{T_1 \cap T_2 : T_1 \in \mathcal{T}_1, T_2 \in \mathcal{T}_2\}$. Then, $v(\mathcal{T}_1 \cap \mathcal{T}_2) \leq v(\mathcal{T}_1) + v(\mathcal{T}_2)$.
4. $\mathcal{T}_1 \times \mathcal{T}_2 := \{T_1 \times T_2 : T_1 \in \mathcal{T}_1, T_2 \in \mathcal{T}_2\}$. Then, $v(\mathcal{T}_1 \times \mathcal{T}_2) \leq v(\mathcal{T}_1) + v(\mathcal{T}_2)$.

The proofs of the above facts are left as an exercise.

Example: For some $(a_1, \cdots a_d) \in \mathbb{R}$, consider the set $\mathcal{T} = \{(-\infty, a_1]\} \times (-\infty, a_2] \cdots \times (-\infty, a_d]\}$. We define $\mathcal{T}_1 = \{(-\infty, a] : a \in \mathbb{R})\}$. Then, $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2 \cdots \mathcal{T}_d$. By the above bound, $v(\mathcal{T}) \leq d$.

9.3 Vector Space Structure

In this section we consider a general class of classifiers and try to upper bound its VC dimension by leveraging the dimensionality of the vector space consisting of all possible decision boundaries.

Proposition: Let $\mathcal{G}$ be a finite dimensional vector space of real valued functions. Then, the class of sets $\mathcal{T} = \{\{x : g(x) \leq 0\} : g \in \mathcal{G}\}$ has $v(\mathcal{T}) \leq dim(\mathcal{G})$.

Proof: Let $n = dim(\mathcal{G}) + 1$ and let $X_1^n = (X_1, \ldots, X_n)$ be the set of samples. Now consider any $g \in \mathcal{G}$. Let $b_1, b_2, \ldots, b_{dim(\mathcal{G})}$ be the basis functions of $\mathcal{G}$. By the virtue of $\mathcal{G}$ being a vector space,

$$g(X) = \sum_{i=1}^{dim(\mathcal{G})} c_{ig} b_i(X)$$

Now, let us define a linear map $L : \mathcal{G} \to \mathbb{R}^n$ as $L(g) = (g(X_1), \ldots, g(X_n))$.

$$\implies L(g) = \begin{bmatrix} b_1(X_1) & b_2(X_1) & \ldots b_{dim(\mathcal{G})}(X_1) \\ \vdots & & \\ b_1(X_n) & b_2(X_n) & \ldots b_{dim(\mathcal{G})}(X_n) \end{bmatrix} \begin{bmatrix} c_{1g} \\ c_{2g} \\ \vdots \\ c_{dim(\mathcal{G})} \end{bmatrix} = BC_g$$

By construction, $rank(B) \leq dim(\mathcal{G})$. Thus, there exists a u such that $\langle u, L(g) \rangle = 0$, with $u \neq 0$.

Assume at least one coordinate of u is positive. Rewriting $\langle u, L(g) \rangle = 0$ as

$$\sum_{\{i|u_i \leq 0\}} (-u_i) g(x_i) = \sum_{\{i|u_i > 0\}} (u_i) g(x_i)$$

We claim that no set in $\mathcal{T}$ can pick out $\{x_i | u_i \leq 0\}$. We prove this by contradiction. Suppose $\exists g \in G$ such that $\{x : g(x) \leq 0\}$ picks out $\{x_i | u_i \leq 0\}$. Then,

$$\{x : g(x) \leq 0\} \cap X_1^n = \{x_i | u_i \leq 0\}$$

We know that RHS of the equation, $\sum_{\{i|u_i>0\}}(u_i)g(x_i) > 0$ since $u_i > 0$ and $g(x_i) > 0$. On the other hand, the LHS of the equation $\sum_{\{i|u_i\leq 0\}}(-u_i)g(x_i) \leq 0$ since $u_i < 0$ making $-u_i > 0$ and $g(x_i) < 0$. Thus,

$$LHS \neq RHS$$

Our assumption that $\exists g \in G$ such that $\{x : g(x) \leq 0\}$ picks out $\{x_i|u_i \leq 0\}$ is false and no set in $\mathcal{T}$ picks out the set $\{x_i|u_i \leq 0\}$. Thus, we can say that

$$|\mathcal{T}(X_1^n)| = |\{T \cap X_1^n | T \in \mathcal{T}\}| < 2^n$$

and,

$$v(\mathcal{T}) \leq dim(\mathcal{G})$$

Example: We define

$$\mathcal{F} = \{X \to \mathbb{I}(\langle a, X\rangle + b \leq 0) \mid (a, b) \in \mathbb{R}^d \times \mathbb{R}\}$$

as the set of all possible linear classifiers. The space defined by $\{\langle a, X\rangle + b\}$ is a $d+1$ dimensional vector space. Thus, $v(\mathcal{F}) \leq d + 1$.

Example: Spheres in $\mathbb{R}^d$. Consider $\mathcal{F} := \{x \to \mathbb{I}(\|x - c\|_2^2 \leq r^2) : (c, r) \in \mathbb{R}^d \times \mathbb{R}\}$. where c is the center and r is the radius of the sphere.

$$\|x - c\|_2^2 - r^2 = \sum_{i=1}^{d} x(i) + (\|c\|^2 - r^2) - 2\sum_{i=1}^{d} c(i)x(i)$$

where $x(i)$ denotes the ith coordinate of x. The vector space is spanned by the following $(d + 2)$ basis vectors: $b_1(X) = \sum_{i=1}^d x(i)^2$, $b_2(x) = x(1)$, $b_3(x) = x(2)$, $\ldots$, $b_{d+1}(x) = x(d)$ and $b_{d+2}(x) = 1$. Thus, the VC dimension of $\mathcal{F}, v(\mathcal{F}) \leq d + 2$.