MDL exercises, fourth handout

Solutions

24 March 2020

1. (a) Let $H(p) = -p\log p - (1-p)\log(1-p)$ denote the binary entropy of a Bernoulli[p] distribution when the probability of observing a zero is p. (The logarithm is base two.) Use Stirling's approximation $\ln(n!) = (n+\frac{1}{2})\ln n - n + \frac{1}{2}\ln 2\pi + O(1/n)$ to show that $\log\binom{n}{\gamma n} = nH(\gamma) - \frac{1}{2}\log n + O(1)$.

   Below, we will abbreviate Stirling's approximation by SA. We see

$$
\begin{aligned}
\ln\binom{n}{\gamma n} &= \ln\left(\frac{n!}{(\gamma n)!(n-\gamma n)!}\right) \\
&= \ln(n!) - ln((\gamma n)!) - \ln((n(1-\gamma))!) \\
&\overset{\text{SA}}{=} (n+\frac{1}{2})\ln n - n + \frac{1}{2}\ln 2\pi + O(1/n) \\
&\quad - (\gamma n + \frac{1}{2})\ln(\gamma n) + \gamma n - \frac{1}{2}\ln 2\pi + O(1/(\gamma n)) \\
&\quad - (n(1-\gamma) + \frac{1}{2})\ln(n(1-\gamma)) + n(1-\gamma) - \frac{1}{2}\ln 2\pi + O(1/(n(1-\gamma))) \\
&= (n + \frac{1}{2} - \gamma n - \frac{1}{2} - n(1-\gamma) - \frac{1}{2})\ln(n) - (\gamma n + \frac{1}{2})\ln(\gamma) \\
&\quad - (n(1-\gamma) + \frac{1}{2})\ln(1-\gamma) - \frac{1}{2}\ln(2\pi) + O(1/n) \\
&= -\frac{1}{2}\ln(n) + n(-\gamma\ln(\gamma) - (1-\gamma)\ln(1-\gamma)) \\
&\quad - \frac{1}{2}\ln(\gamma) - \frac{1}{2}\ln(1-\gamma) - \frac{1}{2}\ln 2\pi + O(1/n) \\
&= -\frac{1}{2}\ln(n) + n(-\gamma\ln(\gamma) - (1-\gamma)\ln(1-\gamma)) + O(1),
\end{aligned}
$$

   where we have used that all constant terms and all $O(1/n)$ terms are $O(1)$. Finally, dividing by $\ln 2$ on both sides, we see

$$
\begin{aligned}
\log\binom{n}{\gamma n} &= -\frac{1}{2}\log n + n(-\gamma\log(\gamma) - (1-\gamma)\log(1-\gamma)) + O(1) \\
&= -\frac{1}{2}\log n + nH(\gamma) + O(1).
\end{aligned}
$$

(b) More generally, consider a sample space $\mathcal{X} = \{1, \ldots, k\}$ and probability mass functions p on $\mathcal{X}$, given in the form of a vector $p = (p_1, \ldots, p_k)$. Let $H(p) = \sum_{i=1}^{k} -p_i \log p_i$ denote the binary entropy of the distribution with mass function p. Use Stirling's approximation to express $\log \binom{n}{p_1 n \ldots p_k n} = n!/((p_1 n)! \ldots (p_k n)!)$ up to an $O(1)$ term.

Analogous to the previous exercise:

$$\begin{aligned}
\ln \binom{n}{p_1 n \ldots p_k n} &= \ln\left(n!/((p_1 n)! \ldots (p_k n)!)\right) \\
&= \ln(n!) - \sum_{i=1}^{k} \ln((p_i n)!) \\
&\overset{\text{SA}}{=} (n + \frac{1}{2}) \ln n - n + \frac{1}{2} \ln(2\pi) + O(1/n) \\
&\quad - \sum_{i=1}^{k} (p_i n + \frac{1}{2}) \ln(p_i n) - p_i n + \frac{1}{2} \ln(2\pi) + O(1/(p_i n)) \\
&= (n + \frac{1}{2}) \ln n - n + \sum_{i=1}^{k} p_i n - \sum_{i=1}^{k} (p_i n + \frac{1}{2}) \ln(p_i n) + O(1) \\
&= (n + \frac{1}{2}) \ln n - \sum_{i=1}^{k} (p_i n + \frac{1}{2}) \ln(p_i n) + O(1) \\
&= (n + \frac{1}{2} - \sum_{i=1}^{k} (p_i n + \frac{1}{2})) \ln n - n \sum_{i=1}^{k} p_i \ln p_i - \frac{1}{2} \sum_{i=1}^{k} \frac{1}{2} \ln p_i + O(1) \\
&= \frac{1-k}{2} \ln n - n \sum_{i=1}^{k} p_i \ln p_i + O(1).
\end{aligned}$$

Dividing by $\ln 2$ on both sides:

$$\begin{aligned}
\log \binom{n}{p_1 n \ldots p_k n} &= \frac{1-k}{2} \log n - n \sum_{i=1}^{k} p_i \log p_i + O(1) \\
&= \frac{1-k}{2} \log n + n H(p) + O(1).
\end{aligned}$$

Note that if we put $k = 2$, we indeed see that this is a generalisation of the formula given in the previous exercise.

2. Consider two codes for coding sequences of 0s and 1s. One is the Bayesian code with lengths $-\log P_M(x^n)$, where P_M is the Bayesian probability based on a uniform prior over the Bernoulli model. The other is the two-stage code where you first code the number of 1s n_1 in x^n using a uniform

code, and then you code the actual sequence with that number of 1's, using again a uniform code over all sequences of length n with n_1 1s.

Which code is better and why?

In the first handout, we proved that

$$P_M(x^n) = \frac{1}{(n+1)\binom{n}{n_1}},$$

so the Bayesian code has code length

$$L_{Bayes}(x^n) = -\log P_M(x^n) = \log(n+1) + \log\binom{n}{n_1}.$$

The two-stage code needs $\log(n+1)$ bits to encode n_1, because $n_1 \in \{0, 1, \ldots n\}$. Since there are $\binom{n}{n_1}$ sequences with n_1 ones, it needs $\log\binom{n}{n_1}$ bits to encode which sequence with n_1 ones it precisely is. Therefore the two-stage code has total code length

$$L_{2-stage}(x^n) = \log(n+1) + \log\binom{n}{n_1}.$$

We thus see that the codes have the same codelength for every x^n and are therefore equally good.

3. Markov chains.

   (a) Compute the maximum likelihood estimator $\hat{\theta} = (p_{0\to1}, p_{1\to1})$ for a binary first order Markov chain.

   By definition of the Markov chain, we have for any sequence x^n:

   $$P(x^n) = \frac{1}{2}\prod_{i=2}^{n} P(x_i|x_{i-1}).$$

   Now, let us denote with n_{ij} $(i, j \in \{0, 1\})$ the number of times a transition $i \to j$ occurs in x^n. Then we can rewrite the probability to

   $$P(x^n) = \frac{1}{2}\prod_{i=0}^{1}\prod_{j=0}^{1} p_{i\to j}^{n_{ij}}.$$

   Using that $p_{0\to0} = 1 - p_{0\to1}$ and $p_{1\to0} = 1 - p_{1\to1}$, we write

   $$P(x^n) = \frac{1}{2} p_{0\to1}^{n_{01}}(1 - p_{0\to1})^{n_{00}} p_{1\to1}^{n_{11}}(1 - p_{1\to1})^{n_{10}}.$$

   Taking the logarithm, we see

   $$\begin{aligned}\log P(x^n) = {} & \log(1/2) + n_{01}\log(p_{0\to1}) + n_{00}\log(1 - p_{0\to1}) \\ & + n_{11}\log(p_{1\to1}) + n_{10}\log(1 - p_{1\to1}).\end{aligned}$$

Differentiating with respect to $p_{0\to 1}$:

$$\frac{\partial}{\partial p_{0\to 1}} \log P(x^n) = \frac{n_{01}}{p_{0\to 1}} - \frac{n_{00}}{1 - p_{0\to 1}}.$$

Setting to zero to find the maximum likelihood value $\hat{p}_{0\to 1}$:

$$\frac{n_{01}}{\hat{p}_{0\to 1}} = \frac{n_{00}}{1 - \hat{p}_{0\to 1}} \Rightarrow \hat{p}_{0\to 1} = \frac{n_{01}}{n_{01} + n_{00}}.$$

Similar for $p_{1\to 1}$:

$$\frac{\partial}{\partial p_{1\to 1}} \log P(x^n) = \frac{n_{11}}{p_{1\to 1}} - \frac{n_{10}}{1 - p_{1\to 1}}.$$

Setting to zero to find the maximum likelihood value $\hat{p}_{1\to 1}$:

$$\frac{n_{11}}{\hat{p}_{1\to 1}} = \frac{n_{10}}{1 - \hat{p}_{1\to 1}} \Rightarrow \hat{p}_{1\to 1} = \frac{n_{11}}{n_{11} + n_{10}}.$$

So the maximum likelihood estimator is given by:

$$\hat{\theta} = \left(\frac{n_{01}}{n_{01} + n_{00}}, \frac{n_{11}}{n_{11} + n_{10}} \right).$$

(b) Draw X_1, X_2, X_3 from an order 1 Markov chain. Are X_1 and X_3 dependent? What if you know the value of X_2?

We see

$$P(X_3 = 1 | X_1 = 0) = p_{0\to 0} p_{0\to 1} + p_{0\to 1} p_{1\to 1} = (1 - p_{0\to 1}) p_{0\to 1} + p_{0\to 1} p_{1\to 1}$$

and

$$P(X_3 = 1 | X_1 = 1) = p_{1\to 0} p_{0\to 1} + p_{1\to 1} p_{1\to 1} = (1 - p_{1\to 1}) p_{0\to 1} + p_{1\to 1}^2.$$

Therefore $P(X_3 = 1 | X_1 = 0) \neq P(X_3 = 1 | X_1 = 1)$, so X_1 and X_3 are dependent.

If we know the value of X_2, then we see

$$P(X_3 = 1 | X_2 = x_2, X_1 = x_1) = p_{x_2 \to 1} = P(X_3 = 1 | X_2 = x_2),$$

so X_3 is independent of X_1, if we know X_2.