Model Selection

Kaiwen Zhou

Contents

1 Two levels of inference

Two levels of inference can often be distinguished in the process of data modeling.

At the first level of inference, we assume that a particular model is true, and we fit that model to the data, i.e., we infer what values its free parameters should plausibly take, given the data. The results of this inference are often summarized by the most probable parameter values, and error bars on those parameters. This analysis is repeated for each model.

The second level of inference is the task of model comparison, in which the goal is to compare the models in the light of the data, and assign some sort of preference or ranking to the alternatives.

Model comparison is a difficult task because it is not possible simply to choose the model that fits the data best: more complex models can always fit the data better, so the maximum likelihood model choice would lead us inevitably to implausible, over-parameterized models, which generalize poorly. Perhaps Bayes' theorem can help with this somewhat-difficult task.

Let us write down Bayes' theorem for the two levels of inference described above. Each model m_i is assumed to have a vector of parameters θ. A model is defined by a collection of probability distributions: a 'prior' distribution

$$p\left(\theta \mid m_i\right)$$

which states what values the model's parameters might be expected to take; and a set of conditional distributions,

$$p\left(D \mid \theta, m_i\right),$$

one for each value of θ, defining the predictions that the model makes about the data D.

1.1 The first level of inference

At the first level of inference, we focus on one model, the i-th model, and we infer what the model's parameters θ might be, given the data D.

Using Bayes' theorem, the posterior probability of the parameters θ is:

$$p\left(\theta \mid D, m_i\right)=\frac{p\left(D \mid \theta, m_i\right) p\left(\theta \mid m_i\right)}{p\left(D \mid m_i\right)}$$

which can also be written

$$\text{posterior }=\frac{\text{likelihood } \times \text{ prior}}{\text{evidence}}$$

The normalizing constant

$$p\left(D \mid m_i\right)$$

is irrelevant to the first level of inference, i.e., the inference of θ; but it becomes important in the second level of inference, and we name it the evidence for m_i.

It is common practice to use gradient-based methods to find the maximum of the posterior, which defines the most probable value for the parameters, $\hat{\theta}_{MAP}$; it is then usual to summarize the posterior distribution by the value of $\hat{\theta}_{MAP}$, and error bars or confidence intervals on these fitted parameters. Error bars can be obtained from the curvature of the negative log-posterior. Evaluating the Hessian at $\hat{\theta}_{\text{MAP}}$,

$$A=-\left.\nabla^2 \ln p\left(\theta \mid D, m_i\right)\right|_{\hat{\theta}_{MAP}} \tag{1}$$

we may then see that the posterior can be locally approximated as a Gaussian with covariance matrix A^{-1}. Indeed, Taylor-expanding the log-posterior around the point $\hat{\theta}_{\text{MAP}}$ and then re-exponentiating gives

$$p\left(\theta \mid D, m_i\right) \approx p\left(\hat{\theta}_{MAP} \mid D, m_i\right) \exp \left(-\frac{1}{2} \Delta \theta^{\top} A \Delta \theta\right) \quad \text{where} \quad \Delta \theta=\hat{\theta}_{MAP}-\theta$$

1.2 The second level of inference

At the second level of inference, we wish to infer which model is the most plausible given the data. This entails computing the posterior over models,

$$p(m \mid D)=\frac{p(D \mid m)p(m)}{\sum_{m \in M} p(m, D)} \tag{2}$$

where D is the data.

Given the posterior over models, one can in principle compute the model with the greatest posterior probability, and also perform pairwise comparisons of models. The model which is most plausible given the data (the maximum a-posteriori or MAP model) is computed as

$$\hat{m}=\underset{m}{\operatorname{argmax}}\, p(m \mid D).$$

This is called Bayesian model selection.

Pairwise comparison of models, say m_1 and m_2, is summarized by the posterior odds

$$\frac{p\left(m_1 \mid D\right)}{p\left(m_2 \mid D\right)}=\frac{p\left(D \mid m_1\right)}{p\left(D \mid m_2\right)} \times \frac{p\left(m_1\right)}{p\left(m_2\right)}$$

The ratio

$$\frac{p\left(D \mid m_1\right)}{p\left(D \mid m_2\right)}$$

is called the Bayes factor.

Thus the data, through the Bayes factor, updates the prior odds to yield the posterior odds.

Example 1.1. *As an example, let's consider Bayesian model selection with a uniform prior over models,*

$$p(m) \propto 1$$

In this special case, MAP model selection amounts to picking the model which maximizes

$$p(D \mid m)=\int p(D \mid \theta) p(\theta \mid m) d \theta \tag{3}$$

This is because the ratio $\frac{p(m_1)}{p(m_2)}$ *is* 1 *in this special case.*

The quantity Equation (3) is called the marginal likelihood, the integrated likelihood, or the evidence for model m. Hence model selection is picking the model which has the most evidence for it. How to compute (an approximation of) this integral will be discussed below.

2 Gaussian approximation

We now engage in a further discussion of the Gaussian approximation to a posterior distribution. One of the main points is that the integral in Equation (3), which represents the "normalizing constant" at the first level of inference,

$$p(D \mid m)=\int p(D \mid \theta) p(\theta \mid m) d \theta$$

is hard to compute.

We will approximate this integral by approximating the posterior distribution, of which this is the normalizing factor, by a Gaussian distribution. As this discussion concerns a single model m, we will in this section sometime suppress the model m from the notation. The approximation works as follows: suppose $\theta \in \mathbb{R}^k$, and let

$$p(\theta \mid D)=\frac{1}{Z} e^{-E(\theta)}$$

where $E(\theta)$ is called an energy function and is equal to the negative log of the unnormalized posterior,

$$E(\theta)=-\log p(\theta \mid D)-\log Z \quad \text{with} \quad Z=p(D)$$

being the normalization constant.

Performing a Taylor series expansion around the mode θ^* (i.e., the lowest energy state) we get

$$E(\theta) \approx E\left(\theta^*\right)+\left(\theta-\theta^*\right)^{\top} \mathbf{g}+\frac{1}{2}\left(\theta-\theta^*\right)^{\top} \mathbf{H}\left(\theta-\theta^*\right)$$

where g is the gradient and $\mathbf{H}$ is the Hessian of the "energy function" evaluated at the mode:

$$\mathbf{g}=\left.\nabla E(\theta)\right|_{\theta^*} \quad \mathbf{H}=\left.\frac{\partial^2 E(\theta)}{\partial \theta \partial \theta^{\top}}\right|_{\theta^*} .$$

Since θ^* is the mode, the gradient is zero there. Hence, defining $\hat{p}(\theta \mid D)$ as the associated normal density,

$$\begin{aligned}
& \hat{p}(\theta \mid D) \approx \frac{1}{Z} e^{-E\left(\theta^*\right)} \exp \left[-\frac{1}{2}\left(\theta-\theta^*\right)^{\top} \mathbf{H}\left(\theta-\theta^*\right)\right] \xlongequal[\text{by only the mean and the covaraince}]{\text{Since the Gaussian distribution is determined}} \mathcal{N}\left(\theta \mid \theta^*, \mathbf{H}^{-1}\right) \\
\Longrightarrow & \frac{1}{Z} e^{-E\left(\theta^*\right)}=\frac{1}{\sqrt{(2 \pi)^k|\mathbf{H}|}} \Longrightarrow Z=p(D)=e^{-E\left(\theta^*\right)}(2 \pi)^{-k / 2}|\mathbf{H}|^{-1 / 2}
\end{aligned}$$

The last line follows from the (well-known) normalization constant of the multivariate Gaussian.

A Gaussian approximation is often reasonable in the large-sample situation, since posteriors often become more "Gaussian-like" as the sample size increases, for reasons analogous to the central limit theorem. The same mathematical approximation method is often used in physics where it is usually referred to as the saddle point approximation, or Laplace's method.

We can use the Gaussian approximation to write the log marginal likelihood as follows, dropping irrelevant constants:

$$\log p(D) \approx \log p\left(D \mid \theta^*\right)+\log p\left(\theta^*\right)-\frac{1}{2} \log |\mathbf{H}| \tag{4}$$

The penalization terms which are added to $\log p\left(D \mid \theta^*\right)$ are a measure of model complexity.

If we have a uniform prior,

$$p(\theta) \propto 1,$$

we can drop the second term, and replace θ^* with the MLE, $\hat{\theta}$.

For many problems the parameter posterior (as distinguished from the posterior on model space)

$$p\left(\theta \mid D, m_i\right) \propto p\left(D \mid \theta, m_i\right) p\left(\theta \mid m_i\right)$$

has a strong peak at the most probable parameters $\hat{\theta}_{\text{MAP}}$.

We can thus use Laplace's method to approximate this posterior near its peak. Taking for simplicity the one-dimensional case, the evidence can be approximated, using Laplace's method, by the height of the peak of the integrand

$$p\left(D \mid \theta, m_i\right) p\left(\theta \mid m_i\right)$$

times its width, $\sigma_{\theta \mid D}$:

$$p\left(D \mid m_i\right) \approx p\left(D \mid \hat{\theta}_{MAP}, m_i\right) \times p\left(\hat{\theta}_{MAP} \mid m_i\right) \sigma_{\theta \mid D} \text{ Evidence } \approx \text{ Best fit likelihood } \times \text{ Occam factor}$$

The quantity $\sigma_{\theta \mid D}$ is the posterior uncertainty in θ. Suppose, for simplicity of exposition, that the prior

$$p\left(\theta \mid m_i\right)$$

is uniform on some rather large interval of size σ_w, representing the range of values of θ that were possible a priori, according to m_i.

Then

$$p\left(\hat{\theta}_{MAP} \mid m_i\right)=\frac{1}{\sigma_w} \quad \text{and} \quad \text{Occam factor } := p\left(\hat{\theta}_{MAP} \mid m_i\right) \sigma_{\theta \mid D}=\frac{\sigma_{\theta \mid D}}{\sigma_w}$$

so the Occam factor is equal to the ratio of the posterior accessible volume of m_i 's parameter space to the prior accessible volume, or the factor by which m_i 's hypothesis space collapses when the data arrive.

Intuition: the model m_i can be viewed as consisting of a certain number of exclusive submodels, of which only one survives when the data arrive.

The Occam factor is the inverse of that number. The logarithm of the Occam factor is a measure of the amount of information we gain about the model's parameters when the data arrive. A complex model having many parameters, each of which is free to vary over a large range σ_w, will typically be penalized by a stronger Occam factor than a simpler model. If the posterior is well approximated by a Gaussian, then the Occam factor is obtained from the determinant of the corresponding covariance matrix

$$\begin{gathered}p(D \mid m) \approx p\left(D \mid \hat{\theta}_{MAP}, m\right) \times p\left(\hat{\theta}_{MAP} \mid m\right) \operatorname{det}^{-1/2}(A / 2 \pi) \\ \text{Evidence } \approx \text{ Best fit likelihood } \times \text{ Occam factor}\end{gathered}$$

where $A=-\left.\nabla^2 \ln p\left(\theta \mid D, m_i\right)\right|_{\hat{\theta}_{MAP}}$, is the Hessian as computed above in Equation (1).

In summary, Bayesian model selection with a uniform prior over models comes down to finding the model m which maximizes

$$p(D \mid m)$$

This can be viewed as an extension of maximum likelihood model selection: the evidence is obtained by multiplying the best-fit likelihood by the Occam factor.

To evaluate the Occam factor we need only the Hessian A, if the Gaussian approximation is good. Thus the Bayesian method of model comparison by evaluating the evidence is no more computationally demanding than the task of finding for each model the best-fit parameters and their error bars.

2.1 Bayesian Occam's razor effect

One might think that using $p(D \mid m)$ to select models would always favor the model with the most parameters. Interestingly, this is false. It would be true if we were to use $p\left(D \mid \hat{\theta}_m\right)$ to select models, where $\hat{\theta}_m$ is the MLE or MAP estimate of the parameters for model m, because models with more parameters will fit the data better, and hence achieve higher likelihood (the probability of the observed data D to appear given our parameters $\hat{\theta}_m$). However, if we integrate out the parameters, rather than maximizing them, we are automatically protected from overfitting: models with more parameters do not necessarily have higher marginal likelihood. This is called the Bayesian Occam's razor effect and was discovered by Jefferys and Berger (1992). See also Mackay (1995).

Intuition: A way to understand the Bayesian Occam's razor is to note that probabilities must sum to one. Hence, the sum over all possible data sets gives

$$\sum_D p(D \mid m)=1$$

Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models.

3 Derivation of the BIC

We now focus on approximating the third term in Equation (4) which is $\log |\mathbf{H}|$. Suppose we have sample size n and D_i is the predictor-response pair in the i-th sample.

$$\mathbf{H}=\sum_{i=1}^{n} \mathbf{H}_i \quad \text{where} \quad \mathbf{H}_i=\nabla^2 \log p\left(D_i \mid \theta\right)$$

Let us approximate each $\mathbf{H}_i$ by a fixed matrix $\hat{\mathbf{H}}$. Then we have

$$\log |\mathbf{H}| \approx \log |n \hat{\mathbf{H}}|=\log \left(n^k|\hat{\mathbf{H}}|\right)=k \log n+\log |\hat{\mathbf{H}}|$$

where $k=\operatorname{dim}(\theta)$ and we have assumed $\mathbf{H}$ is full rank.

We can drop the $\log |\hat{\mathbf{H}}|$ term, since it is independent of n, and thus will get overwhelmed by the likelihood. Putting all the pieces together, the desired approximation is

$$\log p(D) \approx \log p(D \mid \hat{\theta})-\frac{k}{2} \log n \tag{5}$$

Example 3.1. *Consider linear regression. The MLE is given by*

$$\hat{\theta}=\left(X^{\top} X\right)^{-1} X^{\top} y$$

and

$$\hat{\sigma}^2=\frac{\mathrm{RSS}}{n} \quad \textit{where} \quad \mathrm{RSS}=\sum_{i=1}^n\left(y_i-\hat{\theta}^{\top} x_i\right)^2 .$$

The corresponding log likelihood is given by

$$\log p(D \mid \hat{\theta})=-\frac{n}{2} \log \left(2 \pi \hat{\sigma}^2\right)-\frac{n}{2}$$

Hence the approximation Equation (5) gives (dropping constant terms)

$$\log p(D) \approx-\frac{n}{2} \log \left(\hat{\sigma}^2\right)-\frac{k}{2} \log n$$

where k is the number of variables in the model.

The negative of this, up to a factor of 2, is often called the Bayesian information criterion and written

$$\mathrm{BIC}=n \log \left(\hat{\sigma}^2\right)+k \log n$$

in which case, if used in model selection, the model with the lower BIC is to be preferred.

4 Minimum Description Length

We first review the theory of coding for data compression, and then apply it to model selection. If you're curious for a full treatment, the seminal book is Cover and Thomas (2012). We think of a certain block of data, z, as a message that we want to encode and send to someone else (the "receiver"). We think of our model as a way of encoding the datum, and will choose the most parsimonious model, that is the shortest code, for the transmission.

Suppose first that the possible messages we might want to transmit are

$$z_1, z_2, \ldots, z_m$$

We want a way of encoding which from among this finite set of possible messages is actually being transmitted, using some form of binary encoding, i.e. each message will be mapped somewhat arbitrarily onto a sequence of 1s and 0s.

Here is an example with four possible messages:

Message	z_1	z_2	z_3	z_4
Code	0	10	110	111

This code is known as an instantaneous prefix code - no code is the prefix of any other, and the receiver (who knows all of the possible codes), knows exactly when the message has been completely sent. We restrict our discussion to such instantaneous prefix codes. We could permute the codes, for example use codes

$$110,10,111,0$$

for

$$z_1, z_2, z_3, z_4$$

The best way to do this depends on how often we will be sending each of the messages. If, for example, we will be sending z_1 most often, it makes sense to use the shortest code 0 for z_1. Using this kind of strategy-shorter codes for more frequent messages-the average message length will be shorter. In general, if messages are sent with probabilities p_i, Shannon's theorem says we should use code lengths

$$\ell_i=-\log_2 p_i$$

and the average message length satisfies

$$\mathbb{E}(\text{ length })\geq-\sum_i p_i \log_2 p_i$$

The right-hand side above is also called the entropy of the distribution. In other words: To transmit a random variable z having probability density function $p(z)$, we require about

$$-\log_2 p(z)$$

bits of information.

This can be applied to continuous random variables also. With a finite code length we cannot code a continuous variable exactly. However, if we code z within a tolerance δz, the message length needed is the log of the probability in the interval

$$[z, z + \delta z]$$

which is well approximated by $\delta z p(z)$ if δz is small.

Since

$$\log[\delta z p(z)] = \log \delta z + \log p(z)$$

this means we can just ignore the constant $\log \delta z$ and regard

$$\log p(z)$$

as our measure of message length.

Now we apply this result to the problem of model selection in a supervised learning context. We have a model m with parameters θ, and data

$$Z = (X, y)$$

consisting of both inputs and outputs.

Let the (conditional) probability of the outputs under the model be

$$p(y \mid \theta, m, X)$$

assume the receiver knows all of the inputs, and we wish to transmit the outputs.

Then the message length required to transmit the outputs is

$$\text{length } = -\log p(y \mid \theta, m, X) - \log p(\theta \mid m)$$

The second term is the average code length for transmitting the model parameters θ, while the first term is the average code length for transmitting the discrepancy between the model and actual target values. The MDL principle says that we should choose the model that minimizes (12.10). We recognize (12.10) as the (negative) log-posterior distribution, and hence minimizing description length is equivalent to maximizing posterior probability. Hence the BIC criterion, derived as approximation to log-posterior probability, can also be viewed as a device for (approximate) model choice by minimum description length.

5 Model Averaging

Suppose we want to predict a new observation y. Let

$$D = \{y_1, \ldots, y_n\}$$

be the observed data so far. Then

$$p(y \mid D) = \sum_j p\left(y \mid D, m_j\right) p\left(m_j \mid D\right)$$

where the sum is over the different models being considered. Recalling (12.2), we have

$$p\left(m_j \mid D\right) = \frac{p\left(D \mid m_j\right)}{\sum_j p\left(m_j, D\right)}$$

when we have a uniform prior over models.

Moreover, according to (12.8), the log of the numerator of the last equation becomes

$$\log p\left(D \mid m_j\right) \approx \log p\left(D \mid \hat{\theta}, m_j\right) - \frac{k_j}{2} \log n \equiv -2 \cdot \mathrm{BIC}_j$$

This means we can approximate

$$p\left(m_j \mid D\right)$$

by

$$\exp\left(-2 \cdot \mathrm{BIC}_j\right) / \left(\sum_j \exp\left(-2 \cdot \mathrm{BIC}_j\right)\right)$$

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[2] William H. Jefferys and James O. Berger. Ockham's razor and Bayesian analysis. In: *American Scientist*, Vol. 80.1, 1992, pp. 64-72.

[3] David J.C. MacKay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. In: *Network: Computation in Neural Systems*, Vol. 6.3, 1995, pp. 469-505.