CMPUT 654: Theoretical Foundations of Machine Learning, Fall 2023
Homework #2

Instructions

Submissions You need to submit a single PDF file, named `p02_<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: October 8 at 11:55 pm

Problems

As usual, we assume all functions are measurable as needed. The topic of the first group of questions is basic measure concentration.

Question 1.

(a) Calculate the moment-generating function of Gaussian random variable. Show your work.

2 points

(b) Prove Tong's version of Markov's inequality. That is, for any random variable X, any function $h : \mathbb{R} \to [0, \infty)$ and any $t > 0$, $\mathbb{P}(h(X) \geq t) \leq \mathbb{E}[h(X)]/t$.

2 points

(c) Let X be a random variable on $\mathbb{R}$ with density with respect to the Lebesgue measure of $p(x) = |x| \exp(-x^2/2)/2$. Show that $\mathbb{P}(|X| \geq \varepsilon) = \exp(-\varepsilon^2/2)$.

2 points

(d) Let X as in the previous part. Show that X is **not** $\sqrt{(2-\varepsilon)}$-subgaussian for any $\varepsilon > 0$.

2 points

(e) Let X_i be σ_i-subgaussian for $i \in \{1, 2\}$ with $\sigma_i \geq 0$. Prove that $X_1 + X_2$ is $(\sigma_1 + \sigma_2)$-subgaussian. Do *not* assume independence of X_1 and X_2.

2 points

Total: **10 points**

The next questions explores the union bound.

Question 2.

Fix $0 \leq \delta \leq 1$. Show that there exist a finite set W and collection of random variables $(X(w))_{w \in W}$ such that, for any $w \in W$, with probability at least $1 - \delta$, $X(w) \geq 0$ yet it is not true that with probability $1 - \delta$, $\min_{w \in W} X(w) \geq 0$.

Total: **2 points**

The next questions explores PAC learning.

Question 3.

(a) Show that the ERM for the AND class can be efficiently computed.

5 points

(b) Show that the ERM for "Decision Lists" (Example 3.2 in the book) can be efficiently computed.

10 points

(c) Show that the "Decision Lists" class is PAC learnable.

5 points

Total: **20 points**

The next questions explore multiplicative Chernoff inequality (and compares it to the additive one).

In this question $X_1, \ldots, X_n$ are i.i.d. random variables, $X_1 \in [0, 1]$, $\mu = \mathbb{E}[X_1]$ and $\bar{X}_n = (X_1 + \cdots + X_n)/n$. We also fix $0 < \delta < 1$ and let $L = \log(1/\delta)/n$. We use the standard notation $(x)_+ = \max(x, 0)$ $(x \in \mathbb{R})$. In words, $(x)_+$ is called the "positive part" of x.

Question 4.

(a) Show that for any $b, c \geq 0$ and $u \in \mathbb{R}$, $u \leq c + b\sqrt{u}$ implies $u \leq c + b\sqrt{c} + b^2$.

2 points

(b) Show that for any $b, c \geq 0$ and $u \in \mathbb{R}$, $u \geq c - b\sqrt{u}$ implies $u \geq c - b\sqrt{c}$.

2 points

(c) Show that with probability at least $1 - \delta$, $\mu \leq \bar{X}_n + \sqrt{2L\bar{X}_n} + 2L$.

2 points

(d) Show that with probability at least $1-\delta$, $\mu \geq (\bar{X}_n - L/3)_+ - \sqrt{2L\,(\bar{X}_n - L/3)_+}$.

2 points

(e) Assume $\mu > 0$. Show that Hoeffding's inequality (or "additive Chernoff") implies that if $2n \geq \log(1/\delta)/(\mu\varepsilon)^2$ then with probability at least $1-\delta$, $\frac{\mu - \bar{X}_n}{\mu} \leq \varepsilon$.

2 points

(f) Assume $\mu > 0$. Show that the multiplicative Chernoff inequality implies that if $n \geq 2\log(1/\delta)/(\mu\varepsilon^2)$ then with probability at least $1-\delta$, $\frac{\mu - \bar{X}_n}{\mu} \leq \varepsilon$.

2 points

Remark Consider now the results of Parts (e) and (f). From these, we see that Hoeffding's inequality implies that to achieve $\bar{X}_n \geq \mu/2$ hold with high probability, it is sufficient to take $O(1/\mu^2)$ samples. In contrast, the multiplicative Chernoff inequality implies that $O(1/\mu)$ samples are sufficient. Yet another reason to call the multiplicative Chernoff inequality multiplicative is because it gives better results for such multiplicative (or relative) error bounds.

Total: **12 points**

For a measurable set $\mathcal{X}$, $U \subset \mathbb{R}$ measurable, let $\mathcal{M}(\mathcal{X}, U)$ be the set of measurable functions from $\mathcal{X}$ to U. Let $\mathcal{Z}$ be a set and let P be a distribution over $\mathcal{Z}$ (hence, $\mathcal{Z}$ is assumed to be equipped with an appropriate measurability structure). For $c_0, c_1 > 0$, we define

$$\mathrm{Var}_{\mathcal{Z}}(c_0, c_1, P) = \{g \in \mathcal{M}(\mathcal{Z}, \mathbb{R}) \,:\, \mathrm{Var}_P(g) \leq c_0^2 + c_1 Pg\},$$

where recall that $Pg = \int g dP$ and $\mathrm{Var}_P(g) = \int (g - Pg)^2 dP$ (which, by a slight abuse of notation, we may also write as $P\,(g - Pg)^2$).

Question 5. Solve the following problems.

(a) Show that $\mathcal{M}(\mathcal{Z}, [0,1]) \subset \mathrm{Var}_{\mathcal{Z}}(0, 1, P)$.

2 points

(b) For some set $\mathcal{X}$, let $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathbb{R})$ and assume that $\mathcal{F}$ is convex (i.e., for any $\alpha \in [0,1]$, $f, g \in \mathcal{F}$, $\alpha f + (1-\alpha)g \in \mathcal{F}$ also holds). Let $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and

$$\mathcal{G} = \{\ell_f \,:\, \ell_f : \mathcal{Z} \to \mathbb{R}, \ell_f(x, y) = (f(x) - y)^2, f \in \mathcal{F}\}\,.$$

By abusing notation, we also write for this set $\mathcal{G} = \ell_{\mathrm{sq}} \circ \mathcal{F}$. Let $P \in \mathcal{M}_1(\mathcal{Z})$ be such that for some $M > 0$ constant, for any $g \in \mathcal{G}$, $g(Z) \leq M^2$ with probability one, where $Z \sim P$. Define $g_* = \mathrm{argmin}_{g \in \mathcal{G}} Pg$ (which is assumed to exist) and

$$\tilde{\mathcal{G}} = \{g - g_* \,:\, g \in \mathcal{G}\} \quad (= \mathcal{G} - \{g_*\})\,.$$

Then, for some universal constant $c > 0$,

$$\tilde{\mathcal{G}} \subset \mathrm{Var}_{\mathcal{Z}}(0, cM^2, P).$$

10 points

(c) Fix $M > 0$. Let $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, [0, M])$, $\mathcal{Z} = \mathcal{X} \times [0, M]$, $P \in \mathcal{M}_1(\mathcal{Z})$. Let $\mathcal{G} = \ell_{\text{sq}} \circ \mathcal{F}$, $f_*(x) = \mathbb{E}[Y|X = x]$, $x \in \mathcal{X}$ and $\tilde{\mathcal{G}} = \mathcal{G} - \{\ell_{f_*}\}$. Then, for some universal constant $c > 0$, $\tilde{\mathcal{G}} \subset \text{Var}_{\mathcal{Z}}(0, cM^2, P)$.

10 points

Total: **22 points**

Total for all questions: 66. Of this, 16 are bonus marks. Your assignment will be marked out of 50.