Machine Learning B

2021-2022

Home Assignment 4

Yevgeny Seldin **Sadegh Talebi**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **21 December 2021, 22:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your full source code in the PDF file, only selected lines if you are asked to do so.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.
- Do NOT zip the PDF file, since zipped files cannot be opened in speed grader. Zipped PDF submissions will not be graded.
- Your PDF report should be self-sufficient. It should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted, please use the provided latex template to write your report.

1 Numerical comparison of kl inequality with its relaxations and with Hoeffding's inequality (25 points)

Let $X_1, \ldots, X_n$ be a sample of n independent Bernoulli random variables with bias $p = \mathbb{P}(X = 1)$. Let $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical average. In this question we make a numerical comparison of the relative power of various bounds on p we have studied. Specifically, we consider the following bounds:

A. **Hoeffding's inequality**: by Hoeffding's inequality, with probability greater than $1 - \delta$:

$$p \leq \hat{p}_n + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

(This is the bound we would like you to plot.)

B. kl **inequality**: you should take the bound in the form "with probability greater than $1 - \delta$, $p \leq \ldots$". In the lecture notes we provide a bound on $\mathrm{kl}(\hat{p}_n \| p)$. The upper bound on p follows by taking the "upper inverse" of kl. Namely, we define $\mathrm{kl}^{-1^+}(\hat{p}_n, z) = \max\{p : \mathrm{kl}(\hat{p}_n \| p) \leq z\}$. We have that if $\mathrm{kl}(\hat{p}_n \| p) \leq z$ then $p \leq \mathrm{kl}^{-1^+}(\hat{p}_n, z)$.

Some guidance: There is no closed-form expression for computing $\mathrm{kl}^{-1^+}(\hat{p}_n, z)$, so it has to be computed numerically. The function $\mathrm{kl}(\hat{p}_n \| p)$ is convex in p (you are very welcome to pick some value of $\hat{p}_n$ and plot $\mathrm{kl}(\hat{p}_n \| p)$ as a function of p to get a bit of intuition about its shape). We also have $\mathrm{kl}(\hat{p}_n \| \hat{p}_n) = 0$, which is the minimum, and $\mathrm{kl}(\hat{p}_n \| p)$ monotonically increases in p, as p grows from $\hat{p}_n$ up to 1. So you need to find the point $p \in [\hat{p}_n, 1]$ at which the value of $\mathrm{kl}(\hat{p}_n \| p)$ grows above z. You could do it inefficiently by linear search or, exponentially more efficiently, by binary search.

A technicality: In the computation of kl we define $0 \ln 0 = 0$. In the numerical calculations $0 \ln 0$ is undefined. So you should treat $0 \ln 0$ operations separately, either by directly assigning the zero value or by replacing them with $0 \ln 1 = 0$.

C. **Pinsker's relaxation of the** kl **inequality**: the bound on p that follows from kl-inequality by Pinsker's inequality.

D. **Refined Pinsker's relaxation of the** kl **inequality**: the bound on p that follows from kl-inequality by refined Pinsker's inequality.

In this task you should do the following:

1. Write down explicitly the four bounds on p you are evaluating.

2. Plot the four bounds on p as a function of $\hat{p}_n$ for $\hat{p}_n \in [0,1]$, $n = 1000$, and $\delta = 0.01$. You should plot all four bounds in one figure, so that you can directly compare them. Clip all bounds at 1, because otherwise they are anyway meaningless and will only destroy the scale of the figure.

3. Generate a "zoom in" plot for $\hat{p}_n \in [0, 0.1]$.

4. Compare Hoeffding's lower bound on p with kl lower bound on p for the same values of $\hat{p}_n, n, \delta$ in a separate figure (no need to consider the relaxations of kl). The kl lower bound follows from the "lower inverse" of kl defined as $\mathrm{kl}^{-1^-}(\hat{p}_n, z) = \min\{p : \mathrm{kl}(\hat{p}_n \| p) \leq z\}$.

   Some guidance: For computing the "lower inverse" $\mathrm{kl}^{-1^-}(\hat{p}_n, \varepsilon) = \min\{p : \mathrm{kl}(\hat{p}_n \| p) \leq \varepsilon\}$ you can either adapt the function for computing the "upper inverse" you wrote earlier (and we leave it to you to think how to do this), or implement a dedicated function for computing the "lower inverse". Direct computation of the "lower inverse" works the same way as the computation of the "upper inverse". The function $\mathrm{kl}(\hat{p}_n \| p)$ is convex in p with minimum $\mathrm{kl}(\hat{p}_n \| \hat{p}_n) = 0$ achieved at $p = \hat{p}_n$, and monotonically decreasing in p, as p increases from 0 to $\hat{p}_n$. So you need to find the point $p \in [0, \hat{p}_n]$ at which the value of $\mathrm{kl}(\hat{p}_n \| p)$ decreases below z. You can do it by linear search or, more efficiently, by binary search. And, as mentioned earlier, you can save all the code writing if you find a smart way to reuse the function for computing the "upper inverse" to compute the "lower inverse". Whatever way you chose you should explain in your main `.pdf` submission file how you computed the upper and the lower bound.

5. Write down your conclusions from the experiment. For what values of $\hat{p}_n$ which bounds are tighter and is the difference significant?

6. [Optional, not for submission.] You are welcome to experiment with other values of n and δ.

2 Occam's razor with kl inequality (15 points)

Prove the following theorem.

Theorem 1. Let S be an i.i.d. sample of n points, let ℓ be the zero-one loss, let $\mathcal{H}$ be countable, and let $\pi(h)$ be such that it is independent of the sample S and

satisfies $\pi(h) \geq 0$ for all h and $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$. Let $\delta \in (0,1)$. Then

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \mathrm{kl}(\hat{L}(h,S) \| L(h)) \geq \frac{\ln \frac{n+1}{\pi(h)\delta}}{n} \right) \leq \delta.$$

Briefly emphasize in your proof where you are using the assumption that $\pi(h)$ is independent of S and why it is necessary.

3 Refined Pinsker's Lower Bound (10 points)

Prove that if $\mathrm{kl}(p\|q) \leq \varepsilon$ then $q \geq p - \sqrt{2p\varepsilon}$. You are allowed to use Refined Pinsker's inequality in Lemma 2.18 in Yevgeny's lecture notes.

4 Bayesian Inference (50 points)

4.1 Here, we revisit the problem of Bayesian inference of Gaussian random variables studied in the class. Consider we observe a collection $\mathbf{X} = (X_1, \ldots, X_n)$ of n random variables, each with an unknown common mean θ. Assume that given the value of the mean, the X_i's are independent, and each X_i is sampled from a Gaussian with a *known* variance σ_i^2. We wish to infer the value of θ.

Characterize the corresponding posterior distribution.

4.2 Consider the Linear LMS estimation of the bias of a coin, as studied in the class, where the probability of heads of the coin is modeled as a random variable Θ. Assume the prior distribution of Θ is uniform over the interval $[0,1]$. We toss the coin n times, independently, and obtain a random number of heads, denoted by X. Thus, if Θ is equal to θ, the random variable X has a binomial distribution with parameters n and θ.

(i) Show that when $X = k$, the posterior density is a beta distribution with parameters $\alpha = k+1$ and $\beta = n-k+1$.

(ii) Show that the variance of X equals $n(n+2)/12$.

(iii) Show that $\mathrm{cov}(X, \Theta) = n/12$.

(iv) Using Parts (ii) and (iii), derive the linear LMS estimator for X.

(v) For a beta distribution with parameters $\alpha = k+1$ and $\beta = n-k+1$, for any m,

$$\mathbb{E}\big[\Theta^m \big| X = k\big] = \frac{(k+1)(k+2)\cdots(k+m)}{(n+2)(n+3)\cdots(n+m+1)}.$$

Use this relation to show that for any estimate $\hat{\theta}$, the corresponding conditional mean squared error $\mathbb{E}\big[(\hat{\theta}-\Theta)^2\big|X=k\big]$ satisfies:

$$\mathbb{E}\big[(\hat{\theta}-\Theta)^2\big|X=k\big]=\hat{\theta}^2-2\hat{\theta}\frac{k+1}{n+2}+\frac{(k+1)(k+2)}{(n+2)(n+3)}$$

(vi) Using the latter formula, derive the error $\mathbb{E}\big[(\hat{\theta}-\Theta)^2\big|X=k\big]$ of the estimator derived in Part (iv), and plot it as a function of k for $n=20$.

5 [Optional, not for submission] Asymmetry of the kl divergence

Prove that kl is asymmetric in its arguments by providing an example of p and q for which $\mathrm{kl}(p\|q)\neq\mathrm{kl}(q\|p)$.

6 [Optional, not for submission] Fast convergence rates when the empirical loss is zero

In this question we provide a simple and intuitive explanation on why faster convergence rates are possible when the empirical loss is zero. The kl inequality provides a continuous interpolation between fast convergence rates (of order $\frac{1}{n}$) when the empirical loss is zero and slow convergence rates (of order $\sqrt{\frac{1}{n}}$) when it is close to 1/2.

1. Let $X_1,\ldots,X_n$ be a sample of n independent Bernoulli random variables with bias $p=\mathbb{P}(X=1)$. Let $\hat{p}_n=\frac{1}{n}\sum_{i=1}^n X_i$ be the empirical average. Recall that by Hoeffding's inequality

$$\mathbb{P}\left(p\geq\hat{p}_n+\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right)\leq\delta.$$

Prove that if $p\geq\varepsilon$, then $\mathbb{P}(\hat{p}_n=0)\leq e^{-n\varepsilon}$. In other words, if $p\geq\frac{\ln\frac{1}{\delta}}{n}$ then $\mathbb{P}(\hat{p}_n=0)\leq\delta$.

The result means that the probability of observing a sample with $\hat{p}_n=0$ that is non-representative (diverges from the true mean p) by more than $\frac{\ln\frac{1}{\delta}}{n}$ is bounded by δ. (Note that for a sample with $\hat{p}_n=\frac{1}{2}$ we can only

bound divergence from the true mean by $\sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$ with the same probability δ.)

Hint for the proof: what is the probability that we make n independent flips of a coin with bias p and get all zeros? The inequality $1 + x \leq e^x$ is helpful for the proof.

2. Let S be an i.i.d. sample of n points. Let $\mathcal{H}$ be countable and let $\pi(h)$ be such that it is independent of S and $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$. Assume that for all $h \in \mathcal{H}$ we have $L(h) \geq \frac{\ln \frac{1}{\pi(h)\delta}}{n}$. Show that

$$\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S) = 0\Big) \leq \delta.$$

The result means that the probability of observing an empirical loss $\hat{L}(h^*, S) = 0$ that is non-representative by more than $\frac{\ln \frac{1}{\pi(h^*)\delta}}{n}$ is bounded by δ.