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

Homework 0. Due 9/15/2023, 11.00 am

Instructions:

1. Homework is due at 11 am on the due date. Please hand over your homework at the beginning of class. *Please see the course website for the policy on late submission.*
2. I recommend that you typeset your homework 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. One of the objectives of Homework 0 is to test if you are familiar with the necessary background topics to take this class. While I do not expect you to know the solution right away, you should be able to solve most of the questions with reasonable effort after looking up any necessary references. If you find this homework exceedingly difficult (especially problems 1 and 2), please talk to me.
5. **Collaboration:** You are allowed to collaborate on problem 3 of this homework in groups of size up to 3. If you do so, please indicate your collaborators at the top of your solution. You are *not* allowed to collaborate on problems 1 and 2.

1 Estimating the mean of a normal distribution

You are given n independent samples $X_1, \ldots, X_n$ sampled from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with *unknown* mean μ, but known variance σ^2. You wish to estimate the mean μ. One straightforward option is to estimate μ using the *sample mean* $\widehat{\mu}_{\mathrm{SM}}$, defined below:

$$\widehat{\mu}_{\mathrm{SM}} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

1. **[2 pts]** To quantify the performance of this estimator, we define the risk R, which is simply the expected squared error of an estimator,

   $$R(\widehat{\mu}_{\mathrm{SM}}, \mu) = \mathbb{E}[(\widehat{\mu}_{\mathrm{SM}} - \mu)^2].$$

   Here, the expectation $\mathbb{E}$ is with respect to the randomness in the data. Show that the risk is σ^2/n for the sample mean estimator.

2. **[4 pts]** *(Concentration)* While the risk measures how well an estimator does in expectation, sometimes we also wish to know that $\widehat{\mu}_{\mathrm{SM}}$ is within some margin of error ϵ of the true mean μ with high probability. Prove the following result for any $\epsilon > 0$:

   $$\mathbb{P}(|\widehat{\mu}_{\mathrm{SM}} - \mu| > \epsilon) \leq 2\exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).$$

   where the probability $\mathbb{P}$ is with respect to the randomness in the data.

   You may use the following facts about normal random variables:

   - If $X_1, \ldots, X_n$ are normal, then so is $\sum_{i=1}^{n} X_i$. (You will need to compute the mean and variance.)
   - If X are normal, then so is aX for any $a \in \mathbb{R}$. (You will need to compute the mean and variance.)
   - If $Z \sim \mathcal{N}(0,1)$ is a standard normal random variable, then $\mathbb{P}(|Z| > \epsilon) \leq 2e^{-\epsilon^2/2}$.

3. **[2 pts]** *(Sample complexity)* Suppose you are given some $\epsilon > 0$ and $\delta \in (0,1)$. You wish to collect enough samples so that your estimator is within an ϵ margin of error with probability at least δ. Show that if $n \geq \frac{2\sigma^2}{\epsilon^2}\log(2/\delta)$, we will have the following guarantee: $\mathbb{P}(|\widehat{\mu}_{\mathrm{SM}} - \mu| > \epsilon) \leq \delta$.

   N.B: In the first part of the class, we will study a procedure called *empirical risk minimization* for learning, which essentially chooses a model that performs well on the observed data. After learning, we need to demonstrate that our learned model will do well on future unobserved data. Hence, we need to find conditions under which a model's performance on the observed dataset will translate to its future generalization performance. Concentration will be an important tool in such proofs.

2 Which estimator is better?

The sample mean is just one of several possible estimators for μ. Student A proposes the following alternative estimator $\widehat{\mu}_\alpha$ with some parameter $\alpha \in (0,1)$,

$$\widehat{\mu}_\alpha = \frac{\alpha}{n}\sum_{i=1}^{n} X_i.$$

In this question, we will explore if and when $\widehat{\mu}_\alpha$ could be a better estimator than $\widehat{\mu}_{\mathrm{SM}}$.

1. **[2 pts]** *(Bias–variance decomposition)* First, show that the following holds for any estimator $\widehat{\mu}$,

   $$R(\widehat{\mu}, \mu) = \mathbb{E}[(\widehat{\mu} - \mu)^2] = \underbrace{(\mathbb{E}[\widehat{\mu}] - \mu)^2}_{\text{bias}} + \underbrace{\mathbb{E}[(\widehat{\mu} - \mathbb{E}[\widehat{\mu}])^2]}_{\text{variance}}.$$

2. **[2 pts]** Using the result from part 1, compute the risk of the estimator $\widehat{\mu}_\alpha$. Note that, unlike the sample mean, the risk of $\widehat{\mu}_\alpha$ depends on the true mean μ.

3. **[2 pts]** Show that there exists at least one value for μ such that $\widehat{\mu}_\alpha$ is a strictly better estimator than $\widehat{\mu}_{\mathrm{SM}}$. That is, there exists $\mu \in \mathbb{R}$, such that, for all $\alpha \in (0,1)$, we have $R(\widehat{\mu}_\alpha, \mu) < R(\widehat{\mu}_{\mathrm{SM}}, \mu)$.

4. **[4 pts]** *(Maximum risk)* Despite the result from part 3, Student B is not satisfied with Student A's proposition, as an estimator should perform well for all values of μ, and not just for one value of μ. In particular, they argue that the worst-case risk over all μ should be small. They propose the following criterion, the *maximum risk* $\mathcal{R}$, as a way to measure how well an estimator performs.

$$\mathcal{R}(\widehat{\mu}) = \sup_{\mu \in \mathbb{R}} R(\widehat{\mu}, \mu) = \sup_{\mu \in \mathbb{R}} \mathbb{E}[(\widehat{\mu} - \mu)^2].$$

   (a) Compute $\mathcal{R}(\widehat{\mu}_{\mathrm{SM}})$ and $\mathcal{R}(\widehat{\mu}_\alpha)$.

   (b) Based on the above answers, which estimator would you choose?

5. **[5 pts]** *(Maximum risk over a bounded domain)* Suppose we had prior knowledge that $\mu \in [0,1]$. While student A agrees with student B's criterion, they argue that we should modify the definition of the maximum risk to incorporate this prior knowledge. They propose the following definition:

$$\mathcal{R}'(\widehat{\mu}) = \sup_{\mu \in [0,1]} R(\widehat{\mu}, \mu) = \sup_{\mu \in [0,1]} \mathbb{E}[(\widehat{\mu} - \mu)^2].$$

   (a) Compute $\mathcal{R}(\widehat{\mu}_{\mathrm{SM}})$ and $\mathcal{R}(\widehat{\mu}_\alpha)$, the maximum risk for the two estimators discussed above?

   (b) Is there any particular value of α (possibly dependent on n and σ) for which $\mathcal{R}'(\widehat{\mu}_\alpha) < \mathcal{R}'(\widehat{\mu}_{\mathrm{SM}})$?

   (c) Based on the above answer, which estimator would you choose? Intuitively, explain the discrepancy in the conclusions in part 4 and part 5.

N.B: In the second part of the class, we will explore the concept of *minimax optimality*. Here, we will design algorithms that have small maximum risk over a class of distributions, and establish lower bounds to prove that no other algorithm can have a significantly smaller maximum risk. We will start with some simple estimation problems, and extend the ideas to regression, classification, density estimation, online learning, and bandits.

3 Understanding exploration–exploiting trade-offs

This question is more difficult than the two previous questions. You are allowed to collaborate on this question with up to two classmates.

Consider the following game which proceeds over T rounds. You have access to two normal distributions $\nu_1 = \mathcal{N}(\mu_1, \sigma^2)$ and $\nu_2 = \mathcal{N}(\mu_2, \sigma^2)$, where σ^2 is known but $\mu_1, \mu_2 \in [0,1]$ are not. On each round t, you get to choose one distribution $I_t \in \{1,2\}$, where $I_t = i$ corresponds to choosing $\nu_i = \mathcal{N}(\mu_i, \sigma^2)$. Once you draw the sample, you earn a monetary reward that is equal to the value of the sample. If the value of the sample is negative, you should pay that amount instead. Your total cumulative reward, over T rounds is $\sum_{t=1}^T X_t$ where $X_t \sim \nu_{I_t}$. We will measure how well we perform via our *average regret*, defined below:

$$R_T = \max\{\mu_1, \mu_2\} - \frac{1}{T} \sum_{t=1}^{T} X_t.$$

We wish to design an algorithm whose average regret vanishes[1] with T in expectation, i.e $\mathbb{E}[R_T] \to 0$ as $T \to \infty$.

Algorithm: A student proposes the following simple algorithm. First sample each of the distributions N times (where $N < T/2$). Then, for the remaining $T - 2N$ rounds, sample the distribution with the highest observed sample mean

[1] Intuitively, if we knew *a priori* which mean was larger, we will always pull the arm with the highest mean and have $\mathbb{E}[R_T] = 0$ as $\frac{1}{T}\mathbb{E}[\sum_t X_t] = \max\{\mu_1, \mu_2\}$. If $\mathbb{E}[R_T] \to 0$, this means we are able to learn which of the two distributions has a larger mean and converge towards the correct answer as we collect more samples.

using the N samples. That is, I_t is chosen as follows:

$$I_t = \begin{cases} 1 & \text{if } t \leq N, \\ 2 & \text{if } N+1 \leq t \leq 2N, \\ 1 & \text{if } t > 2N \text{ and } \widehat{\mu}_1 \geq \widehat{\mu}_2, \\ 2 & \text{if } t > 2N \text{ and } \widehat{\mu}_1 < \widehat{\mu}_2. \end{cases}$$

$$\text{where, } \widehat{\mu}_1 = \sum_{t=1}^{N} X_t, \qquad \widehat{\mu}_2 = \sum_{t=N+1}^{2N} X_t,$$

For what follows, let $\Delta = |\mu_1 - \mu_2|$ denote the gap between the two means.

1. **[5 pts]** *(Regret decomposition)* Establish the following identity for the expected average regret:

$$\mathbb{E}[R_T] = \frac{N\Delta}{T} + \frac{(T-2N)\Delta}{T}\Phi\left(-\Delta\sqrt{\frac{N}{2\sigma^2}}\right)$$

   Here, $\Phi(x) = \mathbb{P}_{Z\sim\mathcal{N}(0,1)}(Z < x)$ is the CDF of the standard normal distribution.

2. **[2 pts]** Using the result from part 1 and the fact that $\mu_1, \mu_2 \in [0,1]$, show the following upper bound on the expected average regret.

$$\mathbb{E}[R_T] \leq \frac{N}{T} + \Delta \exp\left(\frac{-N\Delta^2}{4\sigma^2}\right).$$

   You may use the following property about standard normals, which is a one-sided version of the inequality given in problem 1. If $Z \sim \mathcal{N}(0,1)$, then $\mathbb{P}(Z < -\epsilon) \leq e^{-\epsilon^2/2}$.

3. **[3 pts]** Use the result in part 2 to show the following upper bound.

$$\mathbb{E}[R_T] \leq \frac{N}{T} + C\frac{\sigma}{\sqrt{N}}, \qquad \text{where, } C = \sqrt{2}e^{-1/2}.$$

   Hint: Consider the function $f(x) = \log(x) - \alpha x^2$, where $\alpha > 0$. What is the maximizer of f?

4. **[2 pts]** *(An optimal choice of N.)* Specify a choice for N, depending only on σ and T, so that the upper bound in part 3 is minimized. Are you able to achieve $\mathbb{E}[R_T] \to 0$ as $T \to \infty$? If so, at what rate does it go to zero?

5. **[2 pts]** *(Exploration–exploitation trade-off.)* Let $N^\star$ denote the optimal choice in part 4. In words, explain what would happen had we chosen $N \ll N^\star$ or $N \gg N^\star$.

N.B: In the third part of the class, we will study several models for adaptive decision-making. The model discussed in this question is an example of a *stochastic bandit*, which is one paradigm for decision-making. In bandit settings, we often have to trade-off between *exploration* (learning about the environment) and *exploitation* (leveraging what we have learned to maximize rewards). The above algorithm is a simple, albeit sub-optimal, procedure where we have an explicit exploration phase (first $2N$ rounds) and exploitation phase (last $T - 2N$ rounds). In class, we will look at better algorithms to manage this trade-off which have faster rates of convergence.