Supervised Learning: Exercise of Information Theory Part 1

Yawei Li

yawei.li@stat.uni-muenchen.de

January 26, 2024

Exercise 1: Kullback-Leibler Divergence

(a) You want to approximate the binomial distribution with n numbers of trials and probability p with a Gaussian distribution with mean μ and variance σ^2. To find a suitable distribution you investigate the Kullback-Leibler divergence (KLD) in terms of the parameters $\boldsymbol{\theta} = (\mu, \sigma^2)^T$.

1. Write down the KLD for the given setup.
2. Derive the gradients w.r.t. $\boldsymbol{\theta}$.
3. Is there an analytic solution for the optimal parameter setting? If yes, derive the corresponding solution. If no, give a short reasoning.
4. Independent of the previous exercise, state a numerical procedure to minimize the KLD.

Solution to Question (a.1)

Question: Write down the KLD for the given setup.

- Let f be pmf of $Bin(n, p)$, we know that $\mathbb{E}_f[X] = np$ and $\mathrm{Var}_f[X] = np(1-p)$.
- Let q be the density of $\mathcal{N}(\mu, \sigma^2)$.
- The KL divergence between f andd q is

$$
\begin{aligned}
D_{KL}(f \,||\, q) = \mathbb{E}_f\left[\log \frac{f(X)}{q(X;\boldsymbol{\theta})}\right] &= \underbrace{\mathbb{E}_f\left[\log f(X)\right]}_{c_1} - \mathbb{E}_f[\log q(X|\boldsymbol{\theta})] \\
&= c_1 - \mathbb{E}_f\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(X-\mu)^2\right)\right)\right] \\
&= c_1 - \mathbb{E}_f\left[\underbrace{-\frac{1}{2}\log(2\pi)}_{c_2} - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(X-\mu)^2\right]
\end{aligned}
$$

Solution to Question (a.1): Continued

$$
\begin{aligned}
D_{KL}(f \,||\, q) &= \underbrace{c_1 + c_2}_{c_3} + \frac{1}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}_f[(X-\mu)^2] \\
&= c_3 + \log(\sigma) + \frac{1}{2\sigma^2}\left(\underbrace{\mathbb{E}_f[X^2]}_{\mathrm{Var}_f[X]+\mathbb{E}_f[X]^2} - 2\mu\mathbb{E}_f[X] + \mu^2\right) \\
&= c_3 + \log(\sigma) + \frac{1}{2\sigma^2}(np(1-p) + (np)^2 - 2\mu np + \mu^2) \\
&= c_3 + \log(\sigma) + \frac{1}{2\sigma^2}(np(1-p) + (np-\mu)^2)
\end{aligned}
$$

Solution to Question (a.2)

Question: Derive the gradients w.r.t. $\boldsymbol{\theta}$.

$$\frac{\partial D_{KL}\left(f \,||\, q\right)}{\partial \mu} = \frac{\partial}{\partial \mu} \frac{(np-\mu)^2}{2\sigma^2} = \frac{2(np-\mu)\cdot(-1)}{2\sigma^2} = \frac{\mu - np}{\sigma^2}$$

$$\begin{aligned}
\frac{\partial D_{KL}\left(f \,||\, q\right)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\left(\frac{1}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}(np(1-p) + (np-\mu)^2)\right) \\
&= \frac{1}{2\sigma^2} - \frac{1}{2(\sigma^2)^2}(np(1-p) + (np-\mu)^2) \\
&= \frac{1}{2\sigma^2} - \frac{1}{2\sigma^4}\mathbb{E}_f[(X-\mu)^2]
\end{aligned}$$

Solution to Question (a.3)

Question: Is there an analytic solution for the optimal parameter setting? If yes, derive the corresponding solution. If no, give a short reasoning.
Solution: Yes, there is. By setting the partial derivatives to zero we can identify the critical points.

$$\frac{\partial D_{KL}\left(f \,||\, q\right)}{\partial \mu} = 0 \Leftrightarrow \hat{\mu} = np = \mathbb{E}_f[X]$$

$$\begin{aligned}\frac{\partial D_{KL}\left(f \,||\, q\right)}{\partial \sigma^2} = 0 &\Leftrightarrow \frac{1}{2\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^2}\mathbb{E}_f[(X-\hat{\mu})^2] \\ &\Leftrightarrow \hat{\sigma}^2 = \mathbb{E}_f[(X-\mathbb{E}_f[X])^2] = \mathrm{Var}_f[X] = np(1-p)\end{aligned}$$

In order to prove that critical point is a minimum, we need to check the Hessian matrix w.r.t. (μ, σ^2) is positive definite at $(\mu, \hat{\sigma}^2)$.

Solution to Question (a.3): Continued

The Hessian matrix is:

$$\nabla^2 = \begin{pmatrix} \frac{1}{\sigma^2} & (np-\mu)\frac{1}{\sigma^4} \\ (np-\mu)\frac{1}{\sigma^4} & -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathbb{E}_f[(X-\mu)^2] \end{pmatrix}$$

Evaluated at $(\hat{\mu}, \hat{\sigma}^2)$, we obtain

$$\begin{aligned} \nabla^2\Big|_{(\hat{\mu},\hat{\sigma}^2)} &= \begin{pmatrix} \frac{1}{np(1-p)} & 0 \\ 0 & \frac{1}{2(np(1-p))^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\mathrm{Var}_f(X)} & 0 \\ 0 & \frac{1}{2\mathrm{Var}_f(X)^2} \end{pmatrix} \end{aligned}$$

So the Hessian at $(\hat{\mu}, \hat{\sigma}^2)$ is P.S.D.

Question (b), (c) and (d)

(b) Sample points according to the true distribution and visualize the KLD for different parameter settings of the Gaussian distribution (including the optimal one if available).

(c) Create a surface plot with axes n and p and colour value equal to the KLD for the optimal normal distribution.

(d) Based on the previous result, how can be behavior for varing p and n be explained, respectively?

Solution: Show the code.

Exercise 2: The Convexity of KL Divergence

Let p and q be the PDFs of a pair of absolutely continuous distributions.
(a) Prove that the KL divergence is convex in the pair (p, q), i.e.,

$$D_{KL}\left(\lambda p_1(1-\lambda)p_2 \,||\, \lambda q_1 + (1-\lambda)q_2\right) \leq \lambda D_{KL}\left(p1 \,||\, q_1\right) + (1-\lambda)D_{KL}\left(p_2 \,||\, q_2\right),$$

where (p_1, q_1) and (p_2, q_2) are two pairs of distribution and $0 \leq \lambda \leq 1$.

Hint: you can use the log sum inequality, namely that
$(a_1 + a_2)\log\left(\frac{a_1+a_2}{b_1+b_2}\right) \leq a_1 \log\frac{a_1}{b_1} + a_2 \log\frac{a_2}{b_2}$ holds for $a_1, a_2, b_1, b_2 \geq 0$.

Solution to Question (a)

$$
\begin{aligned}
& D_{KL}\left(\lambda p_1(1-\lambda)p_2 \,||\, \lambda q_1 + (1-\lambda)q_2\right) \\
&= \int_{\mathcal{X}} \left((\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \right) \mathrm{d}x \\
&\leq \int_{\mathcal{X}} \left(\lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \right) \mathrm{d}x \\
&= \lambda \int_{\mathcal{X}} \left(p_1(x) \log \frac{p_1(x)}{q_1(x)} \right) \mathrm{d}x + (1-\lambda) \int_{\mathcal{X}} \left(p_2(x) \log \frac{p_2(x)}{q_2(x)} \right) \mathrm{d}x
\end{aligned}
$$