CS/ECE/STAT-861: Theoretical Foundations of Machine Learning
University of Wisconsin–Madison, Fall 2023

Homework 1. Due 10/06/2023, 11.00 am

Instructions:

1. Homework is due a [illegible] Please hand over your homework at the beginning of class. *Please see the course webs [illegible] submission.*

2. I recommend that y [illegible] k using LaTeX. You will receive 5 percent extra credit if you do so. If you are submitting hand-written homeworks, please make sure it is cleanly written up and legible. I will not invest undue effort to understand bad handwriting.

3. You *must* hand in a hard copy of the homework. The only exception is if you are out of town in which case you must let me know ahead of time and email me a copy of the homework by 11 am on the due date. If this is the case, your homework *must* be typeset using LaTeX. Please do *not* email written and scanned copies.

4. Unless otherwise specified, you may use any result we have already proved in class. You do not need to prove them from scratch, but clearly state which result you are using.

5. Solutions to some of the problems may be found as examples or exercises in the suggested textbooks or other resources. You are encouraged to try the problems on your own first before searching for the solution. If you find an existing solution, first read and understand the proof, and then write it in your own words. Please indicate any references you have used at the beginning of your solution when you turn in your homework.

6. **Collaboration:** You are allowed to collaborate on in groups of size up to 3 on each problem. If you do so, please indicate your collaborators at the beginning of your solution.

1 PAC Learning and Empirical Risk Minimization

1. **[4 pts]** *(What is wrong with this proof?)* We perform empirical risk minimization (ERM) in a finite hypothesis class $\mathcal{H}$ using an i.i.d dataset S of n points. Let $h^\star \in \text{argmin}_{h\in\mathcal{H}} R(h)$ be an optimal classifier in the class, and let $\widehat{h} \in \text{argmin}_{h\in\mathcal{H}}$ [illegible] ical risk of the dataset S. A student offers the following proof and claims that it is pos [illegible] ion error without any dependence on $|\mathcal{H}|$.

   (i) Let $B_1 = \{\widehat{R}($ [illegible] the bad event that the empirical risk of $h^\star$ is ϵ larger than its true risk. By Hoeff [illegible] $\mathbb{P}(B_1) \le e^{-2n\epsilon^2}$.

   *(ii) Similarly, Let [illegible] $\epsilon\}$ denote the bad event that the empirical risk of $\widehat{h}$ is ϵ smaller than its true r [illegible] ality we have $\mathbb{P}(B_2) \le e^{-2n\epsilon^2}$. (**correction:** This previously said $2e^{-2n\epsilon^2}$. Than [illegible] this out. –KK)*

   As $\widehat{R}(\widehat{h}) \le \widehat{R}(h^\star)$, we have,

$$R(\widehat{h}) - R(h^\star) \le R(\widehat{h}) - \widehat{R}(\widehat{h}) + \widehat{R}(h^\star) - R(h^\star) \le 2\epsilon$$

   under the good event $G = B_1^c \cap B_2^c$ which is true with probability at least $1 - 2e^{-2n\epsilon^2}$. This result does not depend on $|\mathcal{H}|$ and even applies to infinite hypothesis classes provided there exists $h^\star$ which minimizes the risk.

   Which sentence below best describes the mistake (if any) with this proof? State your with an explanation. If you believe there is a mistake, be as specific as possible as to what the mistake is.

   (a) Both statement *(i)* and statement *(ii)* are incorrect.

   (b) Only statement *(i)* is incorrect. Statement *(ii)* is correct.

   (c) Only statement *(ii)* is incorrect. Statement *(i)* is correct.

   (d) Both statements are correct. There is nothing wrong with this proof.

2. **[6 pts]** *(PAC bound)* Prove the following result which was presented but not proved in class.

   Let $\mathcal{H}$ be a hypothesis class with finite $\text{Rad}_n(\mathcal{H})$. Let $\widehat{h}$ be obtained via ERM using n i.i.d samples. Let $\epsilon > 0$. Then, there exists universal constants C_1, C_2 such that with probability at least $1 - 2e^{-2n\epsilon^2}$, we have

$$R(\widehat{h}) \le \inf_{h\in\mathcal{H}} R(h) + C_1 \text{Rad}_n(\mathcal{H}) + C_2\epsilon.$$

3. **[3 pts]** *(Sample complexity based on VC dimension)* Say $\mathcal{H}$ has a finite VC dimension d. Let $\delta \in (0,1)$. Using the result/proof in part 2 or otherwise, show that there exist universal constants C_3, C_4 such that when $n \ge d$, the following bound holds with probability at least $1 - \delta$.

$$R(\widehat{h}) \le \inf_{h\in\mathcal{H}} R(h) + C_3\sqrt{\frac{d\log(n/d) + d}{n}} + C_4\sqrt{\frac{1}{n}\log\left(\frac{2}{\delta}\right)}.$$

4. **[3 pts]** *(Bound on the expected risk)* The above results show that $R(\widehat{h})$ is small with high probability. Using the results/proofs in parts 2 and 3 or otherwise, show that it is also small in expectation. Specifically, show that there exist universal constants C_5, C_6 such that the following bound holds.

$$\mathbb{E}[R(\widehat{h})] \le \inf_{h\in\mathcal{H}} R(h) + C_5\sqrt{\frac{d\log(n/d) + d}{n}} + C_6\sqrt{\frac{\log(4n)}{n}} + \frac{1}{\sqrt{n}}.$$

   Here, the expectation is with respect to the dataset S.

For parts 2, 3, and 4, of this question, if you can prove a bound that has similar higher order terms but differs in additive/multiplicative constants or poly-logarithmic factors, you will still receive full credit.

2 Rademacher Complexity & VC dimension

1. **[5 pts]** *(Empirical Rademacher complexity)* Consider a binary classification problem with the 0–1 loss $\ell(y_1, y_2) = \mathbb{1}(y_1 \neq y_2)$ and where $\mathcal{X} = \mathbb{R}$. Consider the following dataset $S = \{(x_1 = 0, y_1 = 0), (x_2 = 1, y_2 = 1)\}$.

   (a) Let $\mathcal{H}_1 = \{h_a$ [illegible] be the hypothesis class of one-sided threshold functions. Compute the empirical [illegible] $\widehat{\mathrm{Rad}}(S, \mathcal{H}_1)$.

   (b) Let $\mathcal{H}_2 = \{h_a$ [illegible] $\cup \{h_a(x) = \mathbb{1}(x \leq a); a \in \mathbb{R}\}$ be the class of two-sided threshold functions. Co[illegible]macher complexity $\widehat{\mathrm{Rad}}(S, \mathcal{H}_2)$.

   (c) Are the values [illegible]nt with the fact that $\mathcal{H}_1 \subset \mathcal{H}_2$?

2. **[6 pts]** *(VC dimen*[illegible] Consider a binary classification problem where $\mathcal{X} = \mathbb{R}^D$ is the D-dimensional Euc[illegible] of linear classifiers is given by $\mathcal{H} = \{h_{w,b}(x) = \mathbb{1}[w^\top x + b \geq 0]; w \in \mathbb{R}^D, b \in \mathbb{R}\}$. Prove that the VC dimension of this class is $d_{\mathcal{H}} = D + 1$. (**correction:** Previously this said $\mathcal{H} = \{h_{w,b}(x) = w^\top x + b \geq 0 \; w \in \mathbb{R}^d, b \in \mathbb{R}\}$. –KK)

3. *(Interval classifiers)* Let $\mathcal{X} = \mathbb{R}$. Consider the class of interval classifiers, given by

$$\mathcal{H} = \{h_{a,b}(x) = \mathbb{1}(a \leq x \leq b); a, b \in \mathbb{R}, a \leq b\}.$$

   (a) **[4 pts]** What is the VC dimension d of this class?

   (b) **[8 pts]** Show that Sauer's lemma is tight for this class. That is, for all n, show that $g(n, \mathcal{H}) = \sum_{i=0}^{d} \binom{n}{i}$.

4. *(Union of interval classifiers)* Let $\mathcal{X} = \mathbb{R}$. Consider the class of the union of K interval classifiers, given by

$$\mathcal{H} = \{h_{a,b}(x) = \mathbb{1}(\exists k \in \{1, \ldots, K\} \text{ s.t. } a_k \leq x \leq b_k); a, b \in \mathbb{R}^k, a_k \leq b_k \forall k\}.$$

   (a) **[4 pts]** What is the VC dimension d of this class?

   (b) **[8 pts]** Show that Sauer's lemma is tight for this class. That is, for all n, show that $g(n, \mathcal{H}) = \sum_{i=0}^{d} \binom{n}{i}$.

   Hint: The following identity from combinatorics, which we used in the proof of Sauer's lemma, may be helpful.

$$\forall\, m > k, \quad \binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}.$$

5. **[6 pts]** *(Tightness of Sauer's lemma)* Prove the following statement about the tightness of Sauer's lemma when $\mathcal{X} = \mathbb{R}$: For all $d > 0$, there exists a hypothesis class $\mathcal{H} \subset \{h : \mathbb{R} \to \{0, 1\}\}$ with VC dimension $d_{\mathcal{H}} = d$ such that, for all dataset sizes $n > 0$, we have $g(n, \mathcal{H}) = \sum_{i=0}^{d} \binom{n}{i}$.

   Keep in mind that the hypothesis class $\mathcal{H}$ should depend on d but not on n.

   Hint: One approach will be to use the results from part 4 which will allow you to prove the results for even d. You should consider a different hypothesis class to show this for odd d.

3 Relationship between divergences

Let P, Q be probabilities with densities p, q respectively. Recall the following divergences we discussed in class

KL divergence: $\mathrm{KL}(P, Q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x) \mathrm{d}x$.

Total variation distance: $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$.

L_1 distance: $\|P - Q\|_1 = \int |p(x) - q(x)| \mathrm{d}x$.

Hellinger distance: $\mathrm{H}^2(P, Q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$.

Finally, let $\|P \wedge Q\| = \int \min(p(x), q(x))\mathrm{d}x$ denote the affinity between two distributions. When we have n i.i.d observations, let P^n, Q^n denote the product distributions.

(**correction:** Previously, the definition of the Hellinger distance said H and not H^2. Thanks to Yixuan for pointing this out. –KK)

Prove the following stat

1. **[3 pts]** $\mathrm{KL}(P^n, Q^n$
2. **[3 pts]** $\mathrm{H}^2(P^n, Q^n$ $))^n$.
3. **[3 pts]** $\mathrm{TV}(P, Q) =$
   Hint: Can you rela on to the set $A = \{x; p(x) > q(x)\}$?
4. **[3 pts]** $\mathrm{TV}(P, Q) =$
5. **[3 pts]** $\mathrm{H}^2(P, Q) \leq \|P - Q\|_1$.
   Hint: What can you say about $(a-b)^2$ and $|a^2 - b^2|$ when $a, b > 0$?