# HOMEWORK SET 3

CSCI5525 Machine Learning: Analysis and Methods (Fall 2024)

**Due** 11:59 pm, Nov 30 2024

**Instruction** Your writeup, either typeset or scanned, should be a single PDF file. For problems requiring coding, organize all codes for all problems into **ONE** Jupyter notebook file (i.e., `.ipynb` file) with cell execution outputs. Your submission to Gradescope should include the single PDF and the one notebook file—**please DO NOT zip them**! Please assign page(s) to each question to help reduce the TA's navigation time. If your notebook submission does not display properly on Gradescope due to a large file error, please try to remove images and/or figures from your cell outputs and re-upload it. No late submission will be accepted. For each problem, you should acknowledge your collaborators—**including AI tools**, if any.

**About the use of AI tools** You are strongly encouraged to use AI tools—they are becoming our workspace friends, such as ChatGPT (https://chat.openai.com/), Claude (https://claude.ai/chats), and Github Copilot (https://github.com/features/copilot), to help you when trying to solve problems. It takes a bit of practice to ask the right and effective questions/prompts to these tools; we highly recommend that you go through this popular free short course **ChatGPT Prompt Engineering for Developers** offered by https://learn.deeplearning.ai/ to get started.

**If you use any AI tools for any of the problems, you should include screenshots of your prompting questions and their answers in your writeup**. The answers provided by such AI tools often contain factual errors and reasoning gaps. **So, if you only submit an AI answer with such bugs for any problem, you will obtain a zero score for that problem**. You obtain the scores only when you explain the bugs and also correct them in your own writing. You can also choose not to use any of these AI tools, in which case we will grade based on the efforts you have made.

**Reminder about notations** We will use small letters (e.g., $u$) for scalars, small boldface letters (e.g., $\boldsymbol{a}$) for vectors, and capital boldface letters (e.g., $\boldsymbol{A}$) for matrices. For a matrix $\boldsymbol{A}$, $\boldsymbol{a}^i$ (supscripting) means its $i$-th row as a *row vector*, and $\boldsymbol{a}_j$ (subscripting) means the $j$-the column as a column vector, and $a_{ij}$ means its $(i, j)$-th element. $\mathbb{R}$ is the set of real numbers. $\mathbb{R}^n$ is the space of $n$-dimensional real vectors, and similarly $\mathbb{R}^{m\times n}$ is the space of $m \times n$ real matrices. The dotted equal sign $\doteq$ means defining.

**Problem 1 (Learning with finite hypothesis class; 4.5/15)** Consider an input space $\mathcal{X} \in \mathbb{R}^d$, output space $\mathcal{Y} = \{+1, -1\}$, i.e., a binary classification problem, and a hypothesis class $\mathcal{H}$ that consists of a finite number of functions mapping from $\mathcal{X}$ to $\{+1, -1\}$.

In class, we showed the following result for the realizable cases: for any $\varepsilon \in (0, 1)$, any $\delta \in (0, 1)$, any labeling function $f_* : \mathcal{X} \to \{+1, -1\}$, and any distribution $\mathcal{D}_\mathcal{X}$ on $\mathcal{X}$, if the realizability assumption holds, i.e., there exists an $h_* \in \mathcal{H}$ so that $R(h_*) = 0$ where $R(\cdot)$ denotes the risk, every predictor $h_S$ returned by the ERM (i.e., empirical risk minimization) rule satisfies

$$\mathbb{P}\left[R(h_S) \geq \varepsilon\right] \leq \delta \quad \text{(here, } \mathbb{P}\left[\cdot\right] \text{ means the probability of a certain event happens)} \tag{1}$$

over the iid drawing of a training set $S$ of size $N$, provided that $N \geq \varepsilon^{-1} \log(|\mathcal{H}|\, \delta^{-1})$, or equivalently, every predictor $h_S$ returned by the ERM rule satisfies

$$\mathbb{P}\left[R(h_S) \leq \frac{1}{N} \log \frac{|\mathcal{H}|}{\delta}\right] \geq 1 - \delta. \tag{2}$$

(a) To prove the result as stated, we have used the union bound, a celebrated result widely used in probability analysis: for any two events $A, B$, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

(i) Prove the union bound using the fact that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$; When is the upper bound in the union bound tight, i.e., $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$? $(0.5/15)$

(ii) Prove that for any integer $P \geq 1$ and events $A_1, \ldots, A_P$, $\mathbb{P}(\bigcup_{p\in[P]} A_p) \leq \sum_{p\in[P]} \mathbb{P}(A_p)$. $(0.5/15)$

(b) Now consider the problem of rolling two six-sided dice. Let $A_1$ be the event that the first die shows a 6, and $A_2$ be the event that the second die shows a 6.

(i) Estimate $\mathbb{P}(A_1 \cup A_2)$ using the union bound; $(0.5/15)$

(ii) Calculate the exact probability $\mathbb{P}(A_1 \cup A_2)$. How large is the gap between the upper bound obtained in (i) and the exact probability you get here? $(0.5/15)$

(c) Let $\mathcal{H}_{bol}$ be the set of all Boolean functions $\{0,1\}^k \mapsto \{0,1\}$, i.e., $\mathcal{X} = \{0,1\}^k$ (i.e., all binary strings of length $k$) and $\mathcal{Y} = \{0,1\}$, and $f_*$ an arbitrary fixed Boolean function which is the true labeling function. For given $\varepsilon, \delta \in (0,1)$, what's the size of the training set we need to make sure that $\mathbb{P}[R(h_S) \geq \varepsilon] \leq \delta$, where $h_S \in \mathcal{H}_{bol}$ is any predictor returned by the ERM rule? $(0.5/15)$

(d) Due to the Hoeffding's inequality, for any fixed predictor $h : \mathcal{X} \to \{+1, -1\}$ and any $\delta \in (0,1)$,

$$\mathbb{P}\left[\left|R(h) - \widehat{R}_S(h)\right| \leq \sqrt{\frac{\log(2/\delta)}{2N}}\right] \geq 1 - \delta \tag{3}$$

over the iid drawing of a training set $S$ of size $N$. Now, imagine tossing a biased coin that lands heads with probability $p$, and let our predictor $h$ be the one that always guesses tails.

(i) What's the risk $R(h)$? $(0.5/15)$

(ii) Set $\delta = 0.02$. How large does $N$ need to be to ensure that the empirical risk $\widehat{R}(h)$ differs from the risk $R(h)$ by no more than $0.03$? $(0.5/15)$

(e) After lifting the realizability assumption, we derived the following result in class: assume that the loss $\ell$ is bounded in $[0,1]$. For any $\varepsilon \in (0,1)$, any $\delta \in (0,1)$, and any distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$, every predictor $h_S$ returned by the ERM rule satisfies

$$\mathbb{P}\left[\left|R(h) - \widehat{R}(h)\right| \leq \varepsilon \ \forall h \in \mathcal{H}\right] \geq 1 - 2\left|\mathcal{H}\right| \exp(-2N\varepsilon^2), \tag{4}$$

or equivalently,

$$\mathbb{P}\left[\left|R(h) - \widehat{R}(h)\right| \leq \sqrt{\frac{\log(2\left|\mathcal{H}\right|/\delta)}{2N}} \ \forall h \in \mathcal{H}\right] \geq 1 - \delta. \tag{5}$$

Using these results, as well as any relevant results stated above, to solve the following contrived problem.

Assume that in the US Senate, laws are proposed in a random fashion independently and identically according to some distribution $\mathcal{D}$ determined by an unknown group of senators. Any law proposal (i.e., bill) is to be debated and voted on in the Senate, with a Yes/No outcome. Suppose that there is a pool of $\mathcal{H} = 2,500$ independent law experts who can predict whether a law proposal can be passed or not in the Senate.

(i) Select an expert from $\mathcal{H}$ who has consistently predicted correctly for the last $N = 200$ law proposals. What's the probability that this expert **incorrectly** predicts the outcome of the next law proposal? A reasonable upper bound is fine. What's the value of the bound with $95\%$ confidence, i.e., when $\delta = 0.05$? $(0.5/15)$

(ii) Assume now that we select an expert from $\mathcal{H}$ who, for the past $N = 200$ law proposals, has predicted incorrectly 20 times. What is the value of the new bound? $(0.5/15)$

**Problem 2 (Learning with infinite hypothesis class with Rademacher complexity; $7/15$)** For cases with a finite hypothesis class and a bounded loss, the results stated in $1(\mathrm{e})$ are sufficiently general and powerful. However, for learning settings with an infinite hypothesis class, we need to measure the effective complexity of the hypothesis class through, e.g., Rademacher complexity, VC dimensions, growth functions, and so on.

For any set $A \subset \mathbb{R}^d$, the Rademacher complexity of $A$ is defined as

$$\mathsf{RC}(A) \doteq \frac{1}{d}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\mathrm{Rad}} \sup_{\boldsymbol{a}\in A} \langle \boldsymbol{r}, \boldsymbol{a}\rangle \,, \tag{6}$$

where $\mathrm{Rad}$ denotes the Rademacher distribution (https://en.wikipedia.org/wiki/Rademacher_distribution).

Given a hypothesis class $\mathcal{H}$, a training set $S = \{\boldsymbol{z}_i\}_{i\in[N]}$, a predictor $h \in \mathcal{H}$, and a loss $\ell$, we write $\ell \circ \mathcal{H} \doteq \{\ell \circ h : h \in \mathcal{H}\}$, $(\ell \circ h)[S] \doteq [\ell(h, \boldsymbol{z}_1), \ldots, \ell(h, \boldsymbol{z}_N)]$, and $(\ell \circ \mathcal{H})[S] \doteq \{(\ell \circ h)[S] : h \in \mathcal{H}\}$. The empirical Rademacher complexity of $\ell \circ \mathcal{H}$ is defined as

$$\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}) \doteq \mathsf{RC}((\ell \circ \mathcal{H})[S]) = \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\mathrm{Rad}} \sup_{h\in\mathcal{H}} \langle \boldsymbol{r}, (\ell \circ h)[S]\rangle \,, \tag{7}$$

i.e., the Rademacher complexity of the set $(\ell \circ \mathcal{H})[S]$. Moreover, the Rademacher complexity of $\ell \circ \mathcal{H}$ is the expectation of the empirical Rademacher complexity, where the expectation is taken over the randomness of the training set $S$:

$$\mathsf{RC}_N(\ell \circ \mathcal{H}) \doteq \mathbb{E}_{S\sim\mathcal{D}_{\mathcal{Z}}^N} \widehat{\mathsf{RC}}((\ell \circ \mathcal{H})[S]) \,. \tag{8}$$

(a) First, we study several properties of the Rademacher complexity.

(i) **nonnegativity** The famous Jensen's inequality says that: for any random vector $\boldsymbol{v}$ taking values in a convex set $C \in \mathbb{R}^d$ and any convex function $f : \mathbb{R}^d \to \mathbb{R}$ defined over $C$,

$$f(\mathbb{E}\boldsymbol{v}) \le \mathbb{E}f(\boldsymbol{v}). \tag{9}$$

Use Jensen's inequality to prove that the empirical Rademachar complexity $\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H})$ is always non-negative, regardless of the $\ell$ and $\mathcal{H}$. (**Hint: review operation rules of convex functions**) $(1/15)$

(ii) **monotonicity** Assume that $\mathcal{H}_1 \subset \mathcal{H}_2$. Prove $\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}_1) \le \widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}_2)$. $(0.5/15)$

(iii) **summation** Prove that $\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}_1 + \ell \circ \mathcal{H}_2) = \widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}_1) + \widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}_2)$ for any two hypothesis classes $\mathcal{H}_1, \mathcal{H}_2$. $(0.5/15)$

(iv) **affine transform** For real numbers $a, b$, consider $a(\ell \circ \mathcal{H}) + b$. Prove that

$$\widehat{\mathsf{RC}}_S(a(\ell \circ \mathcal{H}) + b) = |a|\,\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}) \,. \tag{10}$$

(**Hint: $\boldsymbol{r}$ and $-\boldsymbol{r}$ have the same distribution for $\boldsymbol{r} \sim_{iid} \mathrm{Rad}$, so for any function $f(\boldsymbol{r})$, $\mathbb{E}_{\boldsymbol{r}\sim_{iid}\mathrm{Rad}} f(\boldsymbol{r}) = \mathbb{E}_{\boldsymbol{r}\sim_{iid}\mathrm{Rad}} f(-\boldsymbol{r})$.**) $(1/15)$

(v) **Talagrand's contraction lemma**[1] Let the loss $\ell : \mathbb{R} \to \mathbb{R}$ be an $L$-Lipschitz function, i.e., $|\ell(a) - \ell(b)| \leq L\,|a - b|$ for all $a, b \in \mathbb{R}$. It holds that

$$\widehat{\mathsf{RC}}_S\left(\ell \circ \mathcal{H}\right) \leq L \cdot \widehat{\mathsf{RC}}_S\left(\mathcal{H}\right). \tag{11}$$

In other words, if the predictors are scalar-valued and the loss $\ell$ is a Lipschitz function, we do not need to worry much about $\ell$, but we can focus on estimating $\widehat{\mathsf{RC}}_S\left(\mathcal{H}\right)$, which makes the measure of complexity relatively independent of the loss chosen—more intuitive. Now, consider two hypothesis classes,

$$\mathcal{H}_1 \doteq \{\boldsymbol{x} \mapsto \sigma(\boldsymbol{u}^\mathsf{T}\boldsymbol{x}) : \|\boldsymbol{u}\|_2 \leq \gamma\} \text{ where } \sigma(z) = 1/(1+e^{-z}) \tag{12}$$

$$\mathcal{H}_2 \doteq \{\boldsymbol{x} \mapsto s(\boldsymbol{u}^\mathsf{T}\boldsymbol{x}) : \|\boldsymbol{u}\|_2 \leq \gamma, s \in \{+1, -1\}\}. \tag{13}$$

Make use of the Talagrand's contraction lemma to derive an upper bound of $\widehat{\mathsf{RC}}_S\left(\mathcal{H}_1\right)$ in terms of $\widehat{\mathsf{RC}}_S\left(\mathcal{H}_2\right)$. (1/15)

(b) As mentioned in class, compared to the Rademacher complexity, Gaussian width/complexity of a set $A \subset \mathbb{R}^d$ replaces the iid Rademacher vector with an iid Gaussian vector, i.e.,

$$\mathsf{G}\left(A\right) \doteq \frac{1}{d}\mathbb{E}_{\boldsymbol{g}\sim_{iid}\mathsf{N}(0,1)} \sup_{\boldsymbol{a}\in A} \langle \boldsymbol{g}, \boldsymbol{a}\rangle. \tag{14}$$

These two complexity measures are closely related and can be shown that for any set $A \subset \mathbb{R}^d$,

$$\sqrt{\frac{2}{\pi}}\mathsf{RC}\left(A\right) \leq \mathsf{G}\left(A\right) \leq 2\sqrt{\log d}\;\mathsf{RC}\left(A\right). \tag{15}$$

But the Gaussian complexity may be easier to estimate in most cases, due to the rich collection of results on Gaussian random processes in the literature; see, e.g., [Ver18].

To define the Gaussian complexity of a hypothesis class, we simply emulate what we do for the Rademacher complexity of a hypothesis class, i.e., to have quantities such as $\widehat{\mathsf{G}}_S(\ell \circ \mathcal{H})$ and $\mathsf{G}_N(\ell \circ \mathcal{H})$. Also, all the properties that we prove in (a) for Rademacher complexities also hold for Gaussian complexities.

Given $N$ data points $\{\boldsymbol{x}_i\}_{i\in[N]}$, or in matrix form $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, provide an upper bound for the empirical Gaussian complexity of the set of linear functions with bounded $\ell_2$ norm (i.e., $\mathcal{H}_{\ell_2} \doteq \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x}\rangle : \|\boldsymbol{w}\|_2 \leq 1\}$), which boils down to upper bounding

$$\frac{1}{d}\mathbb{E}_{\boldsymbol{g}\sim_{iid}\mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_2\leq 1} \langle \boldsymbol{X}\boldsymbol{w}, \boldsymbol{g}\rangle. \tag{16}$$

(**Hint: You can emulate the process of upper bounding the empirical Rademacher complexity of $\mathcal{H}_{\ell_2}$ in the proof of Lemma 26.10 (Section 26.2) of the book [SSBD14][2]; Jensen's inequality implies that for any function $f$ and random variable $\boldsymbol{v}$, $\mathbb{E}_{\boldsymbol{v}}\,\|f\left(\boldsymbol{v}\right)\|_2 \leq \mathbb{E}_{\boldsymbol{v}}\,\|f\left(\boldsymbol{v}\right)\|_2^2)^{1/2}$.** (1.5/15)

(c) Similarly, provide a reasonable upper bound of the empirical Gaussian complexity of $\mathcal{H}_{\ell_\infty} \doteq \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x}\rangle : \|\boldsymbol{w}\|_\infty \leq 1\}$, i.e., the bounded $\ell_\infty$ case. (**Hint: The proof of Lemma 26.11 (Section 26.2) in [SSBD14] is helpful.** ) (1.5/15)

[1] Due to the famous mathematician and Abel Prize laureate Michel Talagrand (https://en.wikipedia.org/wiki/Michel_Talagrand).

[2] Available online: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html

**Problem 3 (Learning with infinite hypothesis class with the growth function and the VC dimension; 3.5/15)** For binary classification problems, the growth function and the VC dimension are useful alternative complexity measures that can be easier to estimate than the Rademacher complexity.

Consider $S = \{\boldsymbol{x}_i\}_{i\in[N]}$, $h(S) \doteq [h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_N)]$, and $\mathcal{H}(S) \doteq \{h(S) : h \in \mathcal{H}\}$. The growth function (also called the shattering coefficient) of $\mathcal{H}$, as a function of $N$ (for any integer $N \geq 1$), is defined as

$$\Pi_{\mathcal{H}}(N) = \max_{S:|S|=N} |\mathcal{H}(S)| \,. \tag{17}$$

In other words, if we call each distinct binary labeling $h(S)$ of $S$ a *dichotomy*, the growth function counts the number of distinct dichotomies realizable by $\mathcal{H}$. Clearly, $\Pi_{\mathcal{H}}(N) \leq 2^N$, as there is a total of $2^N$ dichotomies on an $S$ with $|S| = N$.

We say that a set of points $S$ is *schattered* by a hypothesis class $\mathcal{H}$, if each dichotomy on $S$ can be realized by a certain $h \in \mathcal{H}$, that is, if $\mathcal{H}$ realizes all possible dichotomies of $S$. The VC dimension of $\mathcal{H}$, denoted as $\mathsf{VCdim}(\mathcal{H})$, is the size of the largest set that can be shattered by $\mathcal{H}$, or equivalently,

$$\mathsf{VCdim}(\mathcal{H}) = \max\left\{N : \Pi_{\mathcal{H}}(N) = 2^N\right\}. \tag{18}$$

Obviously, $\Pi_{\mathcal{H}}(N) < 2^N$ for any $N > \mathsf{VCdim}(\mathcal{H})$. Hence, to prove $\mathsf{VCdim}(\mathcal{H}) = d$, one needs to show that

- there exists a set of size $d$ that can be shattered by $\mathcal{H}$, **and**
- no set of cardinality $d+1$ can be shattered by $\mathcal{H}$.

(a) Consider the set of decision stumps in $\mathbb{R}$, i.e., $\mathcal{H}^1_{DS} \doteq \{x \mapsto b\,\mathrm{sign}\,(x-\theta) : \theta \in \mathbb{R}, b \in \{\pm 1\}\}$. What's $\Pi_{\mathcal{H}^1_{DS}}(N)$, and what's $\mathsf{VCdim}(\mathcal{H}^1_{DS})$? (1/15)

(b) Let $\mathcal{H}$ be a finite hypothesis class. Prove that $\mathsf{VCdim}(\mathcal{H}) \leq \lfloor \log_2 |\mathcal{H}| \rfloor$. (0.5/15)

(c) Consider hypothesis classes $\mathcal{A}$ and $\mathcal{B}$, both having finite VC-dimensions, say $\mathsf{VCdim}(\mathcal{A}) = d_{\mathcal{A}}$ and $\mathsf{VCdim}(\mathcal{B}) = d_{\mathcal{B}}$. Let $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$.

(i) Prove that $\Pi_{\mathcal{C}}(N) \leq \Pi_{\mathcal{A}}(N) + \Pi_{\mathcal{B}}(N)$. (0.5/15)

(ii) Sauer's lemma states this: for a hypothesis class $\mathcal{H}$ with $\mathsf{VCdim}(\mathcal{H}) = d$. For any integer $N \geq 1$, $\Pi_{\mathcal{H}}(N) \leq \sum_{i=0}^{d} \binom{N}{i}$. Use Sauer's lemma to show that for $N \geq d_{\mathcal{A}} + d_{\mathcal{B}} + 2$, $\Pi_{\mathcal{C}}(N) < 2^N$, and thereby give an upper bound on $\mathsf{VCdim}(\mathcal{C})$. (0.5/15)

(d) Provide an upper bound of $\widehat{\mathsf{RC}}_S(\mathcal{H})$ where $\mathcal{H}_{HC} = \left\{\boldsymbol{x} \mapsto \mathrm{sign}\,(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b) : \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\}$, i.e., the set of hyperplane classifiers. (**Hint: You may want to use these facts: (1) for any $\mathcal{H}$ consisting of binary classifiers, $\widehat{\mathsf{RC}}_S(\mathcal{H}) \leq \sqrt{2\log_2 \Pi_{\mathcal{H}}(N)/N}$ ;** (2) For any $\mathcal{H}$ with $\mathsf{VCdim}(\mathcal{H}) = d$, $\Pi_{\mathcal{H}}(N) \leq (eN/d)^d$ for all $N \geq d$; (3) $\mathsf{VCdim}(\mathcal{H}_{HC}) = d+1$. ) (1/15)

## References


[SSBD14] Shai Shalev-Shwartz and Shai Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge university press, 2014.

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.