Bayesian model selection

Consider the regression problem, where we want to predict the values of an unknown function $y\colon \mathbb{R}^d \to \mathbb{R}$ given examples $\mathcal{D} = \big\{(\mathbf{x}_i, y_i)\big\}_{i=1}^N$ to serve as training data. In Bayesian linear regression, we made the following assumption about $y(\mathbf{x})$:

$$y(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w} + \varepsilon(\mathbf{x}), \tag{1}$$

where $\phi(\mathbf{x})$ is a now explicitly-written feature expansion of $\mathbf{x}$. We proceed in the normal Bayesian way: we place Gaussian priors on our unknowns, the parameters $\mathbf{w}$ and the residuals $\boldsymbol{\varepsilon}$, then derive the posterior distribution over $\mathbf{w}$ given $\mathcal{D}$, which we use to make predictions.

One question left unanswered is how to choose a good feature expansion function $\phi(\mathbf{x})$. For example, a purely linear model could use $\phi(\mathbf{x}) = [1, \mathbf{x}]^\top$, whereas a quadratic model could use $\phi(\mathbf{x}) = [1, \mathbf{x}, \mathbf{x}^2]^\top$, etc. In general, arbitrary feature expansions ϕ are allowed. How can I select between them? Even more generally, how do I select whether I should use linear regression or a completely different probabilistic model to explain my data? These are questions of *model selection,* and naturally there is a Bayesian approach to it.

Before we continue our discussion of model selection, we will first define the word *model,* which is often used loosely without explicit definition. A model is a parametric family of probability distributions, each of which can explain the observed data. Another way to explain the concept of a model is that if we have chosen a likelihood $p(\mathcal{D} \mid \theta)$ for our data, which depends on a parameter θ, then the model is the set of all likelihoods (each one of which is a distribution over $\mathcal{D}$) for every possible value of the parameter θ.

In the case of linear regression, the weight vector $\mathbf{w}$ defines the parametric family, and the model is the set of distributions

$$\big\{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2)\big\} = \big\{\mathcal{N}(\mathbf{y}; \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})\big\},$$

indexed by all possible $\mathbf{w}$. Each one of these is a potential explanation of the observed values $\mathbf{y}$ given $\mathbf{X}$. In the case of flipping a coin n times with an unknown bias θ and observing the number of heads x, the model is

$$\big\{p(x \mid n, \theta)\big\} = \big\{\mathrm{Binomial}(x, n, \theta)\big\},$$

where there is one binomial distribution for every possible $\theta \in (0, 1)$. In the Bayesian method, we maintain a belief over which elements in the model we consider plausible by reasoning about $p(\theta \mid \mathcal{D})$ via Bayes' theorem.

Suppose now that I have at my disposal a finite set of models $\{\mathcal{M}_i\}_i$ that I may use to explain my observed data $\mathcal{D}$, and let us write θ_i for the parameters of model $\mathcal{M}_i$. How do we know which model to prefer? We work out the posterior probability over the models via Bayes' theorem! We have:

$$\Pr(\mathcal{M}_i \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}_i)\Pr(\mathcal{M}_i)}{\sum_j p(\mathcal{D} \mid \mathcal{M}_j)\Pr(\mathcal{M}_j)}.$$

Here $\Pr(\mathcal{M}_i)$ is a prior distribution over models that we have selected; a common practice is to set this to a uniform distribution over the models. The value $p(\mathcal{D} \mid \mathcal{M}_i)$ may also be written in a more-familiar familiar form:

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \theta_i, \mathcal{M}_i) p(\theta_i \mid \mathcal{M}_i)\,\mathrm{d}\theta_i.$$

This is exactly the denominator when applying Bayes' theorem to find the posterior $p(\theta_i \mid \mathcal{D}, \mathcal{M}_i)$!

$$p(\theta_i \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \theta_i, \mathcal{M}_i)p(\theta_i \mid \mathcal{M}_i)}{\int p(\mathcal{D} \mid \theta_i, \mathcal{M}_i)p(\theta_i \mid \mathcal{M}_i)\,\mathrm{d}\theta_i.} = \frac{p(\mathcal{D} \mid \theta_i, \mathcal{M}_i)p(\theta_i \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)},$$

where we have simply conditioned on $\mathcal{M}_i$ to be explicit. In the context of model selection, the term $p(\mathcal{D} \mid \mathcal{M}_i)$ is known as the *model evidence* or simply *evidence*. One interpretation of the model evidence is the probability that your model could have generated the observed data, under the chosen prior belief over its parameters θ_i.

Suppose now that we have exactly two models for the observed data that we wish to compare: $\mathcal{M}_1$ and $\mathcal{M}_2$, with corresponding parameter vectors θ_1 and θ_2 and prior probabilities $\Pr(\mathcal{M}_1)$ and $\Pr(\mathcal{M}_2)$. In this case it is easiest to compute the *posterior odds*, the ratio of the models' probabilities given the data:

$$\frac{\Pr(\mathcal{M}_1 \mid \mathcal{D})}{\Pr(\mathcal{M}_2 \mid \mathcal{D})} = \frac{\Pr(\mathcal{M}_1)p(\mathcal{D} \mid \mathcal{M}_1)}{\Pr(\mathcal{M}_2)p(\mathcal{D} \mid \mathcal{M}_2)} = \frac{\Pr(\mathcal{M}_1)\int p(\mathcal{D} \mid \theta_1, \mathcal{M}_1)p(\theta_1 \mid \mathcal{M}_1)\,\mathrm{d}\theta_1}{\Pr(\mathcal{M}_2)\int p(\mathcal{D} \mid \theta_2, \mathcal{M}_2)p(\theta_2 \mid \mathcal{M}_2)\,\mathrm{d}\theta_2},$$

which is simply the prior odds multiplied by the ratio of the evidence for each model. The latter quantity is also called the *Bayes factor* in favor of $\mathcal{M}_1$. Publishing Bayes factors allows another practitioner to easily substitute their own model priors and derive their own conclusions about the models being considered.

Example

Wikipedia gives a truly excellent example of Bayesian model selection in practice.[1] Suppose I am presented with a coin and want to compare two models for explaining its behavior. The first model, $\mathcal{M}_1$, assumes that the heads probability is fixed to $1/2$. Notice that this model does not have any parameters. The second model, $\mathcal{M}_2$, assumes that the heads probability is fixed to an unknown value $\theta \in (0, 1)$, with a uniform prior on θ: $p(\theta \mid \mathcal{M}_2) = 1$ (this is equivalent to a beta prior on θ with $\alpha = \beta = 1$). For simplicity, we choose a uniform model prior: $\Pr(\mathcal{M}_1) = \Pr(\mathcal{M}_2) = 1/2$.

Suppose we flip the coin $n = 200$ times and observe $x = 115$ heads. Which model should we prefer in light of this data? We compute the model evidence for each model. The model evidence for $\mathcal{M}_1$ is quite straightforward, as it has no parameters:

$$\Pr(x \mid n, \mathcal{M}_1) = \mathrm{Binomial}(n, x, 1/2) = \binom{200}{115}\frac{1}{2^{200}} \approx 0.005956.$$

The model evidence for $\mathcal{M}_2$ requires integrating over the parameter θ:

$$\begin{aligned}\Pr(x \mid n, \mathcal{M}_2) &= \int \Pr(x \mid n, \theta, \mathcal{M}_2)p(\theta \mid \mathcal{M}2)\,\mathrm{d}\theta \\ &= \int_0^1 \binom{200}{115}\theta^{115}(1-\theta)^{200-115}\,\mathrm{d}\theta \\ &= \frac{1}{201} \approx 0.004975.\end{aligned}$$

The Bayes factor in favor of $\mathcal{M}_1$ is approximately 1.2, so the data give very weak evidence in favor of the simpler model $\mathcal{M}_1$.

[1] http://en.wikipedia.org/wiki/Bayes_factor#Example

An interesting aside here is that a frequentist hypothesis test would reject the null hypothesis $\theta = \frac{1}{2}$ at the $\alpha = 0.05$ level. The probability of generating at least 115 heads under model $\mathcal{M}_1$ is approximately 0.02 (similarly, the probability of generating at least 115 tails is also 0.02), so a two-sided test would give a p-value of approximately 4%.

Occam's razor

One spin on Bayesian decision theory is that it automatically gives a preference towards simpler models, in line with Occam's razor. One way to see this is to consider the model evidence $p(\mathcal{D} \mid \mathcal{M})$ as a probability distribution over datasets $\mathcal{D}$. More complex models can explain more datasets, so the support of this distribution is wider in the sample space. But note that the distribution must normalize over the sample space as well, so we pay a price for generality. When moving from a simpler model to a more complex model, the probability of some datasets that are well explained by the simpler model must inevitably decrease to "give up" probability mass for the newly explained datasets in the widened support of the more-complex model. The model selection process then drives us to select the model that is "just complex enough" to explain the data at hand, an in-build Occam's razor.

In the coin flipping example above, model $\mathcal{M}_1$ can only explain datasets with empirical heads probability reasonably near $\frac{1}{2}$. An observation of 200 heads, for example, would have astronomically small probability under this model. The second model $\mathcal{M}_2$ can explain *any* set of observations by selecting an appropriate θ. The price for this generality, though, is that datasets with a roughly equal number of heads and tails have a smaller prior probability under the model than before.

Model selection for Bayesian linear regression

A common application for model selection is for selecting between feature expansion functions $\phi(\mathbf{x})$ in Bayesian linear regression. Here the model $\mathcal{M}_i$ could for example correspond to order-i polynomial regression with

$$\phi_i(\mathbf{x}) = [1, \mathbf{x}, \mathbf{x}^2, \dots \mathbf{x}^i]^\top.$$

After selecting a set of these models to compare, as well as a prior probability for each, the only remaining task is to compute the evidence for each model in observed data $(\mathbf{X}, \mathbf{y})$. In our discussion of Bayesian linear regression, we have actually already computed the desired quantity:

$$p(\mathbf{y} \mid \mathbf{X}, \sigma^2, \mathcal{M}_i) = \mathcal{N}\big(\mathbf{y}; \phi_i(\mathbf{X})\boldsymbol{\mu}, \phi_i(\mathbf{X})\boldsymbol{\Sigma}\phi_i(\mathbf{X})^\top + \sigma^2\mathbf{I}\big),$$

where I have explicitly written the basis expansion in ϕ_i.

Note that the model ϕ_i can also easily explain all datasets well-explained by the models ϕ_j for $j < i$, by simply setting the weights on higher-order terms to zero. Again, however, the simpler model will be preferred due the Occam's razor effect described above.

Bayesian Model Averaging

Note that a "full Bayesian" treatment of a problem would eschew model selection entirely. Instead, when making predictions, we should theoretically use the sum rule to marginalize the unknown model, e.g.:

$$p(y_* \mid \mathbf{x}_*, \mathcal{D}) = \sum_i p(y_* \mid \mathbf{x}_*, \mathcal{D}, \mathcal{M}_i) \Pr(\mathcal{M}_i \mid \mathcal{D}).$$

Such an approach is called *Bayesian model averaging*. Although this is sometimes seen, model selection is still used widely in practice. The reason is that the computational overhead of using a

single model is much lower than having to continually retrain multiple models, and that Bayesian model averaging uses a mixture distribution for predictions, which can have annoying analytic properties (for example, the predictive distribution could be multimodal).