Advanced Topics in Machine Learning - Assignment 1

Christoffer Thrysøe - dfv107

December 12, 2018

1 Numerical comparison of kl inequality with its relaxations and with Hoeffding's inequality

1

A

Hoeffding's bound on p is as followed:

$$P\left\{p \leq \hat{p} + \sqrt{\frac{ln\frac{1}{\delta}}{2n}}\right\} > 1 - \delta$$

B

By theorem 2.14 from Yevgeny's lecture notes, we have the following bound on the kl-inequality:

$$P\left\{\mathrm{kl}(\hat{p}||p) \geq \epsilon\right\} \leq (n+1)e^{-n\epsilon} \tag{1}$$

Setting the right hand side of (1) equal to δ, we get the following bound on $\mathrm{kl}(\hat{p}||p)$, with probability greater than $1-\delta$:

$$\mathrm{kl}(\hat{p}||p) \leq \frac{ln\frac{n+1}{\delta}}{n} \tag{2}$$

We can derive a bound for p by taking the upper inverse of $\mathrm{kl}(\hat{p}||p)$, which is defined as:

$$\mathrm{kl}^{-1^+}(\hat{p}, z) = \max\left\{p : \mathrm{kl}(\hat{p}||p) \leq z\right\} \tag{3}$$

and if $\mathrm{kl}(\hat{p}||p) \leq z$ then $p \leq \mathrm{kl}^{-1^+}(\hat{p}, z)$. If we denote the right hand side of (2) by z and exchange the definition of $\mathrm{kl}(\hat{p}||p)$ with definition 2.11 from Yevgeny's lecture notes, we get the following bound for p:

$$P\left\{p \leq \max\left\{p : p\ \ln\frac{\hat{p}}{p} + (1-\hat{p})\ \ln\ \frac{1-\hat{p}}{p} \leq \frac{ln\frac{n+1}{\delta}}{n}\right\}\right\} > 1 - \delta$$

Note that for the plot, the above bound on p was done using the provided matlab function `ysidkl.m`, which computes the upper inverse of the kl inequality using a binary search.

C

Pinsker's relaxation of the kl inequality gives the following bound on p:

$$|p - \hat{p}| \leq \sqrt{\frac{\text{kl}(\hat{p}||p)}{2}} \leq \sqrt{\frac{ln\frac{n+1}{\delta}}{2n}}$$

$$p \leq \hat{p} + \sqrt{\frac{\text{kl}(\hat{p}||p)}{2}} \leq \hat{p} + \sqrt{\frac{ln\frac{n+1}{\delta}}{2n}}$$

Thus we have the following bound on p:

$$P\left\{p \leq \hat{p} + \sqrt{\frac{ln\frac{n+1}{\delta}}{2n}}\right\} > 1 - \delta \tag{4}$$

D

The refined Pinsker's relaxation of the kl inequality gives the following bound on p:

$$P\left\{p \leq \hat{p} + \sqrt{\frac{2\hat{p}\ ln\frac{n+1}{\delta}}{n}} + \frac{2ln\frac{n+1}{\delta}}{n}\right\} > 1 - \delta \tag{5}$$

2

Figure 1 shows the four aforementioned upper bounds on p, plotted as a function of $\hat{p}$, these bounds will be discussed in section 4 of the assignment.

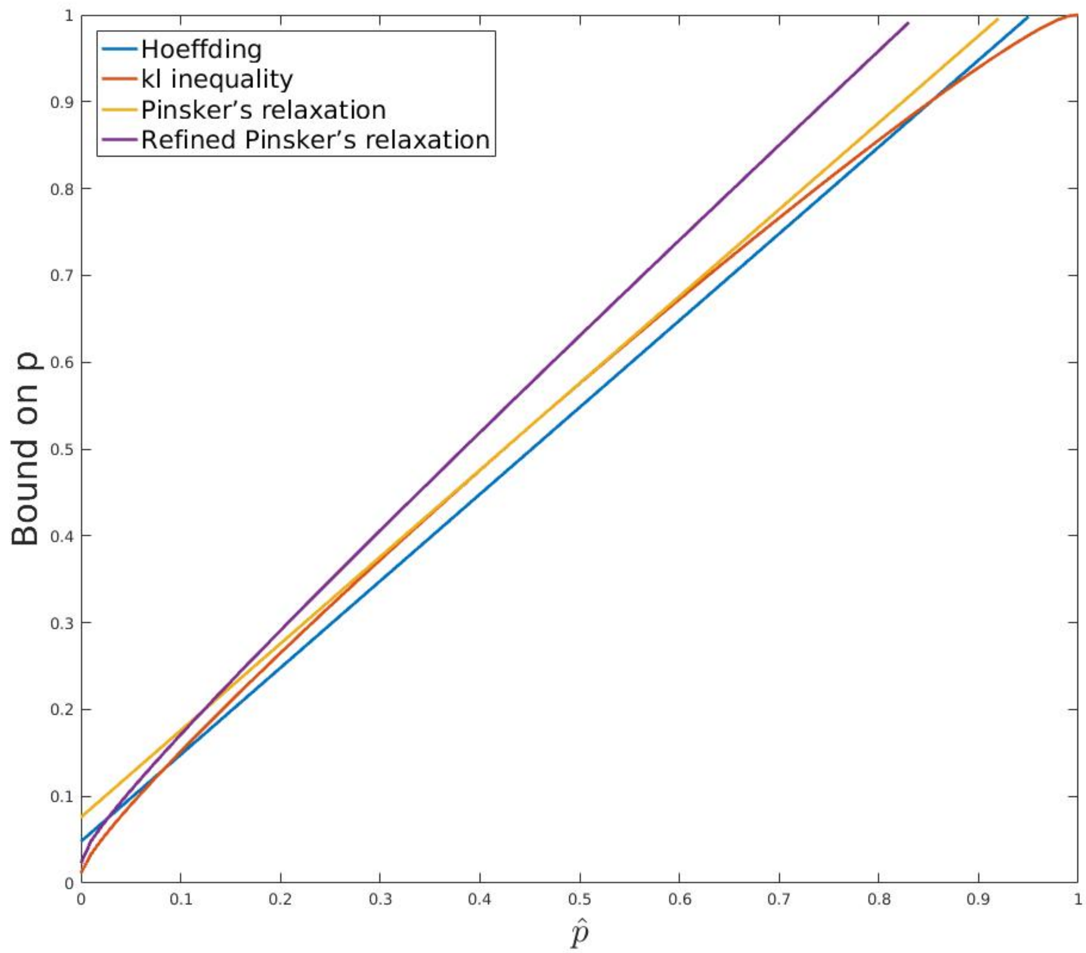

Figure 1: The upper bound on p for the four bounds, plotted as a function of $\hat{p}$

3

Hoeffding's lower bound on p is given by:

$$P\left\{p > \hat{p} - \sqrt{\frac{ln\frac{1}{\delta}}{2n}}\right\} > 1 - \delta$$

The kl inequality lower bound on p is achieved by taking the lower inverse of the kl inequality, which is defined as:

$$kl^{-1^{-}}(\hat{p}, z) = \min\{p : \mathrm{kl}(\hat{p}||p) \leq z\} \tag{6}$$

The lower bound can be stated as followed:

$$\begin{aligned} &P\left\{p > \min\{p : \mathrm{kl}(\hat{p}||p) \leq z\}\right\} > 1 - \delta \\ =&P\left\{p > \min\left\{p : p \ln\frac{\hat{p}}{p} + (1-\hat{p}) \ln \frac{1-\hat{p}}{p} \leq \frac{ln\frac{n+1}{\delta}}{n}\right\}\right\} > 1 - \delta \end{aligned}$$

The above minimization problem was solved using a binary search approach adopted from the existing provided matlab function. The program (`klmin.m`) simply evaluates if the condition $\mathrm{kl}(\hat{p}||p) \leq z$ is satisfied and if so reduces p. If the condition is not met, p is changed to increase $\mathrm{kl}(\hat{p}||p)$. Figure 2 shows the two lower bounds on p, plotted as a function of $\hat{p}$.

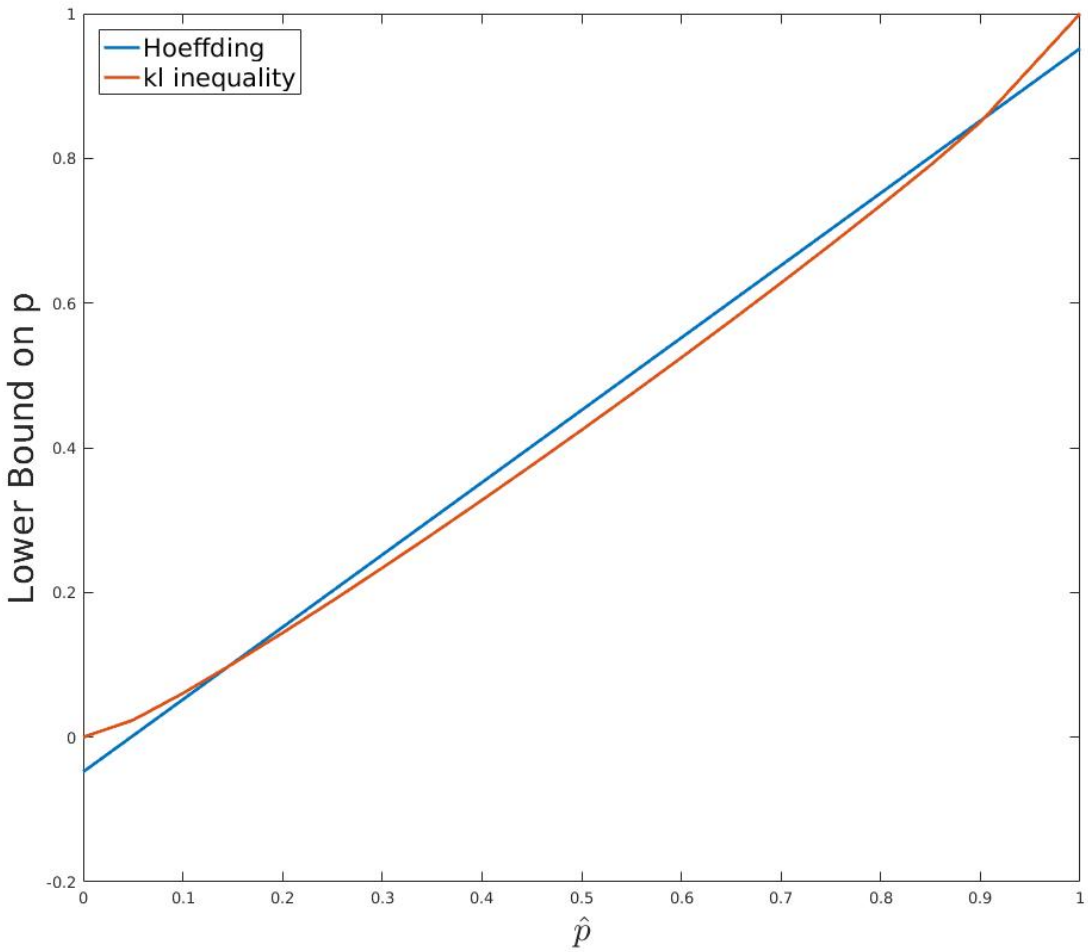

Figure 2: The Hoeffding's and kl inequality lower bound, plotted as a function of $\hat{p}$

4

Comparing Hoeffding's upper bound with Pinsker's relaxation bound in figure 1, it is clear that Hoeffding's bound is tighter due to the $n+1$ term in Pinkser's relaxation. The refined Pinkser's relaxation provides a tighter bound than Hoeffding's when $\hat{p}$ is low, providing a high confidence in the bound, which is due to the second term dominating the first, whereas for larger values of $\hat{p}$, the first term grows large, resulting in a looser bound. As expected, the kl inequality bound is tighter than it's relaxations. The kl inequality bound provides a significantly tighter bound for $\hat{p} \in [0, 0.1]$ and $\hat{p} \in [0.9, 1]$ thus a greater confidence in the bound for small and large empirical errors. Hoeffding's bound is tighter than the kl inequality bound for $\hat{p} \in [0.2, 0.8]$.

The lower bounds on p, shown in figure 2, are very similar to their correspondent upper bounds in the sense that the kl lower bound provides a tighter bound for $\hat{p} \in [0, 0.1]$ and $\hat{p} \in [0.9, 1]$ whereas Hoeffding's lower bound provide a stronger bound for $p \in [0.2, 0.8]$.

2 Occam's razor with kl inequality

We wish to prove that for all $h \in \mathcal{H}$:

$$P\left\{\text{kl}(\hat{L}(h,S)||L(h)) \leq \frac{ln\frac{n+1}{p(h)\delta}}{n}\right\} > 1-\delta \tag{7}$$

where

$$\sum_{h\in\mathcal{H}} p(h) \leq 1$$

First we use theorem 2.14 from Yevgeny's lecture notes, which states:

$$P\left\{\text{kl}(\hat{p}||p) \geq \epsilon\right\} \leq (n+1)e^{-n\epsilon} \tag{8}$$

We can use this theorem as $L(h)$ is our bias and $\hat{L}(h,s)$ is our empirical bias.
replacing ϵ with the desired bound:

$$\epsilon = \frac{ln\frac{n+1}{p(h)\delta}}{n}$$

in (8) we get:

$$P\left\{\text{kl}(\hat{L}(h,S)||L(h)) \geq \frac{ln\frac{n+1}{p(h)\delta}}{n}\right\} \leq (n+1)e^{-n\frac{ln\frac{n+1}{p(h)\delta}}{n}} \tag{9}$$

By utilizing that $e^{log(x)} = x$ we can further expand the right hand side of the last inequality of (9) to the following:

$$\begin{aligned}(n+1)e^{-n\frac{ln\frac{n+1}{p(h)\delta}}{n}} &= (n+1)\frac{1}{\left(\frac{n+1}{p(h)\delta}\right)} \\ &= p(h)\delta\end{aligned}$$

We note that we are able to multiply $p(h)$ and δ because $p(h)$ is independent of the sample S. This is necessary because otherwise the probability of the bound not holding: δ would be dependent on $p(h)$ and we could therefore not multiply the dependent events, therefore $p(h)$ has to be chosen before we observe the sample S.
To show it for all $h \in \mathcal{H}$ we take the union bound over all hypothesis:

$$\begin{aligned}P\left\{\exists h \in \mathcal{H} : \text{kl}(\hat{L}(h,S)||L(h)) \geq \frac{ln\frac{n+1}{p(h)\delta}}{n}\right\} &\leq \sum_{h\in\mathcal{H}} P\left\{\exists h \in \mathcal{H} : \text{kl}(\hat{L}(h,S)||L(h)) \geq \frac{ln\frac{n+1}{p(h)\delta}}{n}\right\} \\ &\leq \sum_{h\in\mathcal{H}} (n+1)e^{-n\frac{ln\frac{n+1}{p(h)\delta}}{n}} \\ &= \sum_{h\in\mathcal{H}} p(h)\delta \\ &\leq \delta\end{aligned}$$

Where the last inequality follows our definition of $p(h)$. Thus we have for all hypothesis in $\mathcal{H}$ that with probability greater than $1-\delta$:

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

3 Refined Pinsker's Lower Bound

We wish to prove that if $\mathrm{kl}(p||q) \leq \epsilon$ then

$$q \geq p - \sqrt{2p\epsilon} \tag{10}$$

For the proof we will prove it for $p > q$, $p < q$ and $p = q$.

p > q

We use Corollary 2.17 from Yevgeny's lecture notes for when $p > q$:

$$\begin{aligned}
\mathrm{kl}(p||q) &\geq \frac{(p-q)^2}{2\ \max\{p,q\}} + \frac{(p-q)^2}{2\ \max\{(1-p),(1-q)\}} && (11)\\
&= \frac{(p-q)^2}{2p} + \frac{(p-q)^2}{2(1-q)} && (12)\\
&\geq \frac{(p-q)^2}{2p} && (13)
\end{aligned}$$

Where (??) follows because the second term of the right hand side of (??) is positive. Since we have the condition: $\mathrm{kl}(p||q) \leq \epsilon$ we can write:

$$\epsilon \geq \mathrm{kl}(p||q) \geq \frac{(p-q)^2}{2p} \tag{14}$$

which we can derive to:

$$\begin{aligned}
\epsilon \geq \frac{(p-q)^2}{2p} &\Rightarrow && (15)\\
2p\epsilon \geq (p-q)^2 &\Rightarrow && (16)\\
\sqrt{2p\epsilon} \geq p-q &\Rightarrow && (17)\\
q \geq p - \sqrt{2p\epsilon}& && (18)
\end{aligned}$$

which proves (??) for $p > q$

q > p

For $q > p$ we have:

$$\begin{aligned}
q > p &\Rightarrow && (19)\\
q - p > 0 &\Rightarrow && (20)\\
q - p + \sqrt{2p\epsilon} > 0 &\Rightarrow && (21)\\
q \geq p - \sqrt{2p\epsilon}& && (22)
\end{aligned}$$

Where the inequality of (??) is satisfied because $\sqrt{2p\epsilon}$ is positive, which proves (??) for $q > p$.

p = q

If we define a new variable t for which $t = p = q$, it is clear to see that (??) holds:

$$t \geq t - \sqrt{2t\epsilon} \tag{23}$$

because $\sqrt{2t\epsilon} \geq 0$, the inequality will hold when $p = q$. Therefore we can write:

$$p \geq p - \sqrt{2p\epsilon} \tag{24}$$

which proves (??) for $q = p$.

4 Convexity

1

We wish to show that for convex functions $f(x)$ and $g(x)$ their weighted sum: $\alpha f(x) + \beta g(x)$ by two non-negative constants α and β is too convex.
First we note that for a function $f(x)$ to be convex it must satisfy the following:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \qquad \text{for } \lambda \in [0,1] \tag{25}$$

Therefore by convexity of f and g we have:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \tag{26}$$
$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2) \tag{27}$$

Combining (11) and (12) and multiplying with the non-negative constants α and β we get the following:

$$\alpha f(\lambda x_1 + (1-\lambda)x_2) + \beta g(\lambda x_1 + (1-\lambda)x_2) \tag{28}$$
$$\leq \alpha(\lambda f(x_1) + (1-\lambda)f(x_2)) + \beta(\lambda g(x_1) + (1-\lambda)g(x_2)) \tag{29}$$
$$= \lambda(\alpha f(x_1) + \beta g(x_1)) + (1-\lambda)(\alpha f(x_2) + \beta g(x_2)) \tag{30}$$

where (15) satisfies (10) and therefore proofs that the weighted sum of two convex functions is convex.

2

We wish to show that if $g(x)$ is convex and $f(x)$ is convex and increasing, then the functional composition $f \circ g = f(g(x))$ is convex. First we note that since g is convex we write:

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2) \tag{31}$$

Then, we use that f is increasing:

$$\begin{aligned} f(g(\lambda x_1 + (1-\lambda)x_2)) &\leq f(\lambda g(x_1) + (1-\lambda)g(x_2)) && (32)\\ &\leq \lambda f(g(x)) + (1-\lambda)f(g(x_2)) && (33) \end{aligned}$$

where the last line follows that f is convex. (18) satisfies the rule of convexity and therefore the composition of f and g is convex.

3

Given that $f(x)$ is a convex function and $g(x)$ is an affine linear function: $g(x) = ax + b$, we wish to show that the composition $f \circ g(x) = f(ax + b)$ is convex. To show this we must prove the following:

$$(f \circ g)(\lambda x_1 + (1-\lambda)x_2) \leq \lambda(f \circ g)(x_1) + (1-\lambda)(f \circ g)(x_2) \tag{34}$$

$$\begin{aligned} (f \circ g)(\lambda x_1 + (1-\lambda)x_2) &= f(a(\lambda x_1 + (1-\lambda)x_2) + \lambda b + (1-\lambda)b) && (35)\\ &= f(\lambda(ax_1 + b) + (1-\lambda)(ax_2 + b)) && (36)\\ &\leq \lambda f(ax_1 + b) + (1-\lambda)f(ax_2 + b) && (37)\\ &= \lambda(f \circ g)(x_1) + (1-\lambda)(f \circ g)(x_2) && (38) \end{aligned}$$

Where (22) follows from f being convex. (23) gives the desired result and the proof is therefore complete.