2022

Lecture 2: Basics of Statistical Learning Theory

Week 2

Lecturer: Tianyu Wang

1 Elementary Probabilistic Inequalities

Theorem 1.1 (McDiarmids inequality)**.** *Let $X_1, \cdots, X_n$ be independent random variables, where $X_i \in \mathcal{X}_i \subseteq \mathbb{R}$. Let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be a function such that:*

$$|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i$$

for all $i = 1, 2, \cdots, n$, and all $(x_1, \cdots, x_i, \cdots, x_n), (x_1, \cdots, x_i', \cdots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For any $t > 0$,

$$\mathbb{P}\left(|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]| \geq t\right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

If we let $f(x_1, \cdots, x_n) = \sum_{i=1}^n x_i$, the McDiarmids inequality gives the Hoeffding's inequality. The proof of McDiarmids inequality is left as homework.

2 Empirical Risk Minimization

Consider an $i.i.d.$ dataset $\{(x_i, y_i)\}_{i=1}^n$, and a hypothesis class $\mathcal{H}$. Empirical risk minimization seeks to find a classifier in $\mathcal{H}$ that solves the following optimization objective

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[f(x_i) \neq y_i]}. \tag{1}$$

The question is: **How good is empirical risk minimization?** More specifically, if the data points are $i.i.d.$ sampled from $\mathbb{Q}$, can we bound the true risk of a function f, which is $R_{true}(f) := \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\mathbb{I}_{[f(x)\neq y]}\right]$, in terms of its empirical risk $R_{emp}(f) := \sum_{i=1}^n \frac{1}{n}\mathbb{I}_{[f(x_i)\neq y_i]}$ and the VC dimension of $\mathcal{H}$?

Again, this is about *generalization* and *learning*. If the model only memorizes the dataset, it cannot generalize to the true risk with respect to the true distribution.

3 Statistical Learning Theory with VC Dimension

Lemma 3.1. *Consider two* $i.i.d.$ *datasets* $\{(x_i, y_i)\}_{i=1}^n \sim \mathbb{Q}$, *and* $\{(x_i', y_i')\}_{i=1}^n \sim \mathbb{Q}$. *We have, for any* $f \in \mathcal{H}$ *and any* $t > 0$

$$\mathbb{P}\left(R_{emp}(f) - R'_{emp}(f) \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2}\right), \tag{2}$$

where $R_{emp}(f)$ *is the empirical risk on* $\{(x_i, y_i)\}_{i=1}^n$, *and* $R'_{emp}(f)$ *is the empirical risk on* $\{(x_i', y_i')\}_{i=1}^n$. *The notations* $R_{emp}(f)$ *and* $R'_{emp}(f)$ *will be used henceforth.*

Proof. By definition,

$$\begin{aligned} R_{emp}(f) - R'_{emp}(f) = &\frac{1}{n}\left(\sum_{i=1}^n \left(\mathbb{I}_{[f(x_i)\neq y_i]} - \mathbb{E}_{(x_i,y_i)\sim\mathbb{Q}}\left[\mathbb{I}_{[f(x_i)\neq y_i]}\right]\right)\right) \\ &+ \frac{1}{n}\left(\sum_{i=1}^n \left(\mathbb{E}_{(x_i',y_i')\sim\mathbb{Q}}\left[\mathbb{I}_{\left[f(x_i')\neq y_i'\right]}\right] - \mathbb{I}_{\left[f(x_i')\neq y_i'\right]}\right)\right). \end{aligned}$$

We can now apply the McDiarmids inequality to the above equation to conclude the proof. □

Lemma 3.2. *Consider two* $i.i.d.$ *datasets* $\{(x_i, y_i)\}_{i=1}^n \sim \mathbb{Q}$, *and* $\{(x_i', y_i')\}_{i=1}^n \sim \mathbb{Q}$. *It holds that*

$$\mathbb{P}\left(\{R_{true}(f) - R_{emp}(f) > t\}\right) \leq \frac{\mathbb{P}\left(\{R'_{emp}(f) - R_{emp}(f) > t/2\}\right)}{\mathbb{P}\left(\{R'_{emp}(f) - R_{true}(f) > -t/2\}\right)}.$$

Proof. Consider the following inclusion of events

$$\begin{aligned} &\{R_{true}(f) - R_{emp}(f) > t\} \cap \{R'_{emp}(f) - R_{true}(f) > -t/2\} \\ \Rightarrow &\{R'_{emp}(f) - R_{emp}(f) > t/2\}. \end{aligned}$$

Thus we have

$$\begin{aligned} &\mathbb{P}\left(\{R'_{emp}(f) - R_{emp}(f) > t/2\}\right) \\ \geq &\mathbb{P}\left(\{R_{true}(f) - R_{emp}(f) > t\} \cap \{R'_{emp}(f) - R_{true}(f) > -t/2\}\right) \\ = &\mathbb{P}\left(\{R_{true}(f) - R_{emp}(f) > t\}\right)\mathbb{P}\left(\{R'_{emp}(f) - R_{true}(f) > -t/2\}\right), \end{aligned}$$

where the last inequality uses independence of $\{(x_i, y_i)\}_{i=1}^n$ and $\{(x_i', y_i')\}_{i=1}^n$. □

Lemma 3.3. *Instate the settings and notations in previous lemmas. Suppose* $\mathcal{H}$ *is closed. Let* $f^* \in \arg\max_{f\in\mathcal{H}}\left(R'_{emp}(f) - R_{emp}(f)\right)$ *with respect to fixed datasets* $\{(x_i, y_i)\}_{i=1}^n$ *and* $\{(x_i', y_i')\}_{i=1}^n$. *For any* $t \in \mathbb{R}$, *it holds that*

$$\begin{aligned} &\mathbb{P}\left(\sup_{f\in\mathcal{H}}\left(R'_{emp}(f) - R_{emp}(f)\right) \geq t/2\right) \\ \leq &\mathbb{P}\left(\sup_{(y_1,\cdots,y_n,y_1'\cdots,y_n')\in\mathcal{H}_{x_1,\ldots,x_n,x_1',\ldots,x_n'}}\left(R'_{emp}(f^*) - R_{emp}(f^*)\right) \geq t/2\right). \end{aligned}$$

Proof. For any $f \in \mathcal{H}$, it holds that

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[f(x_i)\neq y_i]} \leq \sup_{(y_1,\cdots,y_n)\in\mathcal{H}_{x_1,\ldots,x_n}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[f(x_i)\neq y_i]},$$

and similarly,

$$R_{emp}(f) \geq \inf_{(y_1,\cdots,y_n)\in\mathcal{H}_{x_1,\ldots,x_n}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[f(x_i)\neq y_i]}.$$

We use the above observations and we arrive at the following argument. Let $f^* \in \arg\max_{f\in\mathcal{H}} \left(R'_{emp}(f) - R_{emp}(f\right.$ for a fixed realization of datasets. We have

$$\begin{aligned}\sup_{f\in\mathcal{H}} \left(R'_{emp}(f) - R_{emp}(f)\right) &= R'_{emp}(f^*) - R_{emp}(f^*) \\ &\leq \sup_{(y_1,\cdots,y_n,y'_1\cdots,y'_n)\in\mathcal{H}_{x_1,\ldots,x_n,x'_1,\ldots,x'_n}} \left(R'_{emp}(f^*) - R_{emp}(f^*)\right)\end{aligned}$$

□

Note 1. We *cannot* replace $\mathbb{P}\left(\sup_{f\in\mathcal{H}}\left(R'_{emp}(f) - R_{emp}(f)\right) \geq t/2\right)$ by $\mathbb{P}\left(\left(R'_{emp}(f^*) - R_{emp}(f^*)\right) \geq t/2\right)$ since, in a probabilistic setting, f^* depends on the realization of the datasets. However, by taking a supremum over all possible data configurations, the randomness is effectively removed.

Note 2. Lemma 3.3 converts a supremum over a possibly infinite set $\mathcal{H}$ to a maximum over a finite set $\mathcal{H}_{x_1,\ldots,x_n,x'_1,\ldots,x'_n}$. This will be helpful when we later apply a union bound.

Theorem 3.4. *Instate the assumptions and notations in previous lemmas. For any $f \in \mathcal{H}$, it holds that*

$$\mathbb{P}\left(R_{true}(f) \leq R_{emp}(f) + O\left(\sqrt{\frac{\log(S_{\mathcal{H}}(2n)/\delta)}{n}}\right)\right) \geq 1-\delta, \quad \forall\delta \in (0,1).$$

Proof. We gathers results from previous lemmas to give a proof:

$$\begin{aligned}&\mathbb{P}\left(R_{true}(f) - R_{emp}(f) \geq t\right) \\ \leq\ &\frac{\mathbb{P}\left(\{R'_{emp}(f) - R_{emp}(f) > t/2\}\right)}{\mathbb{P}\left(\{R'_{emp}(f) - R_{true}(f) > -t/2\}\right)} && \text{(by Lemma 3.2)} \\ \leq\ &\frac{1}{1-\exp\left(-\frac{nt^2}{8}\right)}\mathbb{P}\left(\{R'_{emp}(f) - R_{emp}(f) > t/2\}\right). && \text{(by the McDiarmids inequality)}\end{aligned}$$

Also, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{H}} R'_{emp}(f) - R_{emp}(f) > t/2\right) && (3)\\
&\le \mathbb{P}\left(\sup_{(y_1,\cdots,y_n,y'_1\cdots,y'_n)\in\mathcal{H}_{x_1,\ldots,x_n,x'_1,\ldots,x'_n}} \left(R'_{emp}(f^*) - R_{emp}(f^*)\right) \ge t/2\right) && \text{(by Lemma 3.3)}\\
&\le \sum_{(y_1,\cdots,y_n,y'_1\cdots,y'_n)\in\mathcal{H}_{x_1,\ldots,x_n,x'_1,\ldots,x'_n}} \mathbb{P}\left(\left(R'_{emp}(f^*) - R_{emp}(f^*)\right) \ge t/2\right) && \text{(by union bound)}\\
&\le S_{\mathcal{H}}(2n)e^{-nt^2/8}.
\end{aligned}
$$

Note that $\frac{1}{1-\exp\left(-\frac{nt^2}{8}\right)} = O(1)$ for n and t larger than some constant. We then let $\delta = S_{\mathcal{H}}(2n)e^{-nt^2/8}$ and rearrange the terms to finish the proof. □

Corollary 3.5. *Instate the assumptions and notations from previous lemmas. Let $f_{\min}$ be a function in $\mathcal{H}$ such that $f_{\min} \in \arg\min_{f\in\mathcal{H}} R_{true}(f)$. Let $f_n \in \arg\min_{f\in\mathcal{H}} R_{emp}(f)$. Then it holds that*

$$
\mathbb{P}\left(R_{true}(f_{\min}) \le R_{emp}(f_n) + O\left(\sqrt{\frac{\log(S_{\mathcal{H}}(2n)/\delta)}{n}}\right)\right) \ge 1-\delta, \quad \forall\delta \in (0,1).
$$

Proof. Since $R_{true}(f_{\min}) \le R_{true}(f_n)$, this corollary follows from the Theorem 3.4. Note that $R_{true}(f)$ does not depend on the realization of the datasets. □

4 Back to Growth Function

First of all, note that VC dimension of a hypothesis class $\mathcal{H}$ can be equivalently defined as

$$
\text{VC-dim}(\mathcal{H}) = \max\{n : S_{\mathcal{H}}(n) = 2^n\}.
$$

This can be verified by checking the definition of shattering.

Lemma 4.1. *Let $\mathcal{H}$ be a class of functions with finite VC dimension d. Then for all positive integers n,*

$$
S_{\mathcal{H}}(n) \le \sum_{i=0}^{d} \binom{n}{i},
$$

where $d :=$ VC-dim$(\mathcal{H})$.

Proof. For any $X = \{x_1, \cdots, x_n\}$, consider a table containing values of $\mathcal{H}_X$. Recall the definition of $\mathcal{H}_X$ in the previous notes.

$h(x_1)$	$h(x_2)$	$h(x_3)$	$\cdots$	$h(x_n)$
-	+	-	$\cdots$	+
+	-	-	$\cdots$	+
-	+	+	$\cdots$	-
+	+	+	$\cdots$	+
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: The evaluation table.

Obviously the number of unique rows in the evaluation table T is the same as the cardinality of $\mathcal{H}_X$. Let $d :=$ VC-dim$(\mathcal{H})$. If one row has i "+"s in it, it must be one of the $\binom{n}{i}$ patterns. Summing over i from 0 to d, we know that the number of unique rows in T is upper bounded by $\sum_{i=0}^{d}\binom{n}{i}$, which means

$$\sup_{X,|X|=n} |\mathcal{H}_X| \leq \sum_{i=0}^{d}\binom{n}{i}.$$

Note that there is no need to sum over $i = d+1, d+2, \cdots, n$. The reason is that $\mathcal{H}$ cannot correctly classify points of size $d+1$ and above.

$\square$

Lemma 4.2. *Let $\mathcal{H}$ be a hypothesis class with VC-dim$(\mathcal{H}) = d$. Then for all $m \geq d$,*

$$S_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d \leq O(n^d).$$

Proof.

$$\begin{aligned}
S_{\mathcal{H}}(n) &\leq \sum_{i=0}^{d}\binom{n}{i} && \text{(by Lemma 4.1)} \\
&\leq \sum_{i=0}^{n}\binom{n}{i}\left(\frac{n}{d}\right)^{d-i} \\
&= \left(\frac{n}{d}\right)^d \sum_{i=0}^{n}\binom{n}{i}\left(\frac{d}{n}\right)^i \\
&= \left(\frac{n}{d}\right)^d \left(1+\frac{d}{n}\right)^n && \text{(by the binomial theorem)} \\
&\leq \left(\frac{n}{d}\right)^d e^d && \text{(since } \left(1+\tfrac{d}{n}\right)^n \text{ converges to } e^d \text{ from below.)}
\end{aligned}$$

$\square$

Theorem 4.3. *Instate the assumptions and notations from previous lemmas. Let $f_{\min}$ be a function in $\mathcal{H}$ such that $f_{\min} \in \arg\min_{f\in\mathcal{H}} R_{true}(f)$. Let $f_n \in \arg\min_{f\in\mathcal{H}} R_{emp}(f)$. Then it holds that*

$$\mathbb{P}\left(R_{true}(f_{\min}) \leq R_{emp}(f_n) + O\left(\sqrt{\frac{\textit{VC-dim}(\mathcal{H})\log(n/\delta)}{n}}\right)\right) \geq 1-\delta, \quad \forall\delta \in (0,1).$$

Proof. This theorem is a consequence of Lemma 4.2 and Corollary 3.5. □

Acknowledgement

A recommended general reference is the machine learning textbook by Kevin Murphy. TW used lecture notes by Cynthia Rudin to compile this notes. The theorems are due to Vapnik, Chervonenkis, Sauer, Shelah. Also, a thank you to wikipedia contributors.