CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Minimum Description Length

Lecturer: Michael I. Jordan *Scribe: Mario Micheli*

1 The MDL perspective

As we said last time, description length may be thought of the number of digits in a binary string used to *code* the data for transmission. Therefore the MDL problem is actually equivalent to the one of code length minimization in a transmission channel (see Figure **??**); for an in-depth discussion of this topic please refer to [**?**].

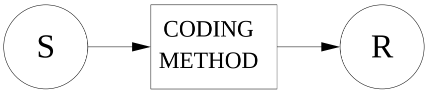

Figure 1: Source coding in a communication channel: S is the source, R is the receiver. The coding method depends on the source probability distribution, i.e. on $p(\mathcal{D}|\theta, M)$ (where M and θ are, respectively, the model and the parameter).

Given a certain probability distribution model M (e.g. Bernoulli) a first task for the sender is to estimate its parameters, and then to code with respect to the corresponding probability distribution. We will now briefly describe three coding algorithms, each assuming that the length n of string $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ is known to both the sender and the receiver.

1.1 Two-stage coding

Two-stage coding consists in the following steps (stages):

- fit a parameter and transmit that parameter estimate $\widehat{\theta}$;
- code using $p(\mathcal{D}|\widehat{\theta}, M)$.

Suppose that the sender uses a Bernoulli model,[1] with parameter p, for the source. Parameter p has to be estimated by the sender from the sequence of symbols and sent first. Let k be the number of ones in the string $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ of n bits: assuming that all symbols are iid then $\widehat{p}_{\mathrm{ML}} = k/n$. In order to transmit estimate $\widehat{p}$, $\log_2 n$ bits are necessary. Furthermore, from the form of distribution $\mathrm{Ber}(\widehat{p})$ it is easy to see that we can encode every symbol in the string at a cost of $-\log_2(k/n)$ bits for a one and $-\log_2(1-k/n)$ for a zero.[2] Thus, transmitting a sequence of n symbols requires a total code length of:

$$log_2 n + \left[-k \log_2 \frac{k}{n} - (n-k)\log_2\left(1 - \frac{k}{n}\right)\right].$$

[1] We will indicate a Bernoulli random variable X with: $X \sim \mathrm{Ber}(p)$, where p is its parameter.

[2] Remember from last time that the way to achieve the lower bound for the expected code length is taking $L_i = -\log_2 p_i$.

1.2 Mixture coding (with a uniform prior)

Suppose again that we postulate an iid Ber(p) model, but rather than estimating p we assign it a uniform prior density $\mathcal{U}[0,1)$; we can then apply the resulting mixture distribution to encode the binary string. The probability model that we get is the following:

$$p(\mathcal{D}|M) = \int_0^1 p^k(1-p)^k \, dp = \frac{k!(n-k)!}{(n+1)!}.$$

Therefore, the total code length of this (uniform) mixture code is given by:

$$-\log_2 k! - \log_2 (n-k)! + \log_2 (n+1)!$$

which is is related, but different, from the one we got with the previous algorithm. They should be compared, and the lower cost should be chosen.

Many mixture codes can be created making different choices for the prior density assigned to p, although for just a few of them an analytic (closed form) expression for distribution $p(\mathcal{D}|M)$ can be found (such prior densities are called *conjugate priors*).

1.3 Predictive coding

The idea this time is to estimate parameter p on-line, i.e. after each data point (*both* the sender *and* the receiver do it). So the key point is that *parameter estimate* $\widehat{p}$ *does not have to be transmitted.* In some sense, the predictive coder is designed to learn about parameter p from the past history, and hence improves the encoding for the next symbol.

Suppose that out of the past $t-1$ symbols in the sequence, k_{t-1} ones have occurred. At "time" t a new Bernoulli coder can be used with the Laplace estimator

$$\widehat{p}_{t-1} = \frac{k_{t-1}+1}{t+1}, \quad \widehat{p}_0 = \frac{1}{2}.$$

Therefore:

- for coding a 1, $-\log_2 \widehat{p}_{t-1}$ bits are needed;
- for coding a 0, $-\log_2(1-\widehat{p}_{t-1})$ bits are needed.

So, indicating with $q(\cdot)$ the joint probability distribution of $x^n \triangleq (x_1, x_2, \ldots, x_n)$, i.e.

$$q(x_1, x_2, \ldots, x_n) = \prod_{t=1}^{n} (\widehat{p}_{t-1})^{x_t} (1-\widehat{p}_{t-1})^{1-x_t},$$

we will have that the total number of bits that are necessary to code the n-symbol string will be given by:

$$\begin{aligned}
\sum_{t=1}^{n} L_t(x_t) &= -\log_2 q(x^n) \\
&= -\log_2 \prod_{t=1}^{n} (\widehat{p}_{t-1})^{x_t} (1-\widehat{p}_{t-1})^{1-x_t}
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{t=1}^{n}\left[x_t \log_2 \widehat{p}_{t-1} + (1-x_t)\log_2(1-\widehat{p}_{t-1})\right] \\
&= -\sum_{t:x_t=1} \log_2 \widehat{p}_{t-1} - \sum_{t:x_t=0} \log_2(1-\widehat{p}_{t-1}) \\
&= -\sum_{t:x_t=1} \log_2(k_{t-1}+1) - \sum_{t:x_t=0} \log_2(t-k_{t-1}) + \sum_{t=1}^{n} \log_2(t+1) \\
&= -\log_2 k! - \log_2(n-k)! + \log_2(n+1)!
\end{aligned}
$$

which is exactly the same expression we got for mixture coding. This is *not* a general fact, but it turns out to be true in the Bernoulli case.

Note that this reasoning makes use of so-called *fractional bits*. In practical terms it is not possible to send less than a bit of information at a time. However, if we delay transmission by a certain fixed number of bits, we can send a larger piece of data at a lower cost. In other words we do not use bits but *blocks*, and parameter estimation is performed after each block.

2 Pointwise lower bound for 2-stage coding

Now, for a given model M let there be k parameters. Discretize each parameter with precision[3] $\delta = 1/\sqrt{n}$. Then the necessary number of bits for transmitting the parameters will be:

$$-k \log_2 \frac{1}{\sqrt{n}} = \frac{k}{2} \log_2 n.$$

The total number of bits (parameter plus encoded string) is therefore given by:

$$-\log_2 p(x^n | [\widehat{\theta}_n]_\delta) + \frac{k}{2}\log_2 n \simeq -\log_2 p(x^n | \widehat{\theta}_{\mathrm{ML}}) + \frac{k}{2}\log_2 n$$

where the approximation is reasonable for smooth likelihoods. Note that such simplified two-stage MDL takes the form of BIC.

References

[1] Mark H. Hansen and Bin Yu, *Model Selection and the Principle of Minimum Description Length*, to appear in the Journal of the American Statistical Association, June 2001. Currently available at: `http://cm.bell-labs.com/who/cocteau/papers/index.html`

[3] This choice of precision is optimal in regular parametric families. An intuitive argument for this is the following: assume we wanted to estimate the mean μ of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$; then (assuming measurements are iid) the estimation error would have standard deviation: $\sqrt{E[\hat{\mu}-\mu]^2} = \sqrt{\sigma^2/n} = \sigma\frac{1}{\sqrt{n}}$. We cannot do any better than that.