MDL exercises, fifth handout

Solutions

31 March 2020

[1 free point]

1. Let $\{p_\theta | \theta \in \Theta\}$ $\Theta \subset \mathbb{R}$ be a smoothly parameterized i.i.d. 1-dimensional model (see page 65 in the book) and let $I(\theta)$ denote the Fisher information at θ. You may assume that, in the exercises below, the order of taking expectations and differentiating can be interchanged, i.e. the expected value of a derivative is the derivative of the expected value.

   (a) **[1 point]** Show that, for θ, θ' in the interior of Θ, the KL divergence (relative entropy) satisfies
   $$D(\theta \| \theta') = \frac{1}{2} I(\theta)(\theta - \theta')^2 + O((\theta - \theta')^3). \tag{1}$$

   Solution:

   We use the fact that $\log p_\gamma$ is infinitely differentiable on Θ, so that we can expand $p_{\theta'}$ around θ:

$$\begin{aligned}
D(\theta \| \theta') &= \mathbf{E}_{z \sim p_\theta}\left[-\log p_{\theta'}(z) + \log p_\theta(z)\right] \\
&= \mathbf{E}_{z \sim p_\theta}\left[-\log p_\theta(z) + (\theta' - \theta)\, \frac{d}{d\gamma} - \log p_\gamma(z)\Big|_{\gamma=\theta}\right. \\
&\quad \left. + \frac{1}{2}(\theta' - \theta)^2\, \frac{d^2}{d\gamma^2} - \log p_\gamma(z)\Big|_{\gamma=\theta} + \log p_\theta(z)\right] + O\left((\theta' - \theta)^3\right) \\
&= \mathbf{E}_{z \sim p_\theta}\left[(\theta' - \theta)\, \frac{d}{d\theta} - \log p_\theta(z)\Big|_{\theta=\theta} + \frac{1}{2}(\theta' - \theta)^2\, \frac{d^2}{d\theta^2} - \log p_\theta(z)\Big|_{\theta=\theta}\right] \\
&\quad + O\left((\theta' - \theta)^3\right) \\
&= (\theta' - \theta)\, \frac{d}{d\gamma} \mathbf{E}_{z \sim p_\theta}\left[-\log p_\gamma(z)\right]\Big|_{\gamma=\theta} + \mathbf{E}_{z \sim p_\theta}\left[\frac{1}{2}(\theta' - \theta)^2\, \frac{d^2}{d\theta^2} - \log p_\theta(z)\Big|_{\theta=\theta}\right] \\
&\quad + O\left((\theta' - \theta)^3\right) \\
&= : (*).
\end{aligned}$$

   By the information inequality, the p_γ that maximizes the expected likelihood $\mathbf{E}_{z \sim p_\theta}[\log p_\gamma(z)]$ is equal to the distribution that generates the data, i.e. p_θ. Therefore, $\frac{d}{d\gamma}\mathbf{E}[\log p_\gamma(z)]\Big|_{\gamma=\theta} = 0$, so we find:

$$\begin{aligned}
(*) &= \frac{1}{2}(\theta' - \theta)\mathbf{E}_{z \sim p_\theta}\left[\frac{d^2}{d\theta^2} - \log p_\theta(z)\Big|_{\theta=\theta}\right] + O\left((\theta' - \theta)^3\right) \\
&= \frac{1}{2}(\theta' - \theta) I(\theta) + O\left((\theta' - \theta)^3\right).
\end{aligned}$$

(b) **[1.5 points]** For a variety of models in their standard parameterizations, including the Poisson, geometric, normal and Bernoulli families, the following facts hold: (1) $I(\theta)$ is a continuous function of θ; (2) for every parameter θ and every sequence $x^n = x_1, \ldots, x^n$ such that both θ and the ML estimator $\hat{\theta}$ fall in the interior of Θ, we have:

$$\frac{1}{n}\left(-\log \frac{p_\theta(x^n)}{p_{\hat{\theta}}(x^n)}\right) = D(\hat{\theta}\|\theta) \tag{2}$$

Now suppose that we restrict the model to a subset Θ' of the interior of Θ where Θ' is some finite interval of length A. We discretize Θ' to a finite set $\ddot{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$ of m parameter values at distance $A/\sqrt{n}$, where $m = \sqrt{n} + 1$.

Now consider the two-part code that works as follows: the data x^n are encoded in two stages: we first code the $\theta \in \ddot{\Theta}$ that maximizes the probability of the data. Here we use a uniform code on $\ddot{\Theta}$. We then code the data using the Shannon-Fano code based on the θ we encoded in the first stage.

Assume that we get data such that, for all large n, $\hat{\theta} \in \Theta'$. Show, using (1) and (2) that the number of bits $L(x^n)$ we need to encode the data in this way satisfies

$$-\log p_{\hat{\theta}}(x^n) < L(x^n) \leq -\log p_{\hat{\theta}}(x^n) + \frac{1}{2}\log n + C$$

for some constant C independent of n.

Solution:

Firstly, we need $\log(m)$ bits to encode the $\theta \in \ddot{\Theta}$ that maximizes the probability of the data. Then the Shanon-Fano code has codelength $-\log p_\theta(x^n)$, so that the total codelength is given by

$$L(x^n) = \log(m) - \log p_\theta(x^n).$$

From $m = \sqrt{n} + 1$ it follows that $m > 1$ and so $\log(m) > 0$. Therefore

$$L(x^n) = \log(m) - \log p_\theta(x^n) > -\log p_\theta(x^n) \geq -\log p_{\hat{\theta}}(x^n),$$

since $\hat{\theta}$ maximizes the probability of the data overall. This concludes the lower bound. For the upper bound, we substitute (1) in (2) to see

$$\frac{1}{n}\left(-\log \frac{p_\theta(x^n)}{p_{\hat{\theta}}(x^n)}\right) = \frac{1}{2}I(\hat{\theta})(\hat{\theta} - \theta)^2 + O((\hat{\theta} - \theta)^3).$$

Rewriting this gives us

$$\log p_{\hat{\theta}}(x^n) - \log p_\theta(x^n) = \frac{n}{2}I(\hat{\theta})(\hat{\theta} - \theta)^2 + nO((\hat{\theta} - \theta)^3)$$

$$\Rightarrow -\log p_\theta(x^n) = -\log p_{\hat{\theta}}(x^n) + \frac{n}{2}I(\hat{\theta})(\hat{\theta} - \theta)^2 + nO((\hat{\theta} - \theta)^3).$$

Now, since $\theta \in \ddot{\Theta}$ maximizes the data in the discretized set and $\hat{\theta} \in \Theta'$ maximizes the data overall, we know $|\theta - \hat{\theta}| \leq \frac{A}{2\sqrt{n}}$. Therefore $n(\hat{\theta} - \theta)^2$ is a constant independent of n. Similarly, $nO((\hat{\theta} - \theta)^3)$ goes to 0 as n goes to infinity. Therefore, for large values of n:

$$-\log p_\theta(x^n) \leq -\log p_{\hat{\theta}}(x^n) + C,$$

from which it follows that

$$L(x^n) = \log(\sqrt{n} + 1) - \log p_\theta(x^n) \le \frac{1}{2}\log n - \log p_{\hat{\theta}}(x^n) + C.$$

This concludes the upper bound.

2. Consider the Bernoulli model. Compute the probability that the first two outcomes are different on the basis of four different universal models/codes:

- **[0.5 points]** The Bayesian model with uniform prior.

  Solution:
  To avoid confusion, we will denote $P_{M,U}$ for the Bayesian marginal probability with uniform prior. We have seen that $P_{M,U}(x^n) = \frac{1}{(n+1)\binom{n}{n_1}}$, where n_1 is the number of ones in x^n. Therefore, the following holds:

  $$P_M((0,1)) + P_M((1,0)) = 2 \cdot \frac{1}{(2+1)\binom{2}{1}} = 2 \cdot \frac{1}{6} = \frac{1}{3}.$$

- **[0.5 points]** The Bayesian model with Jeffrey's prior .

  Solution:
  Let us denote $P_{M,J}$ for the Bayesian marginal probability with Jeffrey's prior. Using the variation of Laplace's rule of succesion that holds for this universal model:

  $$P_{M,J}(X_{n+1} = 1 | X^n = x^n) = \frac{n_1 + \frac{1}{2}}{n+1},$$

  we see:

  $$\begin{aligned}
  &P_{M,J}((0,1)) + P_{M,J}((1,0)) \\
  &\qquad = P_{M,J}(X_1 = 1 | X_0 = 0) P_{M,J}(X_0 = 0) + P_{M,J}(X_1 = 0 | X_0 = 1) P_{M,J}(X_0 = 1) \\
  &\qquad = \frac{1}{4}\frac{1}{2} + (1 - \frac{3}{4})\frac{1}{2} \\
  &\qquad = \frac{1}{4}.
  \end{aligned}$$

- **[0.5 points]** The NML model for sample size 2.

  Solution:
  We use that the maximum likelihood estimator for a given sequence of data is given by $\hat{\theta}(x^n) = \frac{n_1}{n}$ and that $p_\theta(x^n) = \theta^{n_1}(1-\theta)^{n-n_1}$, to see:

  $$\begin{aligned}
  \hat{\theta}((0,0)) = 0 &\Rightarrow P_{\hat{\theta}((0,0))}((0,0)) = 1 \\
  \hat{\theta}((1,1)) = 1 &\Rightarrow P_{\hat{\theta}((1,1))}((1,1)) = 1 \\
  \hat{\theta}((1,0)) = \hat{\theta}((0,1)) = \frac{1}{2} &\Rightarrow P_{\hat{\theta}((0,1))}((0,1)) = P_{\hat{\theta}((1,0))}((1,0)) = \frac{1}{4}.
  \end{aligned}$$

Then the probability of the first two outcomes being different is:

$$\begin{aligned}
&P_{NML}{}^{(}(1,0)) + P_{NML}((0,1)) \\
&\quad = \frac{P_{\hat\theta((0,1))}((0,1)) + P_{\hat\theta((1,0))}((1,0))}{P_{\hat\theta((0,1))}((0,1)) = P_{\hat\theta((1,0))}((1,0)) + P_{\hat\theta((0,0))}((0,0)) + P_{\hat\theta((1,1))}((1,1))} \\
&\quad = \frac{\frac{1}{4}+\frac{1}{4}}{\frac{1}{4}+\frac{1}{4}+1+1} = \frac{1}{5}.
\end{aligned}$$

- **[0.5 points]** Similarly as above, we have

$$\begin{aligned}
&P_{\hat\theta((0,0,0))}((0,0,0)) = 1 \\
&P_{\hat\theta((1,1,1))}((1,1,1)) = 1 \\
&P_{\hat\theta((1,0,0))}((1,0,0)) = P_{\hat\theta((0,1,0))}((0,1,0)) = P_{\hat\theta((0,0,1))}((0,0,1)) = \frac{1}{3}\left(\frac{2}{3}\right)^2 \\
&P_{\hat\theta((1,1,0))}((1,1,0)) = P_{\hat\theta((1,0,1))}((1,0,1)) = P_{\hat\theta((0,1,1))}((0,1,1)) = \frac{1}{3}\left(\frac{2}{3}\right)^2.
\end{aligned}$$

Then the probability of the first two outcomes being different is:

$$P_{NML}((1,0,0)) + P_{NML}((0,1,0)) + P_{NML}((1,0,1)) + P_{NML}((0,1,1)) = \frac{4\cdot\frac{1}{3}\left(\frac{2}{3}\right)^2}{2+6\cdot\frac{1}{3}\left(\frac{2}{3}\right)^2} = \frac{8}{39}.$$

3. **[2 points]** Recall that the NML code is defined such that it has a constant regret of $\log\sum_{x^n} P(x^n|\hat\theta(x^n))$. With n_0 and n_1 defined as usual, show that in the case of the Bernoulli model this is equal to:

$$\log\sum_{x^n\in\mathcal{X}^n}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}.$$

Solution:

We know that $\hat\theta(x^n) = \frac{n_1}{n}$ and $P_\theta(x^n) = \theta^{n_1}(1-\theta)^{n-n_1}$, so that

$$P_{\hat\theta(x^n)}(x^n) = \left(\frac{n_1}{n}\right)^{n_1}\left(1-\frac{n_1}{n}\right)^{n-n_1} = \left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n-n_1}{n}\right)^{n-n_1},$$

using that, by definition, $n_0 = n - n_1$, summing and taking the log, we see:

$$\log\sum_{x^n\in\mathcal{X}^n} P_{\hat\theta(x^n)}(x^n) = \log\sum_{x^n\in\mathcal{X}^n}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}.$$

4. Suppose that we model data with a uniform distribution on the real numbers between 0 and $\theta > 0$.

   (a) **[1 point]** Given outcomes $x_1,\ldots,x_n$, what is the maximum likelihood value for θ?

   Solution:

The likelihood of the data is given by

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}[x_i \leq \theta] = \left(\frac{1}{\theta}\right)^n \mathbb{1}[\theta > \max_i x_i].$$

it is then clear that the maximum is somewhere in the interval $[\max_i x_i, \infty)$. On this interval, the log-likelihood of the data is

$$\log p(x_1, \ldots, x_n) = n \log \left(\frac{1}{\theta}\right).$$

Differentiating wrt θ, we see:

$$\frac{d}{d\theta} \log p(x_1, \ldots, x_n) = -\frac{n}{\theta}.$$

Since the derivative is negative, the likelihood is a decreasing function for $\theta \geq \max_i x_i$. Therefore, the maximum likelihood estimator is given by

$$\hat{\theta} = \max_i x_i.$$

(b) **[0.5 points]** Explain why a formula like (1) cannot be proven for the uniform distributions on $[0, \theta]$. In what way then is the model of uniform distributions crucially different from the Bernoulli and the normal family?

Solution:

As we saw above, the model of uniform distributions is not smoothly parameterized.

(c) **[1 point]** Show that (2) does hold for the uniform model.

Solution:

Let θ and $x^n = x_1, \ldots, x^n$, such that $\theta \geq \max_i x$ (so that $p_\theta(x^n) > 0$). Then:

$$\begin{aligned}
D(\hat{\theta} \| \theta) &= \mathbf{E}_{z \sim p_{\hat{\theta}}} \left[-\log p_\theta(z) + \log p_{\hat{\theta}}(z)\right] \\
&= \mathbf{E}_{z \sim p_{\hat{\theta}}} \left[-\log \frac{1}{\theta} + \log \frac{1}{\hat{\theta}}\right] \\
&= -\log \left(\frac{\left(\frac{1}{\theta}\right)}{\left(\frac{1}{\hat{\theta}}\right)}\right) \\
&= -\frac{1}{n} \log \left(\frac{\left(\frac{1}{\theta}\right)^n}{\left(\frac{1}{\hat{\theta}}\right)^n}\right) \\
&= -\frac{1}{n} \log \left(\frac{p_\theta(x^n)}{p_{\hat{\theta}}(x^n)}\right).
\end{aligned}$$