Bayesian model selection

In this course, we will learn about many kernel functions for probabilistic modeling. One major question that we will need to address is how to choose a good kernel for a given modeling task. There are some settings where prior knowledge can inform this decision but is there a more quantitative way to select which kernel function to use? Even more generally, how do I select what probabilistic model I should use to explain my data in the first place? 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**: a model is a parametric family of probability distributions, each of which which could explain some observed dataset. 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.

As an example, consider the setting of coin flipping: 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 a 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 $\{\mathcal{M}_i\}_{i=1}^n$ 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 can work out the posterior probability over the models via Bayes' theorem:

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

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 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 made the conditioning on $\mathcal{M}_i$ explicit. In the context of model selection, the term $p(\mathcal{D} \mid \mathcal{M}_i)$ is known as the **model evidence** or simply the **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.

If we have only two models for the observed data that we wish to compare, $\mathcal{M}_1$ and $\mathcal{M}_2$, it is easiest to compute the **posterior odds** or 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 (from Wikipedia's article on Bayes factor)

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 $1/2$ (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 $\mathcal{M}_1$.

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 and 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.

In the coin flipping example above, model $\mathcal{M}_1$ can only explain datasets with empirical heads probability reasonably near $1/2$. 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.

Bayesian Model Averaging

Note that a “fully Bayesian” approach to models would eschew model selection entirely. Instead, when making predictions, we should theoretically use the sum rule to marginalize the unknown model, giving rise to the model-marginal predictive distribution:

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

Such an approach is called **Bayesian model averaging.** Although this is sometimes seen, model selection is much more common because the computational overhead of using a single model is much lower than having to continually retrain multiple models. Also, the model-marginal predictive distribution tends to have annoying analytic properties, which can make it difficult to work with.