Machine Learning 3: KL divergence and lower bounds for deviations, PAC learning in the realizable case

Master 2 Computer Science

Aurélien Garivier

2019-2020

Table of contents

1. Kullback-Leibler divergence
2. PAC learning

Kullback-Leibler divergence

Kullback-Leibler divergence

Definition

Let P and Q be two probability distributions on a measurable set Ω. The Kullback-Leibler divergence from Q to P is defined as follows:

- if P is not absolutely continuous with respect to Q, then $\mathrm{KL}(P,Q) = +\infty$;
- otherwise, let $\frac{dP}{dQ}$ be the Radon-Nikodym derivative of P with respect to Q. Then

$$\mathrm{KL}(P,Q) = \int_\Omega \log\frac{dP}{dQ}\,dP = \int_\Omega \frac{dP}{dQ}\log\frac{dP}{dQ}\,dQ\,.$$

Property: $0 \leq \mathrm{KL}(P,Q) \leq +\infty$, $\mathrm{KL}(P,Q) = 0$ iff $P = Q$.

If $P \ll Q$ and $f = \frac{dP}{dQ}$, $\int_\Omega f\log(f)\,dQ = \int_\Omega \left[f\log(f)\right]_+ dQ - \int_\Omega \left[f\log(f)\right]_- dQ$, the later is finite since $\left[f\log(f)\right]_- \leq 1/e$.

Examples:

$\mathrm{KL}\left(\mathcal{B}(p),\mathcal{B}(q)\right) = \mathrm{kl}(p,q)$, $\mathrm{KL}\left(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\right) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}$.

Properties

Tensorization of entropy:

If $P = P_1 \otimes P_2$ and $Q = Q_1 \otimes Q_2$, then

$$\mathrm{KL}(P, Q) = \mathrm{KL}(P_1, Q_1) + \mathrm{KL}(P_2, Q_2)\,.$$

Contraction of entropy data-processing inequality:

Let $(\Omega, \mathcal{A})$ be a measurable space, and let P and Q be two probability measures on $(\Omega, \mathcal{A})$. Let $X : \Omega \to (\mathcal{X}, \mathcal{B})$ be a random variable, and let P^X (resp. Q^X) be the push-forward measures, ie the laws of X wrt P (resp. Q). Then

$$\mathrm{KL}\left(P^X, Q^X\right) \leq \mathrm{KL}(P, Q)\,.$$

Pinsker's inequality:

Let $P, Q \in \mathfrak{M}_1(\Omega, \mathcal{A})$. Then

$$\|P - Q\|_{TV} \stackrel{\text{def}}{=} \sup_{A \in \mathcal{A}} |P(A) - Q(A)| \leq \sqrt{\frac{\mathrm{KL}(P, Q)}{2}}\,.$$

Proof: contraction

Contraction: if $\mathsf{KL}(P,Q)=+\infty$, the result is obvious. Otherwise, $P \ll Q$ and there exists $\frac{dP}{dQ}:\Omega\to\mathbb{R}$ such that for all measurable $f:\Omega\to\mathbb{R}$, $\int_\Omega f\,dP=\int_\Omega f\,\frac{dP}{dQ}\,dQ$.

- We first prove that $P^X \ll Q^X$ and, if $\gamma(x):=\mathbb{E}_Q\left[\frac{dP}{dQ}\middle|X=x\right]$ is the Q-a.s. unique function such that $\mathbb{E}_Q\left[\frac{dP}{dQ}\middle|X\right]=\gamma(X)$, then $\gamma=\frac{dP^X}{dQ^X}$. Indeed, for all $B\in\mathcal{B}$,

$$\begin{aligned}
P^X(B)=P(X\in B)&=\int_{X\in B}\frac{dP}{dQ}dQ=\mathbb{E}_Q\left[\frac{dP}{dQ}\mathbb{1}\{X\in B\}\right]\\
&=\mathbb{E}_Q\left[\mathbb{E}_Q\left[\frac{dP}{dQ}\mathbb{1}\{X\in B\}\middle|X\right]\right]=\mathbb{E}_Q\left[\mathbb{1}\{X\in B\}\mathbb{E}_Q\left[\frac{dP}{dQ}\middle|X\right]\right]\\
&=\mathbb{E}_Q\left[\mathbb{1}\{X\in B\}\gamma(X)\right]=\int_{X\in B}\gamma(X)dQ=\int_B\gamma dQ^X
\end{aligned}$$

  and hence $P^X\ll Q^X$ and $\frac{dP^X}{dQ^X}=\gamma$.
- Now,

$$\begin{aligned}
\mathsf{KL}\left(P^X,Q^X\right)&=\int_{\mathcal{X}}\gamma\log\gamma\ dQ^X=\int_\Omega\gamma(X)\log\gamma(X)\,dQ\\
&=\mathbb{E}_Q\left[\phi\left(E_Q\left[\frac{dP}{dQ}\middle|X\right]\right)\right]\quad\text{where } \phi:=x\mapsto x\log(x)\text{ is convex}\\
&\le\mathbb{E}_Q\left[\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\middle|X\right]\right]\qquad\text{by (conditional) Jensen's inequality}\\
&=\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]=\mathsf{KL}(P,Q)\,.
\end{aligned}$$

Proof: Pinsker

Let $A \in \mathcal{A}$, $p = P(A)$ and $q = Q(A)$. By contraction,

$$\mathrm{KL}(P, Q) \geq \mathrm{KL}(P^{\mathbb{1}_A}, Q^{\mathbb{1}_A}) = \mathrm{KL}\left(\mathcal{B}\big(P(A)\big), \mathcal{B}\big(Q(A)\big)\right) = \mathrm{kl}\left(P(A), Q(A)\right) \geq 2\big(P(A) - Q(A)\big)^2 .$$

Application: Lower bound
"Chernoff's bound is asymptotically almost tight"

Let $\mu \in (0,1)$. $X_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{B}(\mu)$, and let $x \in (\mu, 1]$. Then

$$\liminf_n \frac{1}{n} \log \mathbb{P}(\bar{Y}_n > x) \geq -\operatorname{kl}(x, \mu) \ .$$

Proof: Let $\epsilon > 0$ and on the same probability space let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{B}(x+\epsilon)$ and $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{B}(\mu)$. Then

$$\begin{aligned} n \operatorname{kl}(x+\epsilon, \mu) &= \operatorname{KL}\left(P^{\mathbf{X}}, P^{\mathbf{Y}}\right) \qquad \text{by tensorization} \\ &\geq \operatorname{KL}\left(P^{\mathbb{1}\{\bar{X}_n \geq x\}}, P^{\mathbb{1}\{\bar{Y}_n \geq x\}}\right) \qquad \text{by contraction} \\ &= \operatorname{kl}\left(\mathbb{P}(\bar{X}_n \geq x), \mathbb{P}(\bar{Y}_n \geq x)\right) \\ &\geq \mathbb{P}(\bar{X}_n \geq x) \log \frac{1}{\mathbb{P}(\bar{Y}_n \geq x)} - \log(2) \end{aligned}$$

since $\operatorname{kl}(p,q) = -h(p) + p \log \frac{1}{q} + (1-p)\log\frac{1}{1-q}$. Hence, by Hoeffding's inequality,

$$\liminf_m \frac{1}{n} \log \mathbb{P}(\bar{Y}_n > x) \geq \liminf_n \frac{-n \operatorname{kl}(x+\epsilon, \mu) + \log(2)}{n\big(1 - \exp(-2n\epsilon^2)\big)} = -\operatorname{kl}(x+\epsilon, \mu)$$

for all $\epsilon > 0$, and we conclude by the continuity of $\operatorname{kl}(\cdot, \mu)$.
Note that one can also derive non-asymptotic lower bounds.

PAC learning

Learning framework

- Underlying distribution D on $\mathcal{X} \times \mathcal{Y}$.
- Sample $S \overset{iid}{\sim} D$ (otherwise: transductive learning).
- $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$ hypothesis class.
- loss function $l(y, y')$ (regression, classification)
- generalization error (loss) $L_D(h)$
- training error $L_S(h)$
- Realizable assumption: there exists h^* such that $L_S(h^*) = 0$.
- Antonym: *agnostic* learning.

Empirical risk minimization with inductive bias

Definition

Any learning algorithm $\hat{h}_m$ of the form

$$ERM_{\mathcal{H}}(S) \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

is called a *empirical risk minimizer*.

Risk of overfitting

PAC learnability: "probably approximately correct"

Definition

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_m$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D_X on $\mathcal{X}$ and for every labelling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, D_X, f$ then when $S = \big((X_1, f(X_1)), \ldots, (X_m, f(X_m))\big)$ with $(X_i)_{1 \le i \le m} \overset{iid}{\sim} D_X$,

$$\mathbb{P}\Big(L_{(D_X,f)}\big(\hat{h}_m\big) \ge \epsilon\Big) \le \delta$$

for all $m \ge m_{\mathcal{H}}(\epsilon, \delta)$.
The smallest possible function $m_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

Remark: Valiant's PAC requires also sample complexity and running time polynomial in $1/\epsilon$ and $1/\delta$.

Examples

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbb{1}\{x \leq a\}$ is PAC-learnable with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{log \frac{2}{\delta}}{\epsilon} \right\rceil .$$
Proof: let a^* be such that $L_D(h_{a^*}) = 0$ and let $a_0 = \inf\{a : D_X([a, a^*]) \leq \epsilon\}$ and $a_1 = \sup\{a : D_X([a^*, a]) \leq \epsilon\}$.
An ERM is $\hat{h}_S(x) = \mathbb{1}_{x \leq T}$ where $T \in [B_0, B_1]$, with $B_0 = \max\{x : (x, 1) \in S\}$ and $B_1 = \min\{x : (x, 0) \in S\}$. Then
$P(L(\hat{h}_S) \geq \epsilon) \leq \mathbb{P}(B_0 < a_0) + \mathbb{P}(B_1 > a_1)$. As $D_X(a_0, a^*) \geq \epsilon$,
$\mathbb{P}(B_0 < a_0) \leq (1 - D_X([a_0, a^*]))^m \leq \exp(-m\epsilon)$.
- Exercise: Learning axis-aligned rectangles: given real numbers $a_1 \leq b_1$ and $a_2 \leq b_2$, let
$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2\,; \\ 0 & \text{otherwise}\,. \end{cases}$$
Let $\mathcal{H}^2_{\mathrm{rec}} = \{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1 \text{ and } a_2 \leq b_2\}$. Show that $\mathcal{H}^2_{\mathrm{rec}}$ is PAC-learnable, with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{4 log \frac{4}{\delta}}{\epsilon} \right\rceil .$$

Finite hypothese classes are PAC-learnable

The sample complexity of finite hypothese classes in the realizable case is smaller than $m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$:

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\epsilon, \delta \in (0, 1)$ and let m be an integer that satisfies

$$m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon} \ .$$

Then, for any labeling function f and for any distribution D_X on $\mathcal{X}$, under the realizability assumption, with probability at least $1 - \delta$ over the choice of iid sample S of size m, any ERM hypothesis $\hat{h}_m$ is such that

$$L_{(D_X, f)}(\hat{h}_m) \leq \epsilon \ .$$

Proof

The realizability assumption implies that an ERM $\hat{h}_S$ has empirical risk $L_S(\hat{h}_S) = 0$. Hence,

$$
\begin{aligned}
\mathbb{P}\left(L(\hat{h}_S) \geq \epsilon\right) &= D_X^{\otimes m}\Big(\big\{S \in \mathcal{X}^m : \exists h \in \mathcal{H}, L_S(h) = 0 \text{ and } L_D(h) \geq \epsilon\big\}\Big) \\
&= D_X^{\otimes m}\left(\bigcup_{h : L_D(h) \geq \epsilon} S_h\right) \quad \text{where } S_h = \big\{S \in \mathcal{X}^m : L_s(h) = 0\big\} \\
&\leq \sum_{h : L_D(h) \geq \epsilon} D_X^{\otimes m}(S_h) \\
&= \sum_{h : L_D(h) \geq \epsilon} \prod_{i=1}^{m} \underbrace{D_X\big(\{x \in \mathcal{X} : h(x) = f(x)\}\big)}_{= 1 - L_D(h) \leq 1 - \epsilon} \\
&\leq \sum_{h : L_{(D_X, f)}(h) \geq \epsilon} \prod_{i=1}^{m} (1 - \epsilon) \leq \big|\mathcal{H}\big|(1 - \epsilon)^m \leq \big|\mathcal{H}\big| \exp(-m\epsilon)\ .
\end{aligned}
$$

This quantity is smaller than δ for $m \geq \dfrac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$.