Advanced Topics in Machine Learning 2017-2018

Yevgeny Seldin Christian Igel Tobias Sommer Thune Julian Zimmert

Home Assignment 1

Deadline: Tuesday, 12 September 2017, 23:59

The assignments must be answered individually - each student must write and submit his/her own solution. We encourage you to work on the assignments on your own, but we do not prevent you from discussing the questions in small groups. If you do so, you are requested to list the group partners in your individual submission.

Submission format: *Please, upload your answers in a single* `.pdf` *file and additional* `.zip` *file with all the code that you used to solve the assignment. (The* `.pdf` *should* ***not*** *be part of the* `.zip` *file.)*

IMPORTANT: *We are interested in how you solve the problems, not in the final answers. Please, write down the calculations and comment your solutions.*

Assignment structure: The assignment contains 4 mandatory questions and 2 optional questions. All questions are designed to help you understand the material, however, the optional questions are not for submission and you can get help with the optional questions at the TA session if needed.

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**: again, you should take the bound in a 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 $p \leq \mathrm{kl}^{-1^+}(\hat{p}_n, z)$.

We provide a MATLAB function for numerical computation of the inverse kl^{-1^+}.

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. 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 $\mathrm{kl}^{-1^-}(\hat{p}_n, z) = \min\{p : \mathrm{kl}(\hat{p}_n \| p) \leq z\}$.

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

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

We provide a MATLAB function for inverting the kl-divergence with respect to its second argument. The function computes the "upper inverse" $\mathrm{kl}^{-1^+}(\hat{p}, \varepsilon) = \max\{p : \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$. The inversion is computed by binary search. You are not obliged to use the function and can write your own if you like using any programming language you like. For the "lower inverse" $\mathrm{kl}^{-1^-}(\hat{p}, \varepsilon) = \min\{p : \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$ you can either adapt the "upper inverse" function (and we leave it to you to think how to do this) or write your own function. Whatever way you chose you should explain in your main `.pdf` submission file how you computed the upper and the lower bound. Please, attach all code that you used for solving the assignment in a separate `.zip` file.

2 Occam's razor with kl inequality (25 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 $p(h)$ be such that it is independent of the sample S and satisfies $\sum_{h\in\mathcal{H}} p(h) \leq 1$. Let $\delta \in (0,1)$. Then with probability greater than $1-\delta$, for all $h \in \mathcal{H}$ simultaneously:*

$$\mathrm{kl}(\hat{L}(h,S)\|L(h)) \leq \frac{\ln \frac{n+1}{p(h)\delta}}{n}.$$

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

3 Refined Pinsker's Lower Bound (25 points)

Prove that if $\mathrm{kl}(p\|q) \leq \varepsilon$ then $q \geq p - \sqrt{2p\varepsilon}$.

4 Convexity (25 points)

To get a better understanding of the concept of convexity, prove at least two of the following basic statements:

Theorem 2. *Composition of convex functions:*

- *If $g(x)$ is convex and $f(x)$ is convex, then their weighted sum $\alpha f(x)+\beta g(x)$ is convex for nonnegative constants α and β.*
- *If $g(x)$ is convex and $f(x)$ is convex and increasing, then the functional composition $f \circ g(x) = f(g(x))$ is convex.*
- *If $f(x)$ is convex and $g(x)$ is an affine linear function $g(x) = \langle a, x\rangle + b$, then the functional composition $f \circ g(x) = f(\langle a, x\rangle + b)$ is convex.*

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

For $p, q \in [0,1]$ let $\mathrm{kl}(p\|q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$ be the kl-divergence between two Bernoulli distributions with biases p and q. 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 non-representative sample $\hat{p}_n = 0$ in a world with $p \geq \frac{\ln \frac{1}{\delta}}{n}$ is bounded by δ. Turning it the other side around: with probability greater than $1 - \delta$ the sample is representative and we are in a world with $p < \frac{\ln \frac{1}{\delta}}{n}$. Note that the specialized guarantee on p in the case when $\hat{p}_n = 0$ is much stronger than it is in the general case.

   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 $p(h)$ be such that it is independent of S and $\sum_{h \in \mathcal{H}} p(h) \leq 1$. Assume that for all $h \in \mathcal{H}$ we have $L(h) \geq \frac{\ln \frac{1}{p(h)\delta}}{n}$. Show that

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

   Again, if we turn it the other side around, if for some h^* we observe $\hat{L}(h^*, S) = 0$ then with probability greater than $1 - \delta$ it is representative and we are in a world where $L(h^*) < \frac{\ln \frac{1}{p(h^*)\delta}}{n}$.

Good luck!
Yevgeny, Christian, Tobias & Julian