Bayesian Model Selection

David Barber[1]

University College London

[1] Thanks also to Taylan Cemgil for the description of Occam's razor.

Comparing Models the Bayesian Way

Given an indexed set of models $M_1, \ldots, M_m$, and associated prior beliefs in the appropriateness of each model $p(M_i)$, our interest is the model posterior probability

$$p(M_i|\mathcal{D}) = \frac{p(\mathcal{D}|M_i)p(M_i)}{p(\mathcal{D})}$$

where the likelihood of the data $\mathcal{D}$ is

$$p(\mathcal{D}) = \sum_{i=1}^{m} p(\mathcal{D}|M_i)p(M_i)$$

Model M_i is parameterised by θ_i, and the model likelihood is given by

$$p(\mathcal{D}|M_i) = \int p(\mathcal{D}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i$$

In discrete parameter spaces, the integral is replaced with summation. Note that the number of parameters $\dim(\theta_i)$ need not be the same for each model.

Bayes Factor

Comparing two competing model hypotheses M_i and M_j is straightforward and only requires the Bayes Factor:

$$\frac{p(M_i|\mathcal{D})}{p(M_j|\mathcal{D})} = \underbrace{\frac{p(\mathcal{D}|M_i)}{p(\mathcal{D}|M_j)}}_{\text{Bayes' Factor}} \frac{p(M_i)}{p(M_j)}$$

which does not require integration/summation over all possible models.

Caveat

$p(M_i|\mathcal{D})$ only refers to the probability relative to the set of models specified $M_1, \ldots, M_m$. This is not the *absolute* probability that model M fits 'well'.

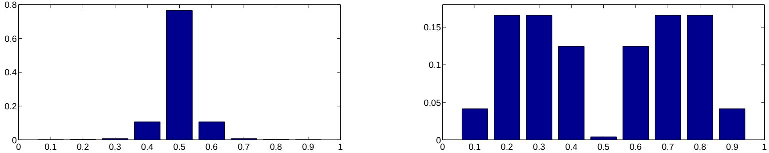
Example: Fair or Biased coin?

Two models:

$$M_{fair} : \text{The coin is fair,} \qquad M_{biased} : \text{The coin is biased}$$

For simplicity we assume $\mathrm{dom}(\theta) = \{0.1, 0.2, \ldots, 0.9\}$.

$p(\theta|M)$



(a)

(b)

Figure: **(a)**: Discrete prior model of a 'fair' coin $p(\theta|M_{fair})$. **(b)**: Prior for a biased 'unfair' coin $p(\theta|M_{biased})$. In both cases we are making explicit choices here about what we consider to be a 'fair' and 'unfair'.

Example: Fair or Biased coin?

The model likelihood

For each model M, the likelihood is given by

$$p(\mathcal{D}|M) = \sum_{\theta} p(\mathcal{D}|\theta, M) p(\theta|M) = \sum_{\theta} \theta^{N_H} \left(1-\theta\right)^{N_T} p(\theta|M)$$

This gives

$$0.1^{N_H} \left(1-0.1\right)^{N_T} p(\theta = 0.1|M) + \ldots + 0.9^{N_H} \left(1-0.9\right)^{N_T} p(\theta = 0.9|M)$$

Bayes factor

Assuming that $p(M_{fair}) = p(M_{biased})$ the Bayes' factor is given by the ratio of the two model likelihoods.

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{fair})} = \frac{p(\mathcal{D}|M_{fair})}{p(\mathcal{D}|M_{fair})}$$

Example: Fair or Biased coin?

5 Heads and 2 Tails

Here $p(\mathcal{D}|M_{fair}) = 0.00786$ and $p(\mathcal{D}|M_{biased}) = 0.0072$. The Bayes' factor is

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{biased}|\mathcal{D})} = 1.09$$

indicating that there is little to choose between the two models.

50 Heads and 20 Tails

Here $p(\mathcal{D}|M_{fair}) = 1.5 \times 10^{-20}$ and $p(\mathcal{D}|M_{biased}) = 1.4 \times 10^{-19}$. The Bayes' factor is

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{biased}|\mathcal{D})} = 0.109$$

indicating that have around 10 times the belief in the biased model as opposed to the fair model.

‘Automatic’ Complexity penalisation

Problem

You are told that the total score given from an unknown number of die is 9. What is the distribution of the number of die?

Model posterior

From Bayes’ rule, we need to compute the posterior distribution over models

$$p(n|t) = \frac{p(t|n)p(n)}{p(t)}$$

Assume $p(n) =$ const.

Likelihood

$$p(t|n) = \sum_{s_1,\ldots,s_n} p(t, s_1, \ldots, s_n|n) = \sum_{s_1,\ldots,s_n} p(t|s_1, \ldots, s_n) \prod_i p(s_i)$$

$$= \sum_{s_1,\ldots,s_n} \mathbb{I}\left[t = \sum_{i=1}^{n} s_i\right] \prod_i p(s_i)$$

where $p(s_i) = 1/6$ for all scores s_i.

‘Automatic’ Complexity penalisation

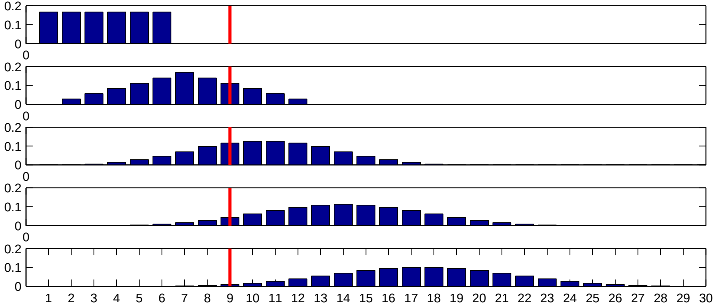
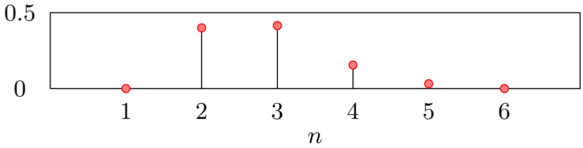


Figure: The likelihood of the total dice score, $p(t|n)$ for $n = 1$ (top) to $n = 5$ (bottom) die. Plotted along the horizontal axis is the total score t. The vertical line marks the comparison for $p(t = 9|n)$ for the different number of die. The more complex models, which can reach more states, have lower likelihood, due to normalisation over t.

‘Automatic’ Complexity penalisation

The posterior $p(n|t = 9)$



Occam’s razor

- A posteriori, there are only 3 plausible models, namely $n = 2, 3, 4$ since the rest are either too complex, or impossible.
- As the models become more ‘complex’ (n increases), more states become accessible and the probability mass typically reduces.
- This demonstrates the Occam’s razor effect which penalises models which are over complex.

Fitting models to continuous data

A model class

Consider an additive set of periodic functions

$$y^0 = w_0 + w_1 \cos(x) + w_2 \cos(2x) + \ldots + w_K \cos(Kx)$$

This can be conveniently written in vector form

$$y^0 = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(x)$$

where

$$\boldsymbol{\phi}(x) = (1, \cos(x), \cos(2x), \ldots, \cos(Kx))^\mathsf{T}$$

Data

We are given a set of data $\mathcal{D} = \{(x^n, y^n), n = 1, \ldots, N\}$ drawn from this distribution, where y is the clean $y^0(x)$ corrupted with additive zero mean Gaussian noise with variance σ^2,

$$y^n = y^0(x^n) + \epsilon^n, \quad \epsilon^n \sim \mathcal{N}\left(\epsilon^n | 0, \sigma^2\right)$$

The task

How many components K should we use?

Fitting models to continuous data

The posterior

Assuming i.i.d. data,

$$
\begin{aligned}
p(K|\mathcal{D}) &= \frac{p(\mathcal{D}|K)p(K)}{p(\mathcal{D})} \\
&= \frac{p(K)\prod_n p(x^n)}{p(\mathcal{D})} p(y^1, \dots, y^N | x^1, \dots, x^N, K)
\end{aligned}
$$

We will assume $p(K) = \text{const.}$.

The likelihood

$$
p(y^1, \dots, y^N | x^1, \dots, x^N, K) = \int_{\mathbf{w}} p(\mathbf{w}|K) \prod_{n=1}^{N} p(y^n | x^n, \mathbf{w}, K)
$$

Evaluating the likelihood

$$p(y^1, \ldots, y^N | x^1, \ldots, x^N, K) = \int_{\mathbf{w}} p(\mathbf{w}|K) \prod_{n=1}^{N} p(y^n | x^n, \mathbf{w}, K)$$

For $p(\mathbf{w}|K) = \mathcal{N}(\mathbf{w}|0, \mathbf{I}_K/\alpha)$, the integrand is a Gaussian in $\mathbf{w}$ for which it is straightforward to evaluate the integral,

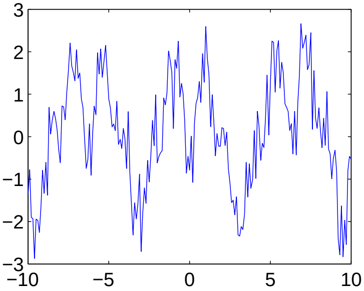
$$N \log\left(2\pi\sigma^2\right) - \sum_{n=1}^{N} \frac{(y^n)^2}{\sigma^2} + \mathbf{b}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{b} - \log\det 2\pi\mathbf{A} + K\log(2\pi\alpha)$$

where

$$\mathbf{A} \equiv \alpha\mathbf{I} + \frac{1}{\sigma^2}\sum_{n=1}^{N} \boldsymbol{\phi}(x^n)\boldsymbol{\phi}^{\mathsf{T}}(x^n), \qquad \mathbf{b} \equiv \frac{1}{\sigma^2}\sum_{n=1}^{N} y^n \boldsymbol{\phi}(x^n)$$

Example

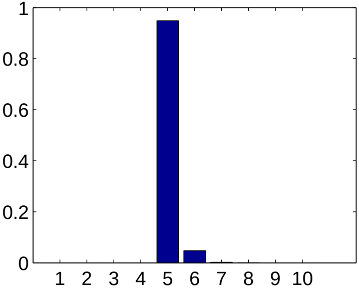
The data



Setting $\alpha = 1$ and $\sigma = 0.5$, we sampled some data from a model with $K = 5$ components.

Example

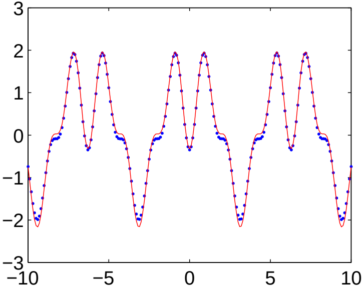
The posterior $p(K|\mathcal{D})$



We assume that we know the correct noise level σ. This is sharply peaked at $K = 5$, which is the ‘correct’ value used to generate the data.

Example

The reconstruction



The clean reconstructions for $K = 5$ are plotted. The reconstruction of the data using $\langle \mathbf{w} \rangle^{\mathsf{T}} \boldsymbol{\phi}(x)$ where $\langle \mathbf{w} \rangle$ is the mean posterior vector of the optimal dimensional model $p(\mathbf{w}|\mathcal{D}, K = 5)$. Plotted in red is the mean reconstruction. Plotted in dots is the true underlying clean data.

Approximating the Model Likelihood

The model likelihood

For a model with continuous parameter vector $\boldsymbol{\theta}$, $\dim(\boldsymbol{\theta}) = K$ and data $\mathcal{D}$, the model likelihood is

$$p(\mathcal{D}|M) = \int p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$

Need for approximations

For a generic expression

$$p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M) = e^{-f(\boldsymbol{\theta})}$$

unless f is of a particularly simple form (quadratic in $\boldsymbol{\theta}$ for example), one cannot compute the integral and approximations are required.

Laplace's method

A simple approximation is given by Laplace's method,

$$\log p(\mathcal{D}|M) \approx -f(\boldsymbol{\theta}^*) + \frac{1}{2}\log\det 2\pi\mathbf{H}^{-1}$$

where $\boldsymbol{\theta}^*$ is the MAP solution

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\; p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)$$

and $\mathbf{H}$ is the Hessian of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$.

i.i.d. data

For i.i.d. data $\mathcal{D} = \{x^1, \ldots, x^N\}$

$$p(\mathcal{D}|M) = \int p(\boldsymbol{\theta}|M)\prod_{n=1}^{N} p(x^n|\boldsymbol{\theta}, M)d\boldsymbol{\theta}$$

In this case Laplace's method computes the optimum of the function

$$-f(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|M) + \sum_{n=1}^{N}\log p(x^n|\boldsymbol{\theta}, M)$$

Bayes Information Criterion

For i.i.d. data the Hessian scales with the number of training examples, N, and a crude approximation is to set $\mathbf{H} \approx N\mathbf{I}_K$ where $K = \dim \boldsymbol{\theta}$. In this case one may take as a model comparison procedure the function

$$\log p(\mathcal{D}|M) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^*, M) + \log p(\boldsymbol{\theta}^*|M) + \frac{K}{2}\log 2\pi - \frac{K}{2}\log N$$

For a simple prior that penalises the length of the parameter vector, $p(\boldsymbol{\theta}|M) = \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}, \mathbf{I}\right)$, the above reduces to

$$\log p(\mathcal{D}|M) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^*, M) - \frac{1}{2}\left(\boldsymbol{\theta}^*\right)^{\mathsf{T}}\boldsymbol{\theta}^* - \frac{K}{2}\log N$$

The BIC approximates ignores the penalty term, giving

$$BIC = \log p(\mathcal{D}|\boldsymbol{\theta}^*, M) - \frac{K}{2}\log N$$

The term $-\frac{K}{2}\log N$ penalises model complexity. In general, the Laplace approximation is to be preferred to the BIC criterion since it more correctly accounts for the uncertainty in the posterior parameter estimate.

Outcome Analysis

Classifiers A and B have been tested on some data, so that we have, for each example in the test set, an outcome pair

$$(o_a(n), o_b(n)), n = 1, \ldots, N$$

where N is the number of test data points, and $o_a \in \{1, \ldots, Q\}$ (and similarly for o_b). That is, there are Q possible types of outcomes that can occur. For example,

$$\mathrm{dom}(o) = \{\mathsf{TruePositive}, \mathsf{FalsePositive}, \mathsf{TrueNegative}, \mathsf{FalseNegative}\}$$

We call $\mathbf{o}_a = \{o_a(n), n = 1, \ldots, N\}$, the outcomes for classifier A, and similarly for $\mathbf{o}_b = \{o_b(n), n = 1, \ldots, N\}$ for classifier B.

Hypothesis testing

1. H_{indep} : $\mathbf{o}_a$ and $\mathbf{o}_b$ are from different categorical distributions.
2. H_{same} : $\mathbf{o}_a$ and $\mathbf{o}_b$ are from the same categorical distribution.

Outcome Analysis

Posterior

$$p(H|\mathbf{o}_a, \mathbf{o}_b) = \frac{p(\mathbf{o}_a, \mathbf{o}_b|H)p(H)}{p(\mathbf{o}_a, \mathbf{o}_b)}$$

where $p(H)$ is the prior belief that H is the correct hypothesis.

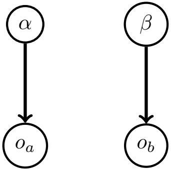
Likelihood

We make the independence of trials assumption

$$p(\mathbf{o}_a, \mathbf{o}_b|H) = \prod_{n=1}^{N} p(o_a(n), o_b(n)|H).$$

To make further progress we need to clarify the meaning of the hypotheses.

H_{indep} : model likelihood



$$p(H_{\text{indep}}|\mathbf{o}_a, \mathbf{o}_b) = \frac{p(\mathbf{o}_a, \mathbf{o}_b|H_{\text{indep}})p(H_{\text{indep}})}{p(\mathbf{o}_a, \mathbf{o}_b)}$$

The outcome model for classifier A is specified using continuous parameters, $\boldsymbol{\alpha}$, giving $p(\mathbf{o}_a|\boldsymbol{\alpha}, H_{\text{indep}})$, and similarly we use $\boldsymbol{\beta}$ for classifier B.

$$\begin{aligned} & p(\mathbf{o}_a, \mathbf{o}_b)p(H_{\text{indep}}|\mathbf{o}_a, \mathbf{o}_b) \\ & = \int p(\mathbf{o}_a, \mathbf{o}_b|\boldsymbol{\alpha}, \boldsymbol{\beta}, H_{\text{indep}})p(\boldsymbol{\alpha}, \boldsymbol{\beta}|H_{\text{indep}})p(H_{\text{indep}})d\boldsymbol{\alpha}d\boldsymbol{\beta} \\ & = p(H_{\text{indep}}) \int p(\mathbf{o}_a|\boldsymbol{\alpha}, H_{\text{indep}})p(\boldsymbol{\alpha}|H_{\text{indep}})d\boldsymbol{\alpha} \int p(\mathbf{o}_b|\boldsymbol{\beta}, H_{\text{indep}})p(\boldsymbol{\beta}|H_{\text{indep}})d\boldsymbol{\beta} \end{aligned}$$

Prior

It is convenient to use the Dirichlet prior, which is conjugate to the categorical distribution:

$$p(\boldsymbol{\alpha}|H_{\text{indep}}) = \frac{1}{Z(\mathbf{u})} \prod_q \alpha_q^{u_q - 1}, \qquad Z(\mathbf{u}) = \frac{\prod_{q=1}^{Q} \Gamma(u_q)}{\Gamma\left(\sum_{q=1}^{Q} u_q\right)}$$

The prior hyperparameter $\mathbf{u}$ controls how strongly the mass of the distribution is pushed to the corners of the simplex. Setting $u_q = 1$ for all q corresponds to a uniform prior. The likelihood of observing $\mathbf{o}_a$ is given by

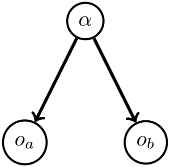
$$\int p(\mathbf{o}_a|\boldsymbol{\alpha}, H_{\text{indep}}) p(\boldsymbol{\alpha}|H_{\text{indep}}) d\boldsymbol{\alpha} = \int \prod_q \alpha_q^{\sharp_q^a} \frac{1}{Z(\mathbf{u})} \prod_q \alpha_q^{u_q - 1} d\boldsymbol{\alpha} = \frac{Z(\mathbf{u} + \sharp^a)}{Z(\mathbf{u})}$$

where $\sharp^a$ is a vector with components $\sharp_q^a$ being the number of times that variable a is in state q in the data. Hence

$$p(\mathbf{o}_a, \mathbf{o}_b) p(H_{\text{indep}}|\mathbf{o}_a, \mathbf{o}_b) = p(H_{\text{indep}}) \frac{Z(\mathbf{u} + \sharp^a)}{Z(\mathbf{u})} \frac{Z(\mathbf{u} + \sharp^b)}{Z(\mathbf{u})}$$

where $Z(\mathbf{u})$ is the Dirichlet normalisation.

H_{same} : model likelihood



In H_{same}, the hypothesis is that the outcomes for the two classifiers are generated from the same categorical distribution. Hence

$$
\begin{aligned}
& p(\mathbf{o}_a, \mathbf{o}_b) p(H_{\text{same}} | \mathbf{o}_a, \mathbf{o}_b) \\
& = p(H_{\text{same}}) \int p(\mathbf{o}_a | \boldsymbol{\alpha}, H_{\text{same}}) p(\mathbf{o}_b | \boldsymbol{\alpha}, H_{\text{same}}) p(\boldsymbol{\alpha} | H_{\text{same}}) d\boldsymbol{\alpha} \\
& = p(H_{\text{same}}) \frac{Z(\mathbf{u} + \sharp^a + \sharp^b)}{Z(\mathbf{u})}
\end{aligned}
$$

Bayes' factor

If we assume no prior preference for either hypothesis, $p(H_{\mathsf{indep}}) = p(H_{\mathsf{same}})$, then

$$\frac{p(H_{\mathsf{indep}}|\mathbf{o}_a, \mathbf{o}_b)}{p(H_{\mathsf{same}}|\mathbf{o}_a, \mathbf{o}_b)} = \frac{Z(\mathbf{u} + \sharp^a) Z(\mathbf{u} + \sharp^b)}{Z(\mathbf{u}) Z(\mathbf{u} + \sharp^a + \sharp^b)}$$

The higher this ratio is, the more likely we are to believe that the data were generated by two different categorical distributions.

Example

Two people classify the expression of each image into happy, sad or normal, using states $1, 2, 3$ respectively. Each column of the data below represents an image classed by the two people (person 1 is the top row and person 2 the second row). Are the two people essentially in agreement?

1	3	1	3	1	1	3	2	2	3	1	1	1	1	1	1	1	1	1	2
1	3	1	2	2	3	3	3	2	3	3	2	2	2	2	1	2	1	3	2

We perform a H_{indep} versus H_{same} test. From this data, the count vector for person 1 is $[13, 3, 4]$ and for person 2, $[4, 9, 7]$. Based on a flat prior for the categorical distribution and assuming no prior preference for either hypothesis, we have the Bayes' factor

$$\frac{p(\text{persons 1 and 2 classify differently})}{p(\text{persons 1 and 2 classify the same})} = \frac{Z([14, 4, 5])Z([5, 10, 8])}{Z([1, 1, 1])Z([18, 13, 12])} = 12.87$$

This is strong evidence the two people are classifying the images differently.