Foundations of Machine Learning II
TP1: Entropy*

Guillaume Charpiat (Teacher) & Gaétan Marceau Caron (Scribe)

Problem 1 (Gibbs' inequality)**.** *Let p and q two probability measures over a finite alphabet $\mathcal{X}$. Prove that* $\mathrm{KL}(p \,\|\, q) \geqslant 0$

Hint: for a concave function f and a random variable X, we have the Jensen's inequality $\mathbb{E}\left[f(X)\right] \leqslant f\left(\mathbb{E}[X]\right)$. ln is a strictly concave function.

Solution: We start by reviewing some useful results from Boyd and Vandenberghe, 2004; Cover and Thomas, 2006.

Definition 1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* convex *if* $\mathbf{dom}\, f$ *is a convex set and if for all $x, y \in \mathbf{dom}\, f$ and $\theta \in [0, 1]$, we have*

$$f(\theta x + (1 - \theta)y) \leqslant \theta f(x) + (1 - \theta)f(y) \tag{1}$$

A function is strictly convex when the equality holds iif $\theta = 0$ or $\theta = 1$. Also, a function f is concave is $-f$ is convex.

Proving that a given function f is convex is usually hard with the previous definition. When f is twice-differentiable, it is easier to use the second-order condition:

Proposition 1. *Let f be a twice-differentiable function, that is, its* Hessian *or second-derivative $\nabla^2 f$ exists at each point in* $\mathbf{dom}\, f$*, which is open. Then, f is convex iif* $\mathbf{dom}\, f$ *is a convex set and $\nabla^2 f$ is positive semidefinite, i.e., for all $x \in \mathbf{dom}\, f$, we have $x^\top \nabla^2 f\, x \geqslant 0$. Moreover, the function is strictly convex if $\nabla^2 f$ is positive definite.*

Proof. Use Taylor expansion (cf. Cover and Thomas, 2006, p.26 for the univariate case) □

By applying this result, we easily find that ln is a strictly concave function.

Now, we state the Jensen's inequality required in the proof of the Gibb's inequality.

Theorem 2. *If f is a convex function and X is a random variable, then*

$$\mathbb{E}\left[f(X)\right] \geqslant f(\mathbb{E}\left[X\right]) \tag{2}$$

*https://www.lri.fr/~gcharpia/machinelearningcourse/

Proof. Induction (cf. Cover and Thomas, 2006, p.27) □

Finally, we recall the definition of the Kullback-Leibler divergence:

Definition 2. *Let p and q be two discrete probability distributions over an alphabet $\mathcal{X}$. The Kullback-Leibler divergence is defined as:*

$$\mathrm{KL}(p \,\|\, q) = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \tag{3}$$

with the convention that $0 \ln \frac{0}{0} = 0$, $0 \ln \frac{0}{b} = 0$ and $0 \ln \frac{a}{0} = \infty$
We finish with the proof given by Cover and Thomas, 2006, p.28.

Theorem 3 (Gibb's Inequality)**.** *Let p and q be two discrete probability distributions over an alphabet $\mathcal{X}$. The Kullback-Leibler divergence has the following property:*

$$\mathrm{KL}(p \,\|\, q) \geqslant 0 \tag{4}$$

with equality iif $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. Let $A = \{x : p(x) > 0\}$, then we have

$$\begin{aligned}
-\mathrm{KL}(p \,\|\, q) &= -\sum_{x \in A} p(x) \ln \frac{p(x)}{q(x)} && (5)\\
&= \sum_{x \in A} p(x) \ln \frac{q(x)}{p(x)} && (6)\\
&\leqslant \ln \sum_{x \in A} p(x) \frac{q(x)}{p(x)} && (7)\\
&= \ln \sum_{x \in A} q(x) && (8)\\
&\leqslant \ln \sum_{x \in \mathcal{X}} q(x) && (9)\\
&= \ln 1 && (10)\\
&= 0 && (11)
\end{aligned}$$

□

To apply the Jensen's inequality, let $u(x) = \frac{q(x)}{p(x)}$ and $f(x) = \ln(x)$ and express eq. (6) in terms of these new functions. Since ln is strictly concave, we have equality in eq. (7) iif $\frac{q(x)}{p(x)}$ is constant, i.e., $p(x) = cq(x)$. Then, we have equality in eq. (9) iif $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x)$. Finally, with both equalities, we have that $c = 1$.

Problem 2 (Evidence Lower bound (ELBO)). *Prove the following inequality*[1]*:*

$$-\ln p(D) \leqslant -\mathbb{E}_{\theta\sim\beta}\left[\ln p(D|\theta)\right] + KL(\beta||\alpha) \tag{12}$$

where D is a dataset, $p(D)$ is the probability of the dataset, $p(D|\theta)$ is the likelihood probability of the dataset given the model parameters θ, β is a distribution over the model parameters approximating the posterior distribution $\pi(\theta) := p(\theta|D)$ and α is the prior distribution over the model parameters.

(a) Write down the natural logarithm of the Bayes' rule in an expanded form:

$$\pi(\theta) = \frac{p(D|\theta)\alpha(\theta)}{p(D)} \tag{13}$$

Solution By applying the properties of the logarithm, we obtain:

$$0 = \ln p(D|\theta) + \ln \alpha(\theta) - \ln p(D) - \ln \pi(\theta) \tag{14}$$

(b) Introduce a new density function β and rewrite the expression in terms of expectation w.r.t. β

Solution

$$\begin{aligned}
0 &= \int \beta(\theta)\left(\ln p(D|\theta) + \ln \alpha(\theta) - \ln p(D) - \ln \pi(\theta)\right) && (15)\\
&= \int \beta(\theta)\ln p(D|\theta) + \int \beta(\theta)\ln \alpha(\theta) - \int \beta(\theta)\ln p(D) && (16)\\
&\quad + \int \beta(\theta)\ln \beta(\theta) - \int \beta(\theta)\ln \beta(\theta) - \int \beta(\theta)\ln \pi(\theta) && (17)\\
&= \int \beta(\theta)\ln p(D|\theta) - \ln p(D) + \int \beta(\theta)\ln \frac{\alpha(\theta)}{\beta(\theta)} + \int \beta(\theta)\ln \frac{\beta(\theta)}{\pi(\theta)} && (18)\\
&= \mathbb{E}_\beta\left[\ln p(D|\theta)\right] - \ln p(D) - \mathrm{KL}(\beta\,\|\,\alpha) + \mathrm{KL}(\beta\,\|\,\pi) && (19)
\end{aligned}$$

which implies

$$-\ln p(D) = -\mathbb{E}_\beta\left[\ln p(D|\theta)\right] + \mathrm{KL}(\beta\,\|\,\alpha) - \mathrm{KL}(\beta\,\|\,\pi) \tag{20}$$

(c) Use the Gibbs' inequality and write down the ELBO

Solution

$$-\ln p(D) \leqslant -\mathbb{E}_\beta\left[\ln p(D|\theta)\right] + \mathrm{KL}(\beta\,\|\,\alpha) \tag{21}$$

(d) Interpret the ELBO in a machine learning framework cf. Variational inference Bishop, 2006, p.462

[1] Further information can be found at `https://www.lri.fr/~bensadon/`

Problem 3 (Entropy)**.** *Compute the differential entropy of the following distributions:*

(a) univariate Normal distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{22}$$

Solution By taking the natural logarithm of the normal distribution, we obtain

$$\ln \mathcal{N}(x|\mu,\sigma^2) = -\ln\sqrt{2\pi}\sigma - \frac{(x-\mu)^2}{2\sigma^2} \tag{23}$$

and with the definition of the differential entropy, we have

$$\begin{aligned}
\operatorname{Ent}\left[\mathcal{N}(\cdot|\mu,\sigma^2)\right] &= -\mathbb{E}\left[\ln \mathcal{N}(x|\mu,\sigma^2)\right] && (24)\\
&= \mathbb{E}\left[\ln\sqrt{2\pi}\sigma\right] + \frac{1}{2\sigma^2}\mathbb{E}\left[(x-\mu)^2\right] && (25)\\
&= \ln\sqrt{2\pi}\sigma + \frac{1}{2} && (26)\\
&= \ln\sqrt{2\pi e\sigma^2} && (27)
\end{aligned}$$

(b) multivariate Normal distribution

$$\mathcal{N}(x|\mu,C) = \frac{1}{\sqrt{(2\pi)^d|C|}} \exp\left[-\frac{1}{2}(x-\mu)^\top C^{-1}(x-\mu)\right] \tag{28}$$

where $x,\mu \in \mathbb{R}^d$ and C is a covariance matrix (assumed to be symmetric positive-definite).

Solution By taking the natural logarithm, we obtain

$$\ln \mathcal{N}(x|\mu,C) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|C| - \frac{1}{2}(x-\mu)^T C^{-1}(x-\mu) \tag{29}$$

and with the definition of the differential entropy, we have

$$\begin{aligned}
\operatorname{Ent}\left[\mathcal{N}(\cdot|\mu,\sigma^2)\right] &= -\mathbb{E}\left[\ln \mathcal{N}(x|\mu,\sigma^2)\right] && (30)\\
&= \mathbb{E}\left[\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|C| + \frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right] && (31)\\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|C| + \frac{1}{2}\mathbb{E}\left[(x-\mu)^T C^{-1}(x-\mu)\right] && (32)\\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|C| + \frac{d}{2} && (33)\\
&= \frac{d}{2}\left(\ln(2\pi) + 1\right) + \frac{1}{2}\ln|C| && (34)\\
&= \ln\sqrt{(2\pi e)^d|C|} && (35)
\end{aligned}$$

Note that we have used to following identity for eq. (32):

$$\begin{aligned}
\mathbb{E}\left[(x-\mu)^T C^{-1}(x-\mu)\right] &= \mathbb{E}\left[\operatorname{tr}((x-\mu)^T C^{-1}(x-\mu))\right] && (36)\\
&= \mathbb{E}\left[\operatorname{tr}((C^{-1}(x-\mu)(x-\mu)^T))\right] && (37)\\
&= \operatorname{tr}(C^{-1}\mathbb{E}\left[(x-\mu)(x-\mu)^T\right]) && (38)\\
&= \operatorname{tr}(C^{-1}C) && (39)\\
&= d && (40)
\end{aligned}$$

where tr is the trace operator. In this identity, we use the fact that the trace of a scalar is equal to the scalar. Then, we use a well-known property of the trace for cycling the variables. Finally, since tr and $\mathbb{E}$ are linear operators, we can switch them.

Problem 4 (Mutual information)**.** *We are interested in computing the mutual information between a multivariate Normal distribution $\beta = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, C)$ where $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^d$ and a product of identical univariate Normal distributions $\alpha = \prod_{i=1}^d \mathcal{N}(x_i|\mu,\sigma)$.*

(a) Express the KL divergence in terms of entropy and expectation w.r.t. β

Solution

$$\begin{aligned}
KL(\beta\|\alpha) &= \int \beta(x)\ln\left[\frac{\beta(x)}{\alpha(x)}\right] && (41)\\
&= -\int \beta(x)\left[\ln\alpha(x) - \ln\beta(x)\right] && (42)\\
&= -\mathbb{E}_{x\sim\beta}\ln\alpha(x) - \operatorname{Ent}(\beta(x)) && (43)
\end{aligned}$$

where $\operatorname{Ent}(\beta(x)) = -\int_x \beta(x)\ln\beta(x)$.

(b) Compute the exact expression of $-\mathbb{E}_{x\sim\beta}\ln\alpha(x)$.

Solution We have

$$\begin{aligned}
\mathbb{E}_{x\sim\beta}\ln\alpha(x) &= \mathbb{E}_{x\sim\beta}\ln\prod_i \mathcal{N}(x_i|\mu,\sigma) && (44)\\
&= -\frac{d}{2}\ln(2\pi\sigma^2) - \mathbb{E}_{x\sim\beta}\left[\sum_i \frac{(x_i-\mu)^2}{2\sigma^2}\right] && (45)\\
&= -\frac{d}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i \mathbb{E}_{x_i\sim\beta_i}\left[(x_i-\mu)^2\right] && (46)\\
&= -\frac{d}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i \left(C_{ii} + (\mu_i-\mu)^2\right) && (47)
\end{aligned}$$

where we have marginalized β for each term of the sum in eq. (45) and where we have used:

$$\begin{aligned}\mathbb{E}_{x_i\sim\beta_i}\left[(x_i-\mu)^2\right] &= \mathbb{E}_{x_i\sim\beta_i}\left[(x_i-\mu_i+\mu_i-\mu)^2\right] && (48)\\ &= \mathbb{E}_{x_i\sim\beta_i}\left[(x_i-\mu_i)^2+2(x_i-\mu_i)(\mu_i-\mu)+(\mu_i-\mu)^2\right] && (49)\\ &= \mathbb{E}_{x_i\sim\beta_i}\left[(x_i-\mu_i)^2\right]+(\mu_i-\mu)^2 && (50)\\ &= C_{ii}+(\mu_i-\mu)^2 && (51)\end{aligned}$$

(c) Compute $KL(\beta||\alpha)$

Solution

$$\begin{aligned}KL(\beta||\alpha) &= -\mathbb{E}_{x\sim\beta}\left[\ln\alpha(x)\right]-\mathrm{Ent}(\beta) && (52)\\ &= \frac{d}{2}\ln(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_i\left(C_{ii}+(\mu_i-\mu)^2\right)-\frac{d}{2}(\ln(2\pi)+1)-\frac{1}{2}\ln|C| && (53)\\ &= \frac{1}{2}\left[d\ln(\sigma^2)-\ln|C|-d+\frac{1}{\sigma^2}\sum_i\left(C_{ii}+(\mu_i-\mu)^2\right)\right] && (54)\end{aligned}$$

(d) Suppose that $\mu_i=\mu$ and $C_{ii}=\sigma^2$ for all i. Simplify the previous expression.

Solution

$$KL(\beta||\alpha)=\frac{1}{2}\left[d\ln(\sigma^2)-\ln|C|\right] \qquad (55)$$

(e) How the mutual information could appear in the ELBO ?

References

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press. ISBN: 0521833787.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience. ISBN: 0471241954.

Programming exercises

Problem 5 (Text entropy)**.** *In the following, we are interested in estimating the entropy of different texts. We will work with the novel Crime and Punishment by Fyodor Dostoyevsky. Other books in different languages are also available.*[2]*. To do so, we compute the entropy of different models:*

1. Compute the entropy of a model based on the frequency of each single symbol in the chosen book (i.i.d. model).
2. Use this model to compute the cross-entropy of the distribution from another book. Compare this value with the previous entropy by computing the KL-divergence.
3. Compute the entropy of a model based on the frequency of pairs of symbols, and compare it with the previous model. Explain the difference.
4. Compute the entropy rate of a Markov chain where each state is a symbol, and transition probabilities are estimated from the chosen book.

[2] The chosen books are available at `https://www.lri.fr/~marceau/Courses/CentraleML2/texts.zip`, thanks to the Gutenberg project.`https://www.gutenberg.org/`