Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

26 Parametric Modeling and Model Selection

In our introduction to inference, the analysis invariably begins with some class of models of the form $p_{\mathsf{y}}(\cdot; x)$ where the parameter $x \in \mathcal{X}$ represents an index into the class. In some cases, making inferences about x is of direct interest, and y is simply the available data. In other cases, making predictions about y is of direct interest, and x is simply a latent variable.

In both of the above cases, the relevant class of models has been assumed known (and therefore given). Not surprisingly, the quality of inference that is possible ultimately depends strongly on this assumption. Thus, how we choose the class of models is important.

In this section, we consider this model selection problem. To make our development concrete, let

$$\mathcal{P}_1 \subset \mathcal{P}_2 \subset \mathcal{P}_3 \subset \cdots$$

denote a nested sequence of model classes. Specifically, for each m, the corresponding $\mathcal{P}_m$ is a set of possible distributions for y. Since the models are nested, the richness of the model class increases with m.

Example 1. Suppose we are interested in discerning the input-output relationship of some system, corresponding to regression analysis, based on the set of inputs $u_1, u_2, \ldots, u_N$ and the corresponding outputs $y_1, y_2, \ldots, y_N$. One option is to consider a simple one-parameter model class

$$\mathsf{y} = x + \mathsf{w}, \tag{1}$$

with i.i.d. Gaussian noise $\mathsf{w} \sim \mathcal{N}(0, 1)$, where $x \in \mathbb{R}$ is the parameter. In this case, there is no dependence on the input. A richer two-parameter model

$$\mathsf{y} = x_1 + x_2 \mathsf{u} + \mathsf{w}, \tag{2}$$

where again $\mathsf{w} \sim \mathcal{N}(0, 1)$ is the i.i.d. noise, introduces dependence on the inputs. Here $(x_1, x_2) \in \mathbb{R}^2$ is a parameter pair. Evidently, if we denote the first and second class of models using $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, then $\mathcal{P}_1 \subset \mathcal{P}_2$. Similarly, one can define the nested model classes $\mathcal{P}_m$ for $m > 2$ via[1]

$$\mathsf{y} = \sum_{k=1}^{m} x_k \mathsf{u}^{k-1} + \mathsf{w}.$$

In such scenarios, the sequence of possible model classes $\{\mathcal{P}_m\}_{m=1}^{M}$ often follows from basic physical considerations about the phenomena of interest. However, the selection of the model class from the sequence is often guided by the empirical evidence (the available data).

[1] Deviating from our usual use of superscript notation to indicate sequences, by u^{k-1} we mean u raised to the $k-1$st power in this case.

26.1 Bayesian Approach

The problem of model class selection is naturally viewed as one of (M-ary) hypothesis testing, where each class $\mathcal{P}_m$ is associated with a distinct hypothesis H_m. To develop the basic insights, and simplify the analysis, we adopt a Bayesian framework for the associated hypothesis testing problem. In particular, it is convenient to associate with each model class a (known) prior distribution over its constituent distributions. Furthermore, we impose a prior distribution over the possible model classes.

Using this formulation, the hypothesis H_m corresponds to the data being characterized by $p_{\mathsf{y}|\mathsf{x},H}(\cdot|\cdot, H_m)$ with $p_{\mathsf{x}|H}(\cdot|H_m)$ and $p_H(H_m)$, all of which are given. We will see later in the development that our insights are not sensitive to this specific formulation or to the form of priors we use for model selection. We should also emphasize that the parameterization under each hypothesis will, in general, be quite different (for example, involving differing numbers of parameters). Accordingly, in our notation $\mathcal{X}_m$ for the alphabet for x, we make the dependence on m explicit; e.g., $\mathcal{X}_m = \mathbb{R}^m$.

An obvious strategy for selecting the model class involves simply selecting the class that contains the distribution that maximizes the likelihood, i.e., $\hat{H} = H_{\hat{m}}$, where

$$\hat{m} = \arg\max_m \Big\{ \max_{p \in \mathcal{P}_m} p(\mathbf{y}) \Big\} = \arg\max_m \Big\{ \max_{\mathbf{a}} p_{\mathsf{y}|\mathsf{x},H}(\mathbf{y}|\mathbf{a}, H_m) \Big\}.$$

However, this kind of “generalized maximum likelihood” rule is generally not optimal. Indeed, it tends to overly favor the richer model classes (larger m). This phenomenon is commonly referred to as “overfitting”, and results in a model with poor predictive capabilities beyond the data set.

An analysis of the underlying hypothesis problem leads to a generally better selection rule. As we will see, such analysis leads to the conclusion that model selection should be guided by a quantitative version of the principle of Occam’s razor – one should always pick the simplest model consistent with the data. This principle suggests that there should be a bias toward selecting the poorer model classes (smaller m).

26.1.1 Bayes Information Criterion

Since the exact analysis is generally prohibitively complex, we instead pursue an approximate analysis that is asymptotically tight. Since we will ultimately be interested in moderately large data sets so that asymptotics can take effect, the choice of prior over the model classes will not play a strong role. For the moment we consider a scalar parameter x, assume a uniform prior over the model classes, and seek m for which

$$p_{\mathsf{y}|H}(\mathbf{y}|H_m) = \int p_{\mathsf{y}|\mathsf{x},H}(\mathbf{y}|x, H_m)\, p_{\mathsf{x}|H}(x|H_m)\, \mathrm{d}x, \tag{3}$$

is the largest among the model classes, which corresponds to the minimum probability-of-error decision rule. We refer to (3) as the *evidence* of the data $\mathbf{y}$ under the model $\mathcal{P}_m$ corresponding to hypothesis H_m.

Evaluating (3) is equivalent to evaluating a partition function associated with $p_{\mathsf{x}|\mathbf{\mathsf{y}}}(\cdot|\mathbf{y})$, which we approximate using Laplace's method. In particular, let

$$q_\circ(x) = p_{\mathbf{\mathsf{y}}|\mathsf{x},H}(\mathbf{y}|x,H)\, p_{\mathsf{x}|H}(x|H) \propto p_{\mathsf{x}|\mathbf{\mathsf{y}},H}(x|\mathbf{y},H), \tag{4}$$

then note that, via the Laplace approximation,

$$q_\circ(x) \cong q_\circ(\hat{x}) e^{-\frac{1}{2} J_{\mathbf{\mathsf{y}}=\mathbf{y}}(\hat{x})(x-\hat{x})^2}, \tag{5}$$

where

$$\hat{x} = \hat{x}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_a p_{\mathbf{\mathsf{y}}|\mathsf{x},H}(\mathbf{y}|a,H)$$

is the ML estimate of x, and $J_{\mathbf{\mathsf{y}}=\mathbf{y}}(x)$ is the observed Fisher information in $\mathbf{\mathsf{y}} = \mathbf{y}$ about x. This implies

$$\begin{aligned} p_{\mathbf{\mathsf{y}}|H}(\mathbf{y}|H) = Z_q &= \int q_\circ(x)\,\mathrm{d}x & (6)\\ &\cong q_\circ(\hat{x})\sqrt{2\pi J^{-1}_{\mathbf{\mathsf{y}}=\mathbf{y}}(\hat{x})} & (7)\\ &= p_{\mathbf{\mathsf{y}}|\mathsf{x},H}(\mathbf{y}|\hat{x},H)\, p_{\mathsf{x}|H}(\hat{x}|H)\,\sqrt{2\pi J^{-1}_{\mathbf{\mathsf{y}}=\mathbf{y}}(\hat{x})}, & (8)\end{aligned}$$

where to obtain (7) we used (5), and to obtain (8) we used (4). Thus to produce evidence $p_{\mathbf{\mathsf{y}}|H}(\mathbf{y}|H)$, the likelihood $p_{\mathbf{\mathsf{y}}|\mathsf{x},H}(\mathbf{y}|x,H)$ is scaled by the Occam factor

$$\sigma_{\hat{x}}(\mathbf{y}) = \sqrt{J^{-1}_{\mathbf{\mathsf{y}}=\mathbf{y}}(\hat{x})}. \tag{9}$$

Note that

$$J_{\mathbf{\mathsf{y}}=\mathbf{y}}(\hat{x}) = -\frac{\partial^2}{\partial x^2} \ln p_{\mathbf{\mathsf{y}}|\mathsf{x},H}(\mathbf{y}|x,H)\bigg|_{x=\hat{x}} \tag{10}$$

is a measure of how well the data is represented by the ML (i.e., asymptotically best) model in the class. So while the likelihood term in (8) favors more complex models, the Occam factor (9) favors simpler ones, and the optimum model class is therefore a compromise between these competing effects.

The above development represents an approximation, but in case of a large number N of i.i.d. observations $\mathbf{\mathsf{y}} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]$, this approximation becomes more and more accurate provided certain smoothness conditions are met. Moreover, the corresponding Occam factor $\sigma_{\hat{x}}(\mathbf{y}) = J^{-1/2}_{\mathbf{\mathsf{y}}=\mathbf{y}}(\hat{x})$ [cf. (9)] satisfies

$$\lim_{N\to\infty} \sqrt{N}\sigma_{\hat{x}}(\mathbf{y}) = \sqrt{J^{-1}_{\mathsf{y}}(\hat{x})}, \tag{11}$$

where $J_{\mathsf{y}}(x)$ is the (expected) Fisher information in any single observation y about x.

It is also straightforward to extend such analysis to the vector parameter case, i.e., when the parameter under the hypothesis of interest is a vector of dimension K_m, i.e., $\mathcal{X}_m = \mathbb{R}^{K_m}$. In particular, with

$$L_{\mathsf{y}}^N(H) \triangleq \log p_{\mathsf{y}|H}^N(\mathbf{y}|H)$$

and

$$L_{\mathsf{y}}^N(x, H) \triangleq \log p_{\mathsf{y}|\mathsf{x},H}^N(\mathbf{y}|\mathbf{x}, H)$$

we obtain [cf. (8)][2]

$$\frac{1}{N} L_{\mathsf{y}}^N(H) = \frac{1}{N} L_{\mathsf{y}}^N(\hat{\mathbf{x}}, H) + \frac{K_m}{2N} \log \frac{2\pi}{N} + \frac{1}{N} \log p_{\mathsf{x}|H}(\hat{\mathbf{x}}|H) - \frac{1}{2N} \log |\mathbf{J}_{\mathsf{y}}(\hat{\mathbf{x}})|, \tag{12}$$

where $\mathbf{J}_{\mathsf{y}}(\mathbf{x})$ is now the Fisher information matrix, and where

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\text{ML}}(\mathbf{y}) = \arg\max_{\mathbf{a}} L_{\mathsf{y}}^N(\mathbf{a}, H) \tag{13}$$

is the ML estimate. Here we have used the multidimensional Laplace approximation where

$$q_\circ(\mathbf{x}) = p_{\mathsf{y}|\mathsf{x},H}^N(\mathbf{y}|\hat{\mathbf{x}}, H)\, p_{\mathsf{x}|H}(\hat{\mathbf{x}}|H) \cong q_\circ(\hat{\mathbf{x}}) e^{-\frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}} \mathbf{J}_{\mathsf{y}=\mathbf{y}}(\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})}, \tag{14}$$

and

$$Z_q = \int q_\circ(x)\, \mathrm{d}x \cong p_{\mathsf{y}|\mathsf{x},H}^N(\mathbf{y}|\hat{\mathbf{x}}, H)\, p_{\mathsf{x}|H}(\hat{\mathbf{x}}|H)\, (2\pi)^{K/2} |\mathbf{J}_{\mathsf{y}=\mathbf{y}}(\hat{\mathbf{x}})|^{-1/2}. \tag{15}$$

Similar to the scalar case, we have

$$\frac{1}{N} \mathbf{J}_{\mathsf{y}=\mathbf{y}}(\hat{\mathbf{x}}) \cong \mathbf{J}_{\mathsf{y}}(\hat{\mathbf{x}}) \tag{16}$$

for large N.

Note that for large N, we can neglect the terms in (12) that are $O(1/N)$ since they decay faster than the others:

$$\frac{1}{N} L_{\mathsf{y}}^N(H_m) \cong \frac{1}{N} L_{\mathsf{y}}^N(\hat{\mathbf{x}}, H_m) - \frac{K_m}{2} \frac{\log N}{N}, \tag{17}$$

where K_m is the dimension of the parameter vector under H_m. This approximation is referred to as the *Bayes information criterion* (BIC) for model selection. In this case the Occam factor that modifies the likelihood function is $N^{-K_m/(2N)}$. Although we assumed a prior distribution $p_{\mathsf{x}|H}$ over the parameter space, this prior plays no role – it is one of the $O(1/N)$ terms that is negligible. Hence, in practice, we need not devise such a prior.

[2] We use $|\cdot|$ to denote the determinant operator.

Note, too, that if we had not assumed a uniform prior on the model classes, then the appropriate decision rule would be the maximum a posteriori (MAP) rule, which corresponds to choosing m such that $p_{\mathsf{H}|\mathbf{y}}(H_m|\mathbf{y})$ is largest. However, since

$$p_{\mathsf{H}||\mathbf{y}}(H_m|\mathbf{y}) \propto p_{\mathsf{H}}(H_m)\, p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_m),$$

we see that when taking logarithms and dividing by N this extension simply adds another $O(1/N)$ term—this time $(1/N)\log p_{\mathsf{H}}(H_m)$—to the objective function, which we can also neglect asymptotically. Hence, we need not devise a prior over model classes either.

Thus, in practice, we compute (17) for each m, then choose the $\mathcal{P}_m$ corresponding to the largest as the best model class. BIC is most useful in the regime where when N is only moderately large, so that the Occam factor is comparable in magnitude to the likelihood. However, when N is especially large, the Occam factor can be neglected and the test degenerates to the generalized ML rule we discussed at the outset. The performance in the limit of large N can be formally evaluated using the asymptotic analysis we developed earlier. In this case, the normalized log-evidence becomes independent of the realized data. Then, provided the true distribution lies in one of the model classes, and because the model classes are nested, the normalized log-evidence increases with m until the model order reaches a threshold value m^* that captures the true model, the remains constant for $m \geq m^*$.

In the special case when each class corresponds to a linear exponential family, i.e.,

$$\ln p_{\mathbf{y}|\mathbf{x},\mathsf{H}}(\mathbf{y}|\mathbf{x}, H) = \mathbf{x}^{\mathrm{T}}\mathbf{t}(\mathbf{y}) - \alpha(\mathbf{x}) + \beta(\mathbf{y}), \tag{18}$$

the observed Fisher information does not depend on the realization $\mathbf{y}$. Specifically,

$$\left[\mathbf{J}_{\mathbf{y}=\mathbf{y}}(\mathbf{x})\right]_{i,j} = -\frac{\partial^2 \alpha(\mathbf{x})}{\partial x_i \partial x_j} = \left[\mathbf{J}_{\mathbf{y}}(\mathbf{x})\right]_{i,j}. \tag{19}$$

In this scenario, the approximation (16) is exact even without going to the large N limit.

26.2 Other Criteria

A variety of other model selection criteria have also proved popular. One example is the minimum description length (MDL) principle, and relies on the strong connection between inference and compression. In particular, with MDL, for each model class $\mathcal{P}_m$, first let $L(H_m)$ be the number of bits required to uniquely identify a model in that class. Next, find the model in the class that when used to describe (i.e., compress) the data $\mathbf{y}$ requires the fewest number of bits. Let the resulting number of bits be $L(\mathbf{y}|\mathsf{H} = H_m)$. Finally, applying the MDL criterion corresponds to choosing the value of m for which $L(H_m) + L(\mathbf{y}|\mathsf{H} = H_m)$ is smallest.

While we will not develop MDL further in these notes, it has a long history and a large literature devoted to it. Moreover, it is useful to note that it is possible to interpret the BIC model selection rule as an approximation to the MDL criterion.

Another popular method for model selection is the Akaike information criterion (AIC). Like BIC and MDL, it attempts to mitigate overfitting effects. However, while the BIC method works well when the true distribution is in one of the model classes, AIC does not have this requirement, but rather seeks to find the best approximation for the true distribution. The approximation problem is naturally formulated as follows. Let p be the true distribution for the data $\mathbf{y}$, and let the models in $\mathcal{P}_m$ be of the form $q(\cdot;\mathbf{x},H_m)$ for some parameter vector $\mathbf{x}$ whose dimension is K_m. Then the AIC criterion corresponds to selecting the model class corresponding to the m that minimizes

$$\mathbb{E}\left[D(p(\cdot)\|q(\cdot;\hat{\mathbf{x}}(\mathbf{y}),H_m))\right] = \int p(\mathbf{y}) \int p(\mathbf{y}') \log \frac{p(\mathbf{y}')}{q(\mathbf{y}';\hat{\mathbf{x}}(\mathbf{y}),H_m)} \,\mathrm{d}\mathbf{y}' \,\mathrm{d}\mathbf{y}, \tag{20}$$

where $\hat{\mathbf{x}}(\mathbf{y})$ is the ML estimate of $\mathbf{x}$ based on the available data $\mathbf{y}$. At least in some cases of interest, when y consists of N i.i.d. observations, this criterion leads to choosing the model class that corresponds to the value of m that maximizes

$$\frac{1}{N} L_{\mathbf{y}}^{N}(\hat{\mathbf{x}},H_m) - \frac{K_m}{N}, \tag{21}$$

where

$$L_{\mathbf{y}}^{N}(\hat{\mathbf{x}},H_m) = \log q(\mathbf{y};\hat{\mathbf{x}}(\mathbf{y}),H_m).$$

In essence, richer models make the divergence in (20) smaller to the extent that the ML estimate of $\mathbf{x}$ provides the best fit within the class. The ML estimate will do this in the limit as $N \to \infty$. However, the rate at which it converges depends on the dimensionality of $\mathbf{x}$, which favors simpler models. Thus AIC balances under- and over-fitting effects. Comparing to (21), we see that in such cases, AIC does not penalize model complexity quite as strongly for moderately large values of N, and thus strikes a different balance between the opposing effects.

26.3 Further Reading

MacKay develops some nice intuition for aspects of this material. The original BIC paper is:

> G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.