CMPUT 655: Theoretical Foundations of Machine Learning, Fall 2023
Homework #1

Instructions

Submissions You need to submit a single PDF file, named `p01_<name>.pdf` where `<name>` is your name. The PDF file should include your typed up solutions (we strongly encourage to use pdfLATEX). Write your name in the title of your PDF file. We provide a LATEXtemplate that you are encouraged to use. To submit your PDF file you should send the PDF file via private message to Csaba on Slack before the deadline.

Collaboration and sources Work on your own. You can consult the problems with your classmates, use books or web, papers, etc. Also, the write-up must be your own and you must acknowledge all the sources (names of people you worked with, books, webpages etc., including class notes.) Failure to do so will be considered cheating. Identical or similar write-ups will be considered cheating as well. Students are expected to understand and explain all the steps of their proofs.

Scheduling Start early: It takes time to solve the problems, as well as to write down the solutions. Most problems should have a short solution (and you can refer to results we have learned about to shorten your solution). Don't repeat calculations that we did in the class unnecessarily.

Other Some problems get zero points. These are practice problems that will not be marked.

Deadline: September 24 at 11:55 pm

Preliminaries

In the problems and throughout the course, we use the notation of standard probability theory. The notation, and the basic rules of calculus with expectations and probabilities are summarized in Chapter 2 of the bandit book, which can be accessed for free here. It may be useful to briefly skim through this chapter before embarking on solving the problems. In the solutions, feel free to use any results mentioned in this chapter, or your favourite book on probabilities. One way I was deviating from the book in the lectures is to write $P(dy)$ instead of $dP(y)$ for integrals over measure P where the integrand uses the variable name y. This is just a variation of notation and you should feel free to use either of them (perhaps not both, at least not in the same expression, or close to each other).

As it will be useful for the first problem, let us recall the following: If X is a random element taking values in some space $\mathcal{X}$ and P_X denotes the pushforward of X[1] then for any real-valued, $f : \mathcal{X} \to \mathbb{R}$ measurable function,

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x) P_X(dx)\,, \tag{1}$$

provided that either the right-hand side, or the left-hand side exist. Some people call this the *law of the unconscious statistician*, or LOTUS. This fact is mentioned in the last paragraph of Section 2.5 of the bandit book. (I find it entertaining and a bit surprising that unconscious statistician would know about push-forward measures and low and behold, usually, this result is given in more elementary texts that state this for some special case only and not for general random element. But this "law" (better: identity) holds in general, so we will just call it LOTUS.)

Another goodie you may need is that if X, Y are independent then the pushforward of $Z = (X, Y)$, P_Z, is a product of the pushforwards of X and Y: $P_Z = P_X \otimes P_Y$. This was left out from the bandit book by accident!

[1] Strictly speaking, we should say P_X is the pushforward of $\mathbb{P}$ under X. This is too long. So we will just say the pushforward of X, or even, the distribution of X.

Problems

In the first problem, we consider the setting of supervised learning: Let

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$$

be a sequence identically distributed, independent random variables[2] taking values in the set $\mathcal{X} \times \mathcal{Y}$. Let the common distribution of these random variables be P.

Elements of $\mathcal{X}$ are called inputs, elements of $\mathcal{Y}$ are called outputs. Let

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$$

be a function, which we call the loss function. For a fixed function $f : \mathcal{X} \to \mathcal{Y}$, let

$$L(f) = \int \ell(f(x), y) P(dx, dy)\,. \tag{2}$$

Here, and in what follows, in these exercises, unless otherwise stated, we assume that the expectations and integrals are (well-)defined.[3] Let

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$$

be a map, which assigns to every n-tuple of input-output pairs a function that maps inputs to outputs. Define

$$f_n = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n))\,.$$

Question 1. Show that the following hold:

(a) For any $f : \mathcal{X} \to \mathcal{Y}$, $L(f) = \mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\ell(f(X_1), Y_1)] = \cdots = \mathbb{E}[\ell(f(X_n), Y_n)]$ whenever any of these terms is well-defined.[4]

2 points

(b) With probability one, it holds that $\mathbb{E}[\ell(f_n(X), Y)|D_n] = L(f_n)$ whenever $\mathbb{E}[L(f_n)]$ exists, where $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$. [5] Here, we must mention that ℓ is a measurable function of its arguments, and so is $\mathcal{A}$.

10 points

Hint: Sometimes the solution becomes clear (and the proof becomes much cleaner) if one proves a more general result. In terms of the tools needed, besides some elementary results, the definition of conditional expectations, Theorem 2.11 from the bandit book and Fubini's theorem are needed. Note that we assume measurability of all maps involved, in particular L (which also follows from the measurability of the other maps). If you are stuck, prove the result for the case when D_n is discrete valued for half the marks.

[2] Purists, like the authors of the bandit book, would call these random elements: According to such purists, random variables are real-valued random elements. This is a bit too much of a formality sometimes. Hence, I may switch between being a purist or not depending on my mood, hoping this won't cause confusion. But in any case, I will *not* rely on that if I say that X is a random variable then it is real-valued. If I do, call me out and I owe you one.

[3] For example, above, for this to hold we would need that $(x, y) \mapsto \ell(f(x), y)$ is measurable with respect to the appropriate measurable spaces and this also means that the integral above gives a finite value. For definitions of these concepts, see Chapter 2 of the bandit book. But in these problems we are not concerned about whether the appropriate maps are measurable. We may occasionally be concerned about whether an integral (expected value) exists. When this is of essence, it will be noted.

[4] Thanks Shuai for this freebee!

[5] Thanks Vlad for this nice problem!

(c) We have $\mathbb{E}[\ell(f_n(X), Y)] = \mathbb{E}[L(f_n)]$ whenever either side exists.

2 points

Hint: Check out Theorem 2.12 (which lists properties of conditional expectations) from the bandit book.

(d) Let $D'_m = (X'_1, Y'_1), \dots, (X'_m, Y'_m)$ be independent, identically distributed sequence of random variables such that the distribution of (X'_1, Y'_1) is also P. Assume that D'_m and D_n are independent. For $f : \mathcal{X} \to \mathcal{Y}$, let

$$L_m(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(X'_i), Y'_i)\,.$$

Show that with probability one it holds that

$$\mathbb{E}[L_m(f_n)|D_n] = L(f_n)\,.$$

assuming $\mathbb{E}[L(f_n)]$ exists.

Hint: Again, Theorem 2.12 from the bandit book, the theorem about the properties of conditional expectations, will be useful.

2 points

(e) In the remaining problems assume that $\mathbb{E}[L(f_n)]$ exist.

Show that with probability one, $L_m(f_n) \to L(f_n)$ as $m \to \infty$. For this problem assume that D_n is a discrete-valued random element (that is, there exists a countable set C such that $\mathbb{P}(D_n \in C) = 1$).

To get full marks, use elementary reasoning together with the strong law of large numbers that states that for independent, identically distributed random variables, the sample mean converges to the common mean with probability one assuming the common mean exist.

10 points

Hint: Recall the law of large numbers.

(f) For the curious minded: Prove that the result of the previous continues to hold even if D_n is not restricted to be discrete-valued.

2 points

Hint: This is hard to show from first principles. If you get stuck, I suggest checking out the book of Chow and Teicher, in particular Section 7.2 in that book. The title of the book is "Probability Theory: Independence, Interchangeability, Martingales" and is available from Springer. I have the third edition.

(g) What changes when D'_m and D_n are not independent? What happens for example when $m = n$ and $D'_m = D_n$? Show that in this case, with probability one,

$$\mathbb{E}[L_m(f_n)|D_n] = \frac{1}{n} \sum_{i=1}^{n} \ell(f_n(X_i), Y_i)\,.$$

2 points

(h) Assuming $D'_m = D_n$, give an example in terms of $(\mathcal{X}, \mathcal{Y}, P, \ell, (\mathcal{A}_n)_{n\geq 1})$ where $\mathcal{A}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$ and P is a joint distribution on $\mathcal{X} \times \mathcal{Y}$, such that for every $n(= m)$, $L_m(f_n) = 0$ with probability one, yet $L(f_n) \geq 0.5$ with probability one. That is, testing on the training set can lead to spurious outcomes.

2 points

Total: **32 points**

The next question is concerned with "trade-offs" or, if you want more drama, no free lunch for statisticians (or machine learners). First, recall that $L(f)$ is the expected loss of a function $f : \mathcal{X} \to \mathcal{Y}$. To discuss tradeoffs, by abusing notation, we overload the definition of L from Eq. (2) to make the dependence of L on P explicit:

$$L(P, f) = \int \ell(f(x), y) P(dx, dy) \,.$$

In class we, informally, defined the goal of choosing $\mathcal{A}$ (the learning rule/map/algorithm/method) to keep the expected losses $L(P, \mathcal{A}) = \mathbb{E}[L(P, \mathcal{A}(D_n))]$ small. Expanding on this expectation, we have

$$L(P, \mathcal{A}) = \int L(P, \mathcal{A}(w)) P^{\otimes n}(dw) = \int \tilde{\ell}(\mathcal{A}(w), z) P^{\otimes n}(dw) P(dz) \,,$$

where, for $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, $\tilde{\ell}(f, z) = \ell(f(x), y)$. Thus, the distribution P appears $n + 1$ times in this expression (why?).

Now, can we design an algorithm $\mathcal{A}^*$ such that

$$L(P, \mathcal{A}^*) = \inf_{\mathcal{A}} L(P, \mathcal{A})$$

holds *for all* P? This is very demanding: $\mathcal{A}^*$, without knowing P, should be as good as an algorithm that is chosen using the "knowledge of P": In the right-hand side above, the value will clearly depend on P (and the choice of $\mathcal{A}$, if any, that minimizes the loss $L(P, \mathcal{A})$, will depend on P). Except for trivial cases, the above goal cannot be achieved. In fact, this is the main reason the field of machine learning is so exciting! If there was a magical solution (and we would know it), the field would be (almost) over.

The next problem will illustrate the core difficulty. As such, the example will be kept simple. In fact, in the example $\mathcal{X}$ is chosen to be a singleton (a set with a single element). As such, any function $f : \mathcal{X} \to \mathcal{Y}$ can be identified with some element of $\mathcal{Y}$ (picking a function is the same as picking an element of $\mathcal{Y}$), and in what follows, we therefore just replace such functions with elements of $\mathcal{Y}$. Moreover, since in the observations (X, Y), the input X carries no information, we will drop the inputs. We also assume there is only one observation, i.e., $n = 1$. Then, learning methods $\mathcal{A}$ are simple (measurable) maps from $\mathcal{Y}$ to $\mathcal{Y}$. Finally, we choose $\mathcal{Y} = \mathbb{R}$ and the loss ℓ is chosen to the quadratic loss:

$$\ell(u, v) = (u - v)^2 \,.$$

It also follows that for $f \in \mathbb{R}$,

$$L(P, f) = \int \ell(f, y) P(dy) = \int (f - y)^2 P(dy) \,. \tag{3}$$

Furthermore, in some part of the next problem we will only consider two distributions, both of them normal with the same variance $\sigma^2 > 0$ but different means, which we will denote by $\mu_1 < \mu_2$. We will denote these by P_1 and P_2.

One way to characterize how different these distributions are is by their relative entropy. By letting $\mathrm{KL}(P,Q)$ denote the relative entropy (also known as the Kullback-Leibler divergence) between two probability measures on the real line P and Q, as it is well known,

$$\mathrm{KL}(P_1,P_2) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2} \tag{4}$$

(and thus, it also follows, that $\mathrm{KL}(P_1,P_2) = \mathrm{KL}(P_2,P_1)$, while, as is well known, the relative entropy is not symmetric). We also let $\mathcal{N}(\mu,\sigma^2)$ denote the normal distribution with mean μ and variance σ^2 over the reals. Thus, $P_i = \mathcal{N}(\mu_i,\sigma^2)$. We will use $\mathcal{M}_1(\mathbb{R})$ to denote the set of probability distributions over the reals.[6] Further, for $P \in \mathcal{M}_1(\mathbb{R})$, we let $\mathrm{Var}(P)$ denote the variance underlying P: $\mathrm{Var}(P) = \int x^2 P(dx) - (\int x P(dx))^2$.

In the next problem, we will need the following result (Theorem 14.2 in the bandit book).

Theorem 1 (Bretagnolle–Huber inequality)**.** *Let P and Q be probability measures on the same measurable space $(\Omega,\mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp\left(-\mathrm{KL}(P,Q)\right) , \tag{5}$$

where $A^c = \Omega \setminus A$ is the complement of A.

To show that there is no method that simultaneously over all P is optimal, we will study

$$R(P,\mathcal{A}) = L(P,\mathcal{A}) - \inf_{\mathcal{A}'} L(P,\mathcal{A}') ,$$

which is the excess loss suffered while using $\mathcal{A}$ on distribution P instead of using the algorithm is achieves the best loss on P. Clearly, $R(P,\mathcal{A}) \geq 0$ for any P and $\mathcal{A}$ (why?). Sometimes, $R(P,\mathcal{A})$ is also called the *regret* of $\mathcal{A}$ on P (which is a shorthand for how much one regrets using $\mathcal{A}$ instead of the optimal P-specific method). The "overall optimal" method $\mathcal{A}_{\text{fictional}}$ would achieve zero regret over all $P \in \mathcal{M}_1(\mathbb{R})$. A short way of expressing this is to write

$$\sup_{P\in\mathcal{M}_1(\mathbb{R})} R(P,\mathcal{A}_{\text{fictional}}) = 0$$

(why?). Conversely, if

$$R^* = \inf_{\mathcal{A}} \sup_{P\in\mathcal{M}_1(\mathbb{R})} R(P,\mathcal{A}) > 0 ,$$

then there is no "overall optimal" method (why?). Our angle to show that there is no optimal method will be to show this latter lower bound on R^*. The quantity R^* is the minimax regret, or minimax excess loss.

We will also need $R^*(\sigma^2)$, which is the minimax regret when the set of distributions is restricted to those whose variance does not exceed σ^2. Formally, this could be defined as follows:

$$R^*(\sigma^2) = \inf_{\mathcal{A}} \sup_{P\in\mathcal{M}_1(\mathbb{R}):\mathrm{Var}(P)\leq\sigma^2} R(P,\mathcal{A}) .$$

Question 2. Solve the following problems.

(a) Show that for any $\mathcal{A}$ and distribution P with finite variance σ^2 and mean μ,

$$L(P,\mathcal{A}) = \int (\mathcal{A}(y)-\mu)^2 P(dy) + \sigma^2 .$$

2 points

[6] For $u \geq 0$, $\mathcal{M}_u(\mathbb{R})$ denotes the set of distributions P over the reals such that $\int dP = u$.

(b) Consider the "identity algorithm" $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ that uses $\mathcal{A}(y) = y$. Evaluate $L(P, \mathcal{A})$, where P is any distribution with finite variance σ^2.

Hint: Use algebra and elementary properties of integrals/expectations.

2 points

(c) Evaluate $L^*(P) = \inf_{\mathcal{A}} L(P, \mathcal{A})$ where P has finite variance σ^2. Is there a method $\mathcal{A}$ that achieves $L^*(P)$ (i.e., $L^*(P) = L(P, \mathcal{A})$ and what is it?).

2 points

(d) Now fix $\mu_1 < \mu_2$ and define $R^*(\mu_1, \mu_2) = \inf_{\mathcal{A}} \max(R(P_1, \mathcal{A}), R(P_2, \mathcal{A}))$, where recall that $P_i = \mathcal{N}(\mu_i, \sigma^2)$. (As usual, in the argument of inf we only consider $\mathbb{R} \to \mathbb{R}$ measurable maps.) Show that $R^* \geq R^*(\mu_1, \mu_2)$. (Hence, if $R^*(\mu_1, \mu_2) > 0$ then it follows that $R^* > 0$.)

2 points

(e) What algorithm would you use if you knew that the algorithm will only be evaluated on either P_1 or P_2, that is, the algorithm's quality is judged based on $\max(R(P_1, \mathcal{A}), R(P_2, \mathcal{A}))$? Is there an algorithm that does better than the "identity" algorithm?

2 points

(f) Show that the following hold: *(i)* for any (measurable) $\mathcal{A} : \mathbb{R} \to \mathbb{R}$, $\mu_1 \neq \mu_2$,

$$\max(R(P_1, \mathcal{A}), R(P_2, \mathcal{A})) \geq \frac{\Delta^2}{16} \exp\left(-\tfrac{\Delta^2}{2\sigma^2}\right) ,$$

where $\Delta = |\mu_1 - \mu_2|$ and then conclude that *(ii)*

$$R^*(\mu_1, \mu_2) \geq \frac{\Delta^2}{16} \exp\left(-\tfrac{\Delta^2}{2\sigma^2}\right) .$$

15 points

Hint: Use Theorem 1. While the previous part is not immediately applicable, it may give you some ideas of how to lower bound the maximum loss by the sum (or maximum) of "failure probabilities" (the logic is that if the data is from P_1, the loss is large when the prediction is in some well-chosen set A, while if the data is from P_2, the loss is large when the prediction is in the complement of A). Also, recall that for a, b reals, $\max(a, b) \geq (a + b)/2$.

(g) Show that

$$R^*(\sigma^2) \geq \frac{\sigma^2}{8} \exp(-1) . \tag{6}$$

Explain the implications of this result. What did we learn from all this? Be concise, i.e., a couple of sentences should suffice. Do not write more than 4 short paragraphs. In particular, what did we learn about R^*?

2 points

(h) Show that

$$R^*(\sigma^2) \leq \sigma^2 \,.$$

2 points

(i) **Practice**: Use the Neyman-Pearson lemma instead of Theorem 1 to prove an analogue of Eq. (6). Compare with Eq. (6); the numerical constant in the lower bound should be at least three times larger than there.

0 points

(j) **Practice**: How would the results look like if the quadratic loss was replaced with its scaled version: $\ell_c(u,v) = c(u-v)^2$, $u, v \in \mathbb{R}$, $c > 0$.

0 points

(k) **Practice**: Recalculate the minimax lower bound from Part (g) for the absolute value loss: $\ell(u,v) = |u-v|$, $u, v \in \mathbb{R}$. Also show that $R^*(\sigma^2) \leq \sigma$. What differences do you observe compared to the squared loss.

Hint: Switch to the Laplace distributions for tractability. The Neyman-Pearson lemma may also be simpler to use than Theorem 1.

0 points

Total: **29 points**

Total for all questions: 61. Of this, 11 are bonus marks. Your assignment will be marked out of 50.