Encoding Model Parameters

Jason Rennie
jrennie@ai.mit.edu

May 23, 2003

1 Introduction

Consider the problem of encoding parameters as described in [2]. Our overall goal is to encode the labels of documents from training data for a text classification problem. We encode the labels in two parts: (1) the parameters of our model, and (2) the labels given the knowledge of the model. We choose to encode the parameters stochastically. Instead of chosing an exact value, we choose a distribution and transmit a random value by drawing from the selected distribution. By the "bits-back" argument, our net expected encoding length is the KL-divergence between the selected distribution and a prior distribution. If we use parameterizations of the Gaussian as our distribution class, the KL-divergence is not only analytic, but simple to evaluate.

2 The "Bits-back" Argument

Let x be a parameter of our model that we would like to learn. Let $\mathcal{X}$ be the set of values that we may consider using. Let $-\log p(x)$ be the encoding length of value x, where p is a mass function on $\mathcal{X}$. Let q be a second mass function on $\mathcal{X}$. Instead of sending a single value, consider randomly drawing a value according to q. our expected encoding length is

$$E_{x\sim q}[-\log p(x)] = -\sum_{x\in\mathcal{X}} q(x)\log p(x). \tag{1}$$

But, since we don't care exactly what values are transmitted (as long as they are drawn from q), we can transmit additional information to the tune of $H(q)$ bits. This argument was introduced in [1]. Our net encoding length is

$$l(q) = -\sum_{x\in\mathcal{X}} q(x)\log p(x) - H(q) = \sum_{x\in\mathcal{X}} q(x)\log\frac{q(x)}{p(x)}. \tag{2}$$

We naturally arrive at the KL-divergence as a measure of the encoding length for the parameter.

3 Continuously Valued Parameters

This argument can be extended to continuously valued sets. We will use the family of Gaussian distributions as an example. Let $\mathcal{X} = \{\ldots, -2\delta, -\delta, 0, \delta, 2\delta, \ldots\}$, $\delta > 0$. Let

$$p(x) = \frac{1}{Z_p\delta\sqrt{\gamma^2}}\exp\left(-\frac{x^2}{2\gamma^2}\right) \tag{3}$$

be a mass function on $\mathcal{X}$, where γ^2 is a parameter, and let $-\log p(x)$ be the encoding length for the value x. Let

$$q(x) = \frac{1}{Z_q\delta\sqrt{\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

be a second mass function on $\mathcal{X}$ where σ^2 and μ are parameters. As discussed above, if we randomly draw values from q, our net expected encoding length is

$$\begin{aligned} l(q) &= \sum_{x\in\mathcal{X}} q(x)\log\frac{q(x)}{p(x)}, && (5) \\ &= \sum_{x\in\mathcal{X}} q(x)\left(-\frac{1}{2}\log\frac{\sigma^2 Z_q}{\gamma^2 Z_p} + \frac{x^2(\sigma^2-\gamma^2)+2x\mu\gamma^2-\mu^2\gamma^2}{2\sigma^2\gamma^2}\right). && (6) \end{aligned}$$

Now, note that as $\delta \to 0^+$, we observe the following limits: $Z_q \to \frac{1}{\sqrt{2\pi}}$, $Z_p \to \frac{1}{\sqrt{2\pi}}$, $\sum_x q(x)x \to \mu$, and $\sum_x q(x)x^2 \to \mu^2+\sigma^2$. Hence,

$$\lim_{\delta\to 0^+} l(q) = \frac{1}{2}\log\frac{\gamma^2}{\sigma^2} + \frac{\mu^2+\sigma^2-\gamma^2}{2\gamma^2}. \tag{7}$$

We have arrived at a net encoding length for transmitting values drawn randomly from a Gaussian distribution. The advantage of this is that we do not need to worry about selecting a discretization of the real number line. The variance on the distribution we choose (σ^2) acts as a knob for setting the precision with which we wish to transmit the parameter.

4 Summary

We have used the "bits-back" argument, introduced by [1], to argue that a continuously valued parameter can be encoded in a finite number of bits. We make use of stochastic encoding—randomly choosing the value according to a distribution. This allows us to avoid the problem of determining a discretization of the real number line. The encoding length is a KL-divergence between the chosen distribution and the prior. In the case of the Gaussian family, the encoding length is analytic and easy to compute.

References

[1] Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, minimum description length, and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*, 1994.

[2] Jason Rennie. Stochastic encoding and the "bits-back" argument. http://www.ai.mit.edu/~jrennie/writing/bitsback.pdf, May 2003.