Foundations of machine Learning II

Project: Entropy

(Analytical part)

Flavie Vampouille

MSc in DS&BA

ESSEC Business School & CentralSupélec

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 $\mathbb{E}[f(X)] \leqslant f(\mathbb{E}[X])$. ln is a strictly concave function.

Concavity of log: Let f be a function f : dom f $\subseteq \mathbb{R}^n \to \mathbb{R}$, we know that: -f strictly convex <=> f strictly concave.

And if f is twice differentiable, f strictly convex <=> dom f convex and $\nabla^2 f$ positive definite ($\forall\, x \in dom$ f \ {0}, $x^T.\nabla^2 f.x$ > 0).

Hence log is a strictly concave function.

Definition of Kullback-Leibler distance: The relative entropy or Kullback-Leibler distance between two probability mass function $p(x)$ and $q(x)$ over a finite alphabet χ is:

$$\mathrm{KL}(\,p\;||\;q\,) \;=\; \sum_{x\,\in\,\chi} p(x)\log\frac{p(x)}{q(x)}$$

With $0.\log\frac{0}{0}$ = 0 , $0.\log\frac{0}{q}$ = 0 and $p.\log\frac{p}{0}$ = ∞.

Application of Jensen's inequality with log strictly concave function: Let $p(x)$, $q(x)$, $x \in \chi$ be two probability mass function. Then

$$\sum_{x\,\in\,\chi} p(x)\log\frac{q(x)}{p(x)} \leq \log \sum_{x\,\in\,\chi} p(x)\,\frac{q(x)}{p(x)}$$

with equality iif $\frac{q(x)}{p(x)}$ = c constant.

Gibbs' inequality: Let $p(x)$, $q(x)$, $x \in \chi$ be two probability mass function. Then

$$\mathrm{KL}(\,p\;||\;q\,) \geq 0$$

With equality iif $p(x)$ = $q(x)$ for all $x \in \chi$.

Proof: (Elements of Information Theory, Cover & Thomas, page 28, information inequality)

Let A = { $x : p(x) > 0$ } be the support set of $p(x)$. Then

$$
\begin{aligned}
- \mathrm{KL}(p \,||\, q) &= - \textstyle\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \textstyle\sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \textstyle\sum_{x \in A} p(x) \frac{q(x)}{p(x)} \\
&= \log \textstyle\sum_{x \in A} q(x) \\
&\leq \log \textstyle\sum_{x \in \chi} q(x) \\
&= \log (1) \\
&= 0
\end{aligned}
$$

We have equality iif

$\frac{q(x)}{p(x)}$ = c constant for all $x \epsilon \chi$ (Jensen)

and

$$\textstyle\sum_{x \in A} q(x) = \sum_{x \in \chi} q(x)$$

We hence have that c = 1, which means that for all $x \epsilon \chi$, $p(x) = q(x)$.

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

$$- \ln p(D) \leqslant -\mathbb{E}_{\theta \sim \beta} [\ln p(D|\theta)] + KL(\beta||\alpha) \tag{1}$$

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:

$$\pi(\theta) = \frac{p(D|\theta)\alpha(\theta)}{p(D)} \tag{2}$$

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

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

(d) Interpret the ELBO in a machine learning framework

(a)

$$\log \pi(\theta) = \log p(D \mid \theta) + \log \alpha(\theta) - \log p(D)$$

<=> $$0 = \log p(D \mid \theta) + \log \alpha(\theta) - \log p(D) - \log \pi(\theta)$$

(b)

$$\begin{aligned}
0 &= \int \beta(\theta) \left(\log p(D \mid \theta) + \log \alpha(\theta) - \log p(D) - \log \pi(\theta) \right) \\
&= \int \beta(\theta) \log p(D \mid \theta) + \int \beta(\theta) \log \alpha(\theta) - \int \beta(\theta) \log p(D) \\
&\qquad - \int \beta(\theta) \log \pi(\theta) + \int \beta(\theta) \log \beta(\theta) - \int \beta(\theta) \log \beta(\theta) \\
&= \int \beta(\theta) \left(\log p(D \mid \theta) \right) - \log p(D) - \int \beta(\theta) \log \frac{\beta(\theta)}{\alpha(\theta)} + \int \beta(\theta) \log \frac{\beta(\theta)}{\pi(\theta)} \\
&= \mathbb{E}_\beta\left(\log p(D \mid \theta) \right) - \log p(D) - \mathrm{KL}(\beta \,||\, \alpha) + \mathrm{KL}(\beta \,||\, \pi)
\end{aligned}$$

Thus

$$-\log p(D) = -\mathbb{E}_\beta\left(\log p(D \mid \theta) \right) + \mathrm{KL}(\beta \,||\, \alpha) - \mathrm{KL}(\beta \,||\, \pi)$$

(c)

Since we have (b) and $\mathrm{KL}(\beta \,||\, \pi) \geq 0$ (Gibbs' inequality). Then

$$\begin{aligned}
-\log p(D) &= -\mathbb{E}_\beta\left(\log p(D \mid \theta) \right) + \mathrm{KL}(\beta \,||\, \alpha) - \mathrm{KL}(\beta \,||\, \pi) \\
&\leq -\mathbb{E}_\beta\left(\log p(D \mid \theta) \right) + \mathrm{KL}(\beta \,||\, \alpha)
\end{aligned}$$

(d)

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

(a) univariate Normal distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{3}$$

(b) multivariate Normal distribution

$$\mathcal{N}(x|\mu, C) = \frac{1}{\sqrt{(2\pi)^d |C|}} \exp\left[-\frac{1}{2}(x-\mu)^\top C^{-1}(x-\mu)\right] \quad (4)$$

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

(a)

We take the logarithm of the univariate Normal distribution:

$$\log \mathcal{N}(\, x \mid \mu, \sigma^2 \,) \;=\; -\log(\, \sqrt{2\pi}\sigma \,) \;-\; \frac{(x-\mu)^2}{2\sigma^2}$$

The differential entropy of the univariate Normal distribution is then

$$\begin{aligned} \mathrm{H}(\, \mathcal{N}(\, x \mid \mu, \sigma^2 \,)\,) &= -\,\mathbb{E}[\, \log(\, \mathcal{N}(\, x \mid \mu, \sigma^2 \,)\,)\,] \\ &= \mathbb{E}[\, \log(\, \sqrt{2\pi}\sigma \,)\,] \;+\; \frac{1}{2\sigma^2}\,\mathbb{E}[\,(x-\mu)^2\,] \\ &= \log(\, \sqrt{2\pi}\sigma \,) \;+\; \frac{1}{2} \\ &= \log(\, \sqrt{2\pi e}\sigma \,) \end{aligned}$$

(b)

We take the logarithm of the multivariate Normal distribution:

$$\log \mathcal{N}(\, x \mid \mu, C \,) \;=\; -\,\frac{d}{2}\log 2\pi \;-\; \frac{1}{2}\log|C| \;-\; \frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)$$

The differential entropy of the multiivariate Normal distribution is then

$$\begin{aligned} \mathrm{H}(\, \mathcal{N}(\, x \mid \mu, C \,)\,) &= -\,\mathbb{E}[\, \log(\, \mathcal{N}(\, x \mid \mu, C \,)\,)\,] \\ &= \mathbb{E}[\, \frac{d}{2}\log 2\pi \;+\; \frac{1}{2}\log|C| \;+\; \frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\,] \\ &= \frac{d}{2}\log 2\pi \;+\; \frac{1}{2}\log|C| \;+\; \frac{1}{2}\mathbb{E}[\, \mathrm{tr}\,((x-\mu)^T C^{-1}(x-\mu))\,] \\ &= \frac{d}{2}\log 2\pi \;+\; \frac{1}{2}\log|C| \;+\; \frac{1}{2}\mathrm{tr}\,(C^{-1}\,\mathbb{E}[\,(x-\mu)^T(x-\mu)\,]\,) \\ &= \frac{d}{2}\log 2\pi \;+\; \frac{1}{2}\log|C| \;+\; \frac{1}{2}\mathrm{tr}\,(C^{-1}C) \\ &= \frac{d}{2}\log 2\pi \;+\; \frac{1}{2}\log|C| \;+\; \frac{d}{2} \\ &= \frac{d}{2}(\, 1 + \log 2\pi \,) \;+\; \frac{1}{2}\log|C| \\ &= \log(\sqrt{(2\pi e)^d|C|}\,) \end{aligned}$$

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

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

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

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

(a)

Entropy w.r.t. β: H($\beta(x)$) = $- \int \beta(x) \log \beta(x)$ then

$$\text{KL} (\beta \,||\, \alpha) = \int \beta(x) \log \frac{\beta(x)}{\alpha(x)}$$

$$= \int \beta(x) \log \beta(x) - \int \beta(x) \log \alpha(x)$$

$$= - \mathbb{E}_{x\sim\beta} [\log \alpha(x)] - \text{H}(\beta(x))$$

(b)

Note: in the introduction of problem 4 we have $\alpha = \prod_{i=1}^{d} \mathcal{N}(x_i \mid \mu, \sigma)$ and in question (d) $C_{ii} = \sigma^2$. So I suppose that $\alpha = \prod_{i=1}^{d} \mathcal{N}(x_i \mid \mu, \sigma^2)$

$$\mathbb{E}_{x\sim\beta} \log \alpha(x) = \mathbb{E}_{x\sim\beta} \log \prod_{i=1}^{d} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

$$= \mathbb{E}_{x\sim\beta} \sum_{i=1}^{d} \log \mathcal{N}(x_i \mid \mu, \sigma^2)$$

$$= \mathbb{E}_{x\sim\beta} [\sum_{i=1}^{d} (- \log (\sqrt{2\pi}\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2})]$$

$$= - d \log \left(\sqrt{2\pi}\sigma \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{d} (\mathbb{E}_{x_i\sim\beta_i}[(x_i - \mu)^2])$$

$$= - d \log \left(\sqrt{2\pi}\sigma \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{d} (\mathbb{E}_{x_i\sim\beta_i}[(x_i - \mu_i + \mu_i - \mu)^2])$$

$$= - d \log \left(\sqrt{2\pi}\sigma \right)$$
$$- \frac{1}{2\sigma^2} \sum_{i=1}^{d} (\mathbb{E}_{x_i\sim\beta_i}[(x_i - \mu_i)^2 + 2((x_i - \mu_i)(\mu_i - \mu) + (\mu_i - \mu)^2])$$

$$= - d \log \left(\sqrt{2\pi}\sigma \right)$$
$$- \frac{1}{2\sigma^2} \sum_{i=1}^{d} (\mathbb{E}_{x_i\sim\beta_i}[(x_i - \mu_i)^2] + \mathbb{E}_{x_i\sim\beta_i}[(\mu_i - \mu)^2])$$

$$= - d \log \left(\sqrt{2\pi}\sigma \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{d} (C_{ii} + (\mu_i - \mu)^2)$$

(c)

$$\mathrm{KL}(\beta \,||\, \alpha) = -\mathbb{E}_{x\sim\beta}[\log \alpha(x)] - \mathrm{H}(\beta(x))$$

$$= d \log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{d}\left(C_{ii} + (\mu_i - \mu)^2\right) - \log\left(\sqrt{(2\pi e)^d |C|}\right)$$

(d)

We suppose that $\mu_i = \mu$ and $C_{ii} = \sigma^2$. Then

$$\mathrm{KL}(\beta \,||\, \alpha) = d \log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{d}\sigma^2 - \log\left(\sqrt{(2\pi e)^d |C|}\right)$$

$$= d\left(\log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2}\right) - \log\left(\sqrt{(2\pi e)^d |C|}\right)$$

$$= \log\left(\sqrt{(2\pi e)^d}\sigma^d\right) - \log\left(\sqrt{(2\pi e)^d |C|}\right)$$

$$= \log\left(\frac{\sigma^d}{\sqrt{|C|}}\right)$$