CS861: Theoretical Foundations of Machine Learning Lecture 3 - 09/11/2023
University of Wi[illegible] Fall 2023

Lectu[illegible]ion to Radamacher complexity

Lecturer: Kirthe[illegible] *Scribed by: Justin Kiefel, Albert Dorador*

Disclaimer: *These n[illegible]ected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we will first finish up **Bayes optimal classifier** from the previous class. Then, we will introduce **McDiarmid's Inequality**. This tool is then applied to derive the **uniform convergence** in probability of the empirical risk to the true risk for any hypothesis in a given hypothesis class. Lastly, we will introduce the **Empirical Radamacher complexity** as a tool to more explicitly bound the difference between empirical risk and true risk, with large probability.

1 Bayes optimal classifier

We will begin this lecture by introducing the Bayes optimal classifier. This classifier always selects the class with the highest probability conditioned on the input. For binary classification, the classifier is as follows:

$$h_B(x) = \arg\max_{y\in\{0,1\}} \mathbb{P}(Y = y|X = x)$$
$$= \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X = x) \geq \frac{1}{2} \\ 0 & \text{if } \mathbb{P}(Y = 0|X = x) \geq \frac{1}{2} \end{cases}$$

We can show that the Bayes optimal classifier produces the minimum risk across all potential classifiers. The intuition behind this result is that selecting the highest probability class will always minimize the expected prediction error.

Theorem: $\forall h \in \{h : X \to Y\}$, $R(h) \geq R(h_B)$

Proof:

$$R(h) = \mathbb{E}_{XY}[\mathbb{1}(h(X) \neq Y)]$$

Applying the law of iterated expectation:

$$= \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbb{1}(h(X) \neq Y)|X]]$$

Using the law of total expectation:

$$= \mathbb{E}_X[\mathbb{E}_Y[\mathbb{1}(h(X) \neq Y)|Y = 0, X]\,\mathbb{P}(Y = 0|X) + \mathbb{E}_Y[\mathbb{1}(h(X) \neq Y)|Y = 1, X]\,\mathbb{P}(Y = 1|X)]$$
$$= \mathbb{E}_X[\mathbb{1}(h(X) \neq 0)\,\mathbb{P}(Y = 0|X) + \mathbb{1}(h(X) \neq 1)\,\mathbb{P}(Y = 1|X)]$$
$$= \mathbb{E}_X[\mathbb{1}(h(X) = 1)\,\mathbb{P}(Y = 0|X) + \mathbb{1}(h(X) = 0)\,\mathbb{P}(Y = 1|X)]$$

Applying the definition of expectation:

$$= \int_x (\mathbb{1}(h(x) = 1)\,\mathbb{P}(Y = 0|X = x) + \mathbb{1}(h(x) = 0)\,\mathbb{P}(Y = 1|X = x))\,dP_X(x)$$

Observe that the above integrand is minimized pointwise if $h(x)$ obeys the following scheme: if $P(Y=0|X=x) \geq P(Y=1|X=x)$ then set $h(x)=0$ so that only the smaller of two summands in the integrand is "activated". A symmet[illegible]hat if $P(Y=1|X=x) \geq P(Y=0|X=x)$ then we should choose $h(x)=1$. In ot[illegible]d, and therefore the risk of h, is minimized for all x if h is chosen to be:

$$[illegible] = 1|X=x) \geq \mathbb{P}(Y=0|X=x)$$
$$[illegible] = 0|X=x) \geq \mathbb{P}(Y=1|X=x)$$

Notice, that this classifier is identical to the Bayes classifier, h_B. Since the Bayes classifier minimizes this term for all x, the integral will be minimized, and $h_B(x)$ will produce the minimum possible value of $R(h)$. $\square$

2 McDiarmid's inequality

Next, we will introduce McDiarmid's inequality. This concentration inequality bounds the difference between a function's sampled value and its expected value. We will use McDiarmid's inequality when working with Radamacher complexity. To begin, let us define the bounded difference property. This property is necessary to apply McDiarmid's inequality.

Definition 1 (Bounded Difference Property). *Let $f : \mathbb{R}^n \to \mathbb{R}$. f satisfies the bounded difference property when $\exists c_1, ..., c_n \in \mathbb{R}$ such that $\forall k \in \{1, ..., n\}$:*

$$\sup_{z_1,...,z_k,...,z_n,\tilde{z}_k} |f(z_1, ..., z_k, ..., z_n) - f(z_1, ..., \tilde{z}_k, ..., z_n)| \leq c_k$$

Intuitively, the bounded difference property states that changing any input to a function will lead to a finite difference in the function's output. Now we define McDiarmid's Inequality.

Theorem 1 (McDiarmid's Inequality). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying the bounded difference property with bounds $c_1, ..., c_n \in \mathbb{R}$. Let $Z_1, ..., Z_n$ be n independent random variables. Then for all $\varepsilon > 0$:*

$$\mathbb{P}(f(Z_1, ..., Z_n) - \mathbb{E}[f(Z_1, ..., Z_n)] > \varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{k=1}^n c_k^2}\right)$$

Similarly,

$$\mathbb{P}(\mathbb{E}[f(Z_1, ..., Z_n)] - f(Z_1, ..., Z_n) > \varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{k=1}^n c_k^2}\right)$$

To demonstrate the application of McDiarmid's inequality we present the following example:

Example 1. We will use McDiarmid's inequality to show $\mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$. For this, first, we will show that the function $\widehat{R}(h)$ satisfies the bounded difference property. Let $X_1, ..., X_n \in \mathbb{R}^d$ and $Y \in \{0, 1\}$ be random variables. Now we define the random variable $Z_i = \mathbb{1}(h(X_i) \neq Y_i)$.

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(h(X_i) \neq Y_i) = \frac{1}{n}\sum_{i=1}^n Z_i$$

Hence, we can represent $R(h)$ as a function of $Z_1, ..., Z_n$. Let $k \in \{1, ..., n\}$. Next, we can see that the maximum difference in $\widehat{R}(h)$ from changing Z_k is bounded by 1.

$$\sup_{Z_1,\ldots,Z_k,\ldots} \ldots Z_i + \tilde{Z}_k + \sum_{i=k+1}^{n} Z_i)\right| = \sup_{Z_k,\tilde{Z}_k} \left|\frac{1}{n}Z_k - \frac{1}{n}\tilde{Z}_k\right| = \frac{1}{n}$$

Hence, the bounded diff… to $\widehat{R}(h)$, and the maximum difference for changing any input is $\frac{1}{n}$. Now we can apply… .

$$\mathbb{P}(\widehat{R}(h) - R \ldots [\widehat{R}(h)] > \varepsilon\Big) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{k=1}^{n}(1/n)^2}\right) = \exp(-2n\varepsilon^2)$$

Applying the same reasoning we get:

$$\mathbb{P}(R(h) - \widehat{R}(h) > \varepsilon) \leq \exp(-2n\varepsilon^2)$$

Using the union bound, we get our desired result.

$$\mathbb{P}(|R(h) - \widehat{R}(h)| > \varepsilon) \leq 2\exp(-2n\varepsilon^2)$$

3 Uniform convergence

We would like to have, for any small $\varepsilon > 0$

$$\mathbb{P}(\forall h \in \mathcal{H}, |\widehat{R}(h) - R(h)| \leq \varepsilon) \geq \gamma$$

where $\gamma \in [0, 1]$ is a large quantity (i.e. close to 1).
In a previous lecture, we have considered for an arbitrary $h \in \mathcal{H}$,

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) \leq \delta$$

where $\delta \in [0, 1]$ is a small quantity (i.e. close to 0), and that bound was derived e.g. by Hoeffding's inequality. Then, we applied a union bound and obtained:

$$\mathbb{P}(\exists h \in \mathcal{H} |\widehat{R}(h) - R(h)| > \varepsilon) \leq \sum_{h \in \mathcal{H}} \mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) \leq |\mathcal{H}| \cdot \delta$$

Of course, the above statement is vacuous if $|\mathcal{H}| \cdot \delta \geq 1$, which will necessarily be the case if $|\mathcal{H}| = \infty$ no matter how small $\delta > 0$ is. Therefore, we wish to derive a bound that is still useful in the presence of a potentially non-finite hypothesis class. We will proceed by considering the following quantity

$$f(S) := \sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h))$$

where

$$S := \{(X_1, Y_1), \ldots, (X_k, Y_k), \ldots, (X_n, Y_n)\}$$

with $X_i \in \mathcal{X}$ and $Y_i \in \{0, 1\}$, and $\widehat{R}_S(h) := \frac{1}{n} \sum_{(X_i,Y_i) \in S} \mathbb{I}_{\{h(X_i) \neq Y_i\}}$

To apply McDiarmid's inequality, additionally define

$$\tilde{S} := \{(X_1, Y_1), \ldots, (\tilde{X}_k, \tilde{Y}_k), \ldots, (X_n, Y_n)\}$$

Then, equipped with the above definitions,

$$
\begin{aligned}
\sup_{S\cup\tilde{S}} \text{[illegible]} & \text{[illegible]} \left|\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h)) - \sup_{h\in\mathcal{H}}(\widehat{R}_{\tilde{S}}(h) - R(h))\right| \\
& \text{[illegible]} \sup_{h\in\mathcal{H}} \left|(\widehat{R}_S(h) - R(h)) - (\widehat{R}_{\tilde{S}}(h) - R(h))\right| \\
& \text{[illegible]} \sup_{h\in\mathcal{H}} \left|\widehat{R}_S(h) - \widehat{R}_{\tilde{S}}(h)\right| \\
& \text{[illegible]} \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\left(\mathbb{I}_{\{h(X_k)\neq Y_k\}} - \mathbb{I}_{\{h(\tilde{X}_k)\neq \tilde{Y}_k\}}\right)\right| \\
& \leq \frac{1}{n}
\end{aligned}
$$

where the inequality above follows from the fact that for any functions f_1, f_2,

$$|\sup_a f_1(a) - \sup_a f_2(a)| \leq \sup_a |f_1(a) - f_2(a)|$$

Noting that $\frac{1}{n}$ plays the role of c_k for all k in the bounded difference property introduced in the previous section, we can see that said property holds and so we can apply McDiarmid's inequality as follows:

$$
\begin{aligned}
\mathbb{P}_{S\sim P_{X,Y}}(f(S) - \mathbb{E}[f(S)] > \varepsilon) &= \mathbb{P}_{S\sim P_{X,Y}}\left(\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h)) - \mathbb{E}[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))] > \varepsilon\right) \\
&\leq \exp(-2n\varepsilon^2)
\end{aligned}
$$

Observe that the above probability is equivalent to

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h)) > \mathbb{E}[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))] + \varepsilon\right)$$

which in turn is equivalent to

$$\mathbb{P}\left(\exists h\in\mathcal{H}, (\widehat{R}_S(h) - R(h)) > \mathbb{E}\left[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))\right] + \varepsilon\right)$$

Hence, by McDiarmid's inequality, we can equivalently claim that, $\forall h \in \mathcal{H}$, with probability at least $1 - \exp(-2n\varepsilon^2)$,

$$\widehat{R}_S(h) - R(h) \leq \mathbb{E}[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))] + \varepsilon$$

It would be more illuminating if we had a way to quantify or at least bound the term $\mathbb{E}[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))]$, which, given the non-linear nature of the supremum operator, is in general not equal to $\sup_{h\in\mathcal{H}}\mathbb{E}[(\widehat{R}_S(h) - R(h))] = \sup_{h\in\mathcal{H}} \cdot 0 = 0$.
Next, we will introduce Radamacher complexity to help us bound the term above.

4 Radamacher complexity

Definition 2. *A Radamacher random variable $\sigma \in \{-1, 1\}$ is such that $\mathbb{P}(\sigma = -1) = \mathbb{P}(\sigma = 1) = 1/2$*

Definition 3 (Empirical Radamacher Complexity)**.** *Let $S := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be an observed sample of n points, and let $\sigma := (\sigma_1, \ldots, \sigma_n) \in \{-1, 1\}^n$ be n independent Radamacher random variables. Then, the Empirical Radamacher Complexity is*

$$\widehat{Rad}(S, \mathcal{H}) := \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(h(x_i), y_i)\right]$$

where $\ell(\cdot)$ is our choice of loss function, e.g. $\ell(h(x_i), y_i) = \mathbb{I}_{\{h(x_i)\neq y_i\}}$ in case of a classification problem.

The above definition can be intuitively interpreted as follows: let $\underset{\sim}{\ell} := (\ell(h(x_1), y_1), \ldots, \ell(h(x_n), y_n))$. Then,

$$\ldots \mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sigma \cdot \underset{\sim}{\ell} \right]$$

measures how well the h[...] correlate[1] with a random Radamacher vectors. More flexible hypothesis classes will h[...]e with random vectors.

[1]Recall that for any two vectors a, b, their dot product is equal to the cosine of the angle between them, scaled by the product of their norms; if those vectors have mean zero, their cosine coincides with their correlation coefficient.