A note on the evidence and Bayesian Occam's razor

Iain Murray **Zoubin Ghahramani**
Gatsby Computational Neuroscience Unit
University College London
http://www.gatsby.ucl.ac.uk/
{i.murray,zoubin}@gatsby.ucl.ac.uk

Gatsby Unit Technical Report GCNU-TR 2005-003

***Abstract*— In his thesis, MacKay (1991) introduced figure 1, explaining how Bayes rule provides an automatic "Occam's razor" effect, penalizing unnecessarily complex models. This figure has been adopted by several authors in the same schematic form. Here, after briefly reviewing necessary material, we compute a realization of the plot for a toy data modeling problem. We discuss interesting aspects of this plot and their implications for understanding model complexity.**

I. Introduction

One framework for statistical data modeling starts with writing down parametric forms for possible models, $\mathcal{H}_i$, and prior distributions over the parameters of those models, $P(\mathbf{w}|\mathcal{H}_i)$. Given these, Bayes' rule provides posterior beliefs about the models' parameters after observing data, D:

$$P(\mathbf{w}|D,\mathcal{H}_i) = \frac{P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)}{P(D|\mathcal{H}_i)}. \qquad (1)$$

Bayes' rule also provides a posterior distribution over models:

$$P(\mathcal{H}_i|D) \propto P(D|\mathcal{H}_i)P(\mathcal{H}_i), \qquad (2)$$

where the *evidence* or *marginal likelihood*, $P(D|\mathcal{H}_i) = \int P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)\mathrm{d}\mathbf{w}$ is the normalizing constant from (1).

Figure 1 illustrates the meaning of the evidence, $P(D|\mathcal{H}_i)$. Each model distributes unit probability mass over all possible data sets. When the $\mathcal{H}_i$ do not explicitly model the amount of data generated, we only consider data sets of a particular size. The D-axis has been chosen so that generally more probable, or "simple", data sets are near the center of the plot.

Simple models choose to concentrate their probability mass around a limited number of data sets. Complex models predict that data will be drawn from a large range of possibilities. If observed data can be explained well by a simple model then more complicated models, which have spent more of their available probability mass elsewhere, will be automatically penalized.

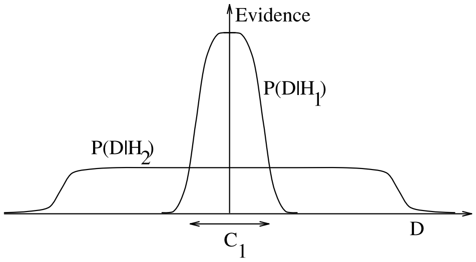


Fig. 1. This figure is reproduced with permission from MacKay (1991). It has also appeared in MacKay (1992) and MacKay (2003, chapter 28). The D-axis indexes all possible data sets (under some idealized ordering). Each curve gives a probability distribution over data sets, so must enclose an area of 1. $\mathcal{H}_1$ is a simple model focusing on data in region $\mathcal{C}_1$. Given data is this region, $\mathcal{H}_1$ has more evidence than a more powerful model $\mathcal{H}_2$, which would be favored given more complex data (outside $\mathcal{C}_1$).

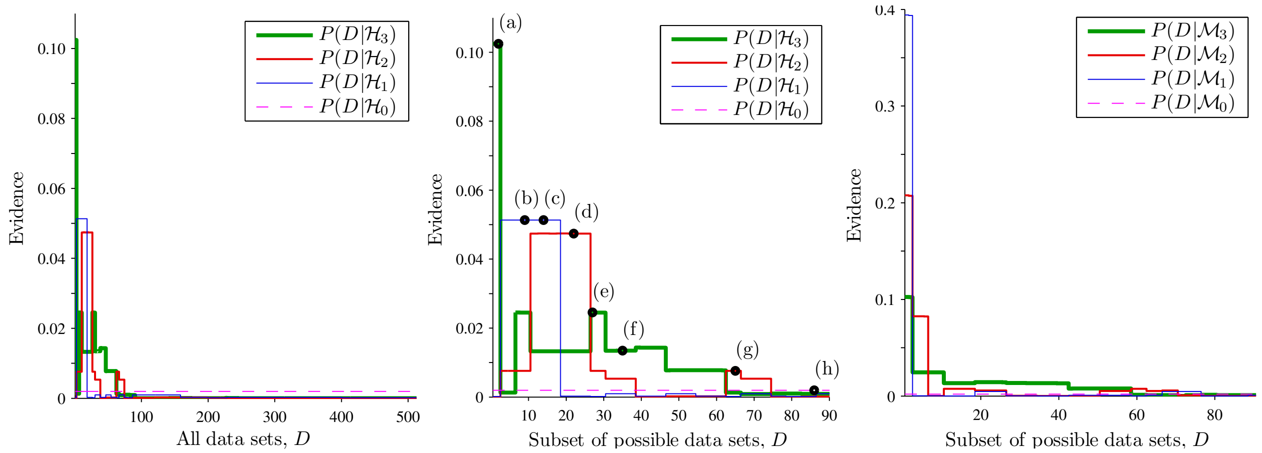
II. An explicit example

Here we consider four models for a simple data modeling problem. An observation consists of a labeling of nine binary observations $D \equiv \{y^{(n)} = \pm 1\}_{n=1}^{9}$, corresponding to a grid of known input locations:

$$\begin{array}{lll} \mathbf{x}^{(1)} = (-1,+1), & \mathbf{x}^{(2)} = (0,+1), & \mathbf{x}^{(3)} = (+1,+1) \\ \mathbf{x}^{(4)} = (-1,\ \ 0), & \mathbf{x}^{(5)} = (0,\ \ 0), & \mathbf{x}^{(6)} = (+1,\ \ 0) \\ \mathbf{x}^{(7)} = (-1,-1), & \mathbf{x}^{(8)} = (0,-1), & \mathbf{x}^{(9)} = (+1,-1)\,. \end{array} \qquad (3)$$

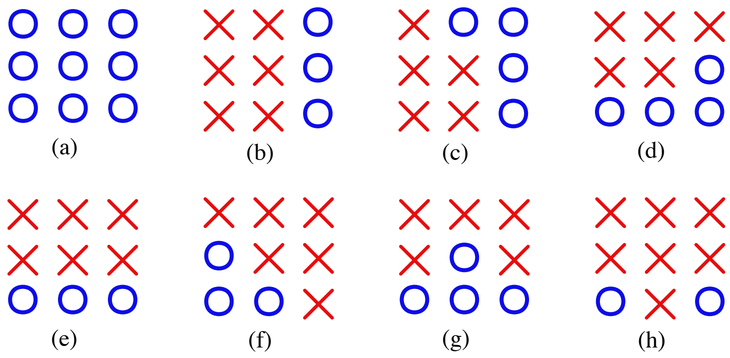
There are $2^9 = 512$ possible labelings (ie data sets D) of these nine locations. We define the following models:

$$\begin{aligned} P(D|\mathbf{w},\mathcal{H}_0) &= \frac{1}{512} \\ P(D|\mathbf{w},\mathcal{H}_1) &= \prod_{n=1}^{9} \frac{1}{1+e^{-y^{(n)}w_1x_1^{(n)}}} \\ P(D|\mathbf{w},\mathcal{H}_2) &= \prod_{n=1}^{9} \frac{1}{1+e^{-y^{(n)}\left(w_1x_1^{(n)}+w_2x_2^{(n)}\right)}} \\ P(D|\mathbf{w},\mathcal{H}_3) &= \prod_{n=1}^{9} \frac{1}{1+e^{-y^{(n)}\left(w_0+w_1x_1^{(n)}+w_2x_2^{(n)}\right)}}\,. \end{aligned} \qquad (4)$$



(i) (ii) (iii)

Fig. 2. (i) Plot of evidence for all possible data sets for the models in section II. (ii) Detail of the previous plot. Each label (a)–(h) sits on the line of the model best predicting the corresponding data set (shown in figure 3). These data sets are discussed in section III. (iii) detail of an evidence plot using an alternative set of models, see section IV.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. Labelings for the grid of inputs in equation (3) corresponding to data sets (a)–(h) labeled in figure 2(ii). Due to symmetry, the assignment of × and ○ to ±1 is arbitrary.

$\mathcal{H}_0$ treats all data sets equally; $\mathcal{H}_3$ is standard logistic regression; $\mathcal{H}_2$ is the same as $\mathcal{H}_3$ but without the bias weight w_0; $\mathcal{H}_1$ is the same as $\mathcal{H}_2$ except it ignores the second dimension of $\mathbf{x}$. Notice we have chosen the subscripts to reflect the number of parameters in each model. The prior over parameters was chosen to be $p(w_j|\mathcal{H}_i) = \mathcal{N}(0, 10^2) \quad \forall i, j$. For this toy problem, these choices are somewhat arbitrary; we chose models that would be interesting to compare.

Figure 2(i) shows the evidence $P(D|\mathcal{H}_i)$ for each model over all possible 512 data sets. The ordering of data sets was chosen heuristically (see appendix). Evidences were approximated by simple Monte Carlo; we drew $S = 10^8$ samples from the prior and computed:

$$P(D|\mathcal{H}_i) \approx \sum_{s=1}^{S} P(D|\mathbf{w}^{(s)}, \mathcal{H}_i), \quad \mathbf{w}^{(s)} \sim P(\mathbf{w}|\mathcal{H}_i). \tag{5}$$

The same samples were used for every computation. (Another interpretation is that our prior was uniform over the discrete set of 10^8 parameter vectors we considered.)

III. Discussion

Figure 2(ii) shows a detail of the whole evidence plot 2(i). The model with the largest evidence has been identified for each of eight data sets, (a)–(h), which are illustrated in figure 3. We briefly check that the most likely model for each data set make sense; then we discuss some more general properties of the models.

Data set (a) has a very unequal distribution of $+1$ and -1. Full logistic regression $\mathcal{H}_3$ is the only model to have a bias term w_0 to account for this, so it makes sense that $\mathcal{H}_3$ gives higher probability to (a) than the other models.

In data set (b) the decision boundary is a function of x_1 but not x_2. Model $\mathcal{H}_1$ is a simple model that captures such decision boundaries, so it makes sense that it gives (b) high probability. Data set (c) is equivalent to (b) under $\mathcal{H}_1$ as the different point has $x_1 = 0$, but now $\mathcal{H}_2$ becomes a close competitor. Data set (d) is equivalent to (c) under

$\mathcal{H}_2$ due to rotation invariance, whereas $\mathcal{H}_1$ cannot model decision boundaries with this orientation.

Data sets (e) and (f) are favored by $\mathcal{H}_3$ with the bias term, as w_0 allows decision boundaries to be offset from the origin. The point at the origin in data set (g) is always ignored by models without a bias weight, so here $\mathcal{H}_2$ is favored over $\mathcal{H}_3$. Finally data set (h) is not well modeled by any sharp linear boundary, in this case the uniform model $\mathcal{H}_0$ is most likely.

The above discussion assumed fairly sharp decision boundaries are typical; this results from the prior $w_i \sim \mathcal{N}(0, 10^2)$. In models $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$, large settings of the weights, $\mathbf{w}$, correspond to a sharp linear boundary in $\mathbf{x}$ space, one side of which has $y = +1$ with high probability, the other side preferring $y = -1$. The prior on the parameters has width 10, which is very vague in parameter space. However, in terms of data sets the prior is *not* vague: it puts most of its mass on settings of the parameters that give sharp linear boundaries. If the priors on parameters had had smaller widths, then the models would have become more like $\mathcal{H}_0$.

The full logistic regression model, $\mathcal{H}_3$, could be considered the most complex model of the four: it has the most parameters and can realize the other models by setting some of its parameters to zero. This means that we expect it to spread the bulk of its unit probability mass over a wider range of data sets than the other models. Figures 2(i) and 2(ii) show this intuition is correct. This flexibility comes at the expense of sometimes losing out to simpler models. Data set (b) was a good example: by moving the decision boundary from the origin some carefully chosen parameter settings of $\mathcal{H}_3$ give (b) higher probability than any settings in $\mathcal{H}_2$. However, $\mathcal{H}_2$ is more likely given (b), because the data are more typical of $\mathcal{H}_2$ than $\mathcal{H}_3$.

The simplest model in terms of parameter counting is $\mathcal{H}_0$, as it has no free parameters. The model simply defines a single distribution over data sets, assigning them all probability $1/152$. Figure 2(i) shows that this model has the largest evidence over a large range of data sets. As a result, it is unable to assign as much probability mass to "simple" data sets (a)–(g) as the other models. In some sense $\mathcal{H}_0$ is a complex model, it assigns many different types of behaviors similar probability. How can a model with no free parameters be described as complex?

We could have made other models with zero parameters. One such model assigns probability $1/8$ to each of data sets (a)–(h) and zero to all other outcomes. This model would have higher evidence than any model we have considered given any data set from figure 3. Other models with "zero parameters" assign exactly the same probability distributions as the marginal distributions, $P(D|\mathcal{H}_i)$, of models $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$. In other words, parameter counting has no real meaning.

It could be argued that $\mathcal{H}_0$ has special status as the only model to give no preference amongst possible data sets. However we might have chosen to record only how many times we observe $y = +1$. Putting a uniform prior over this quantity yields a different supposedly "assumption-free" model from $\mathcal{H}_0$. In other problems, where outcomes are continuous values from some range, the meaning of a "uniform prior over outcomes" depends not only on which quantities are measured, but also their parameterization.

IV. Alternative models

Figures 2(i), 2(ii) were qualitatively different from figure 1. In the schematic figure there was a single ordering from simple data sets out to complex: the most probable data sets according to $\mathcal{H}_1$ were also the most probable according to $\mathcal{H}_2$. The models just disagreed by how much. In our experiments the two most probable data sets according to $\mathcal{H}_3$ were amongst the very least probable data sets under $\mathcal{H}_1$ and $\mathcal{H}_2$.

The behavior we observed was due to $\mathcal{H}_3$ being the only model with a bias weight. The models were nested, in that $\mathcal{H}_i$ could set some parameters to zero to become $\mathcal{H}_j$ for $i > j$, but some basic flexibility was introduced last. An alternative hierarchy of models includes the bias weight first:

$$\begin{aligned}
P(D|\mathbf{w}, \mathcal{M}_0) &= \frac{1}{512} \\
P(D|\mathbf{w}, \mathcal{M}_1) &= \prod_{n=1}^{9} \frac{1}{1 + e^{-y^{(n)} w_0}} \\
P(D|\mathbf{w}, \mathcal{M}_2) &= \prod_{n=1}^{9} \frac{1}{1 + e^{-y^{(n)}\left(w_0 + w_1 x_1^{(n)}\right)}} \\
P(D|\mathbf{w}, \mathcal{M}_3) &= \prod_{n=1}^{9} \frac{1}{1 + e^{-y^{(n)}\left(w_0 + w_1 x_1^{(n)} + w_2 x_2^{(n)}\right)}} \,.
\end{aligned} \tag{6}$$

Results using these models are shown in figure 2(iii). This is more similar to figure 1 in that the top few most probable data sets are the same according to models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. Although there is still not one common ordering from simple data to complex data.

V. Conclusions

Some important points illustrated in this paper are:

- There is not necessarily a relationship between number of parameters and complexity.
- A hierarchy of models does not necessarily have any universal ordering from simple to complex.

- Vague priors on parameters may not be vague in terms of the predictive distributions we care about.
- Bayesian statistics makes coherent inferences from our data based on explicit modeling assumptions. There is no need for additional complexity control.

We hope working through an explicit example will help a wider audience understand the automatic Occam's razor effect in Bayesian model comparison.

Acknowledgments

We thank David MacKay for helpful comments.

References

D. J. C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1991.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. www.inference.phy.cam.ac.uk/mackay/itila/.

Appendix

For completeness we define how the data sets were ordered in figure 2. We used a greedy heuristic that populated the plots from right to left by choosing data sets without replacement from the 512 possibilities. For this algorithm distance was defined as $\sum_i (P(D|\mathcal{H}_i) - P(D'|\mathcal{H}_i))$.

Algorithm to order data sets, $D \in \mathcal{D}$ for figure 2

```
Choose data set $L = \text{argmin}_{D \in \mathcal{D}} \sum_i P(D|\mathcal{H}_i)$
Remove $L$ from $\mathcal{D}$
while ($\mathcal{D}$ is not empty)
    $\mathcal{N}$ = set of points in $\mathcal{D}$ with $L$ as nearest neighbor
    if ($\mathcal{N}$ is empty)
        Choose $L$ = nearest neighbor in $\mathcal{D}$ to $L$
    else
        Choose $L$ = furthest point from $L$ in $\mathcal{N}$
    Remove $L$ from $\mathcal{D}$
```

This heuristic procedure was just one of many possible ways of creating legible plots. We should comment that many more obvious schemes, such as sorting by the evidence of one of the models, give wildly oscillating plots. This reflects the property noted in the discussion: there is not one universal ordering of the data sets or models from simple to complex. It is therefore very fortunate the Bayesian procedure does not require us to specify such an ordering.