```
using Pkg; Pkg.activate("/Users/pavanchaggar/ResearchDocs/Presentations/model-
selection-1402")
```

```
include("functions.jl");
```

```
html"""<style>
main {
max-width: 900px;
}"""
```

present

Understanding Predictive Information Critera for Bayesian Models

Pavanjit Chaggar, December 2021
pavanjit.chaggar@maths.ox.ac.uk
@ChaggarPavan on Twitter
DPhil student at the Mathematical Institute. Supervised by Alain Goriely and Saad Jbabdi, with support from Stefano Magon and Gregory Klein at Roche.

Outline

1. Log Predictive Density as a Measure of Predictive Accuracy
2. Information Criteria
3. Example

Some definitions

Expectation

The expected value of a random variable, $\mathbf{X}$, is given by:

$$\mathbb{E}[\mathbf{X}] = \int x f(x) dx$$

The conditional expectation of a random variable, $\mathbf{X}$, w.r.t. some other random variable, $\mathbf{Y}$, is:

$$\mathbb{E}[\mathbf{X} \mid \mathbf{Y} = y] = \int x f(x \mid y) dx$$

Entropy

Given a perfect encoding scheme, the **expected** number of bits needed to encode a random variable, $\mathbf{X}$, is given by:

$$\mathbf{H}[\mathbf{X}] = \int x \log f(x) dx$$

The cross entropy between two distributions, $f(x)$ and $g(x)$ for $x \in \mathbf{X}$:

$$\begin{aligned} \mathbf{H}[f, g] &= -E[f \mid g] \\ &= \int \log g(x) f(x) dx \end{aligned}$$

It can also be reformulated as:

$$\mathbf{H}[f, g] = \mathbf{H}[f] + \mathbf{D}_{kl}[f \mid\mid g]$$

Measure of Predictive Accuracy

On what basis do we compare models?

Log Predictive Density

- Fancier name for log-likelihood, $\log p(y, \theta)$
- For model comparison, we are interested in how well the model describes the data and gneeralises. Therefore, we are not interetsed in the impact of the prior and can use the log-likelihood as opposed to the log-posterior.
- The prior is still useful for finding good maps between parameters and data.

Predictive Density

The posterior predictive density given posterior distribution $p(\theta \mid y)$ and new data point $\hat{y}_i$ is:

$$p(\hat{y_i} \mid y) = \int p(\hat{y_i} \mid \theta) p(\theta \mid y) d\theta$$

The log predictive densisty is simply the logarithm of this and is refered to as the log predictive density (lpd).

Since future data are unknown, we should define an expectation over y_i, called the expected log predictive densisty (elpd).

$$\mathbb{E}[\log p(\hat{y_i} \mid y) \mid f(\hat{y_i})] = \int \log p(\hat{y_i} \mid y) f(\hat{y_i}) d\hat{y_i}$$

Where $f(\cdot)$ is the *true* generative data distribution.

Predictive Density

$$-\mathbb{E}[\log p(\hat{y_i} \mid y) \mid f(\hat{y_i})] = -\int \log p(\hat{y_i} \mid y) f(\hat{y_i}) d\hat{y_i}$$

The negative elpd has the same form as a cross entropy:

$$\mathbf{H}[f, g] = \int \log g(x) f x dx.$$

and can therefore be interpreted as the cross-entropy between the true data generating process $f(\hat{y_i})$ and the predictive model $p(\hat{y_i} \mid y)$. Or how much information the model captures about the true generative process.

Similarly, it may be interpreted as the KL divergence between the model and the true generative process, since,

$$\mathbf{H}[f, g] = \mathbf{H}[f] + \mathbf{D}_{kl}[f \mid\mid g]$$

and $\mathbf{H}[f]$ is an unknown constant.

Predictive Density

For a new dataset, $\hat{y} = \{y_1, y_2, \ldots, y_n\}$, we can define the **pointwise) elpd, or the expected log pointwise predictive density (ellpd) as simply the sum over the elpd's

$$\mathbb{E}[\log p(\hat{y} \mid y) \mid f(\hat{y})] = \sum_{i=1}^{n} \int \log p(\hat{y_i} \mid y) f(\hat{y_i}) d\hat{y_i}$$

Predictive Density

Since we do not know what the true data generating process is, we leave it out and estimate the ellpd up to a constant and so we no longer have an expectation, just a summation over the log pointwise predictive density (lppd)

$$\mathbb{E}[\log p(\hat{y} \mid y) \mid f(\hat{y})] \approx \sum_{i=1}^{n} \int \log p(\hat{y_i} \mid \theta) p(\theta \mid y) d\theta$$

and in practice, this is computed using the posterior samples:

$$\sum_{i=1}^{n} log\left(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta_s)\right)$$

Predictive Density

- We can assess predictive accuracy as the cross entropy between the true data generating process and the predictive distribution.
- We can estimate this using the log pointwise predictive density
- This is typically a biased estimate for the predictive accuracy against future observations and thus are need of bias correction. This is usually what information criteria attempt to address.

Information Criteria

- Helps us assess predictive accuracy
- Helps us choose a model

What can we do?

- Within-sample predictive accuracy
  - Summary of predictive accuracy on the training data. Definitely biased...
- Adjusted within-sample predictive accuracy
  - Try to unbias the predictive accuracy by penalising model complexity (over-fitting).
  - AIC, BIC, DIC, WAIC.
- Out-of-sample predictive accuracy (cross validation)
  - We usually don't have out-of-sample data. validation.
  - Leave-p-out cross validation.

AIC

The goldren retriever. Simple and a bit goofy.

AIC

Recall that the elpd is the negative cross entropy between the true data generating process $f(\cdot)$ and the estimated model $g(\cdot) = p(\Theta \mid \mathbf{y})$.

$$elpd = -\mathbf{H}[f, g] = -\mathbf{H}[f] - \mathbf{D}_{kl}[f \mid\mid g]$$

Since we don't know $f(\cdot)$, we can't estimate $H[f]$. However, this is just an unknown constant. So, to maximimise the elpd we can just minimise the KL-divergence bewteen $f(\cdot)$ and $g(\cdot)$!

AIC

We can approximate the minimum KL-divergence by using the maximum likelihood estimate.

$$\hat{elpd}_{AIC} \approx min(\mathbf{D}_{kl}[f \mid\mid g]) = log(\mathbf{y} \mid \Theta_{MLE})$$

The AIC is the negative of this with a penalty for the number of parameters, multiplied by two for good measure.

$$AIC = 2k - 2log(\mathbf{y} \mid \Theta_{MLE})$$

For some family of models, we would want to choose the model with the lowest AIC score.s

BIC

A slightly more well groomed golden retriever. But still simple and goofy.

BIC

Almost identical to the AIC with one *significant* change:

$$BIC = k\log(n) - 2log(\mathbf{y} \mid \Theta_{MLE})$$

The penalty term is scaled by $\log(n)$, where n is the number of data points.

DIC

A poodle like information criteria. Generally liked, a bit of a pain sometimes.

DIC

A more Bayesian version of the same principle as AIC, but with some modifications.

- First, instead of using the MLE estimate, we use the posterior mean, $\Theta_B = \mathbb{E}[\Theta \mid \mathbf{y}]$.
- Second, there's a longer penalty term for the effective number of parameters.

$$\begin{align}
DIC &= 2p_{DIC} - 2\log p(\mathbf{y} \mid \bar{\Theta}) \\
p_{DIC} &= 2\big[\log p(\mathbf{y} \mid \bar{\Theta}) - \frac{1}{S}\sum_s \log p(y \mid \Theta_s)\big] \\
 &=2\text{var}[\log p(\mathbf{y}, \Theta) \mid p(\Theta \mid \mathbf{y})]
\end{align}$$

This is especially useful when the number of parameters is not obvious, e.g. when there are lots of covarying parameters.

```
md"
## DIC

A more Bayesian version of the same principle as AIC, but with some
modifications.

* First, instead of using the MLE estimate, we use the posterior mean,
$\Theta_B = \mathbb{E}[\Theta \mid \mathbf{y}]$.
* Second, there's a longer penalty term for the effective number of parameters.

>```math
>\begin{align}
>DIC &= 2p_{DIC} - 2\log p(\mathbf{y} \mid \bar{\Theta}) \\
>p_{DIC} &= 2\big[\log p(\mathbf{y} \mid \bar{\Theta}) - \frac{1}{S}\sum_s \log
p(y \mid \Theta_s)\big] \\
> &=2\text{var}[\log p(\mathbf{y}, \Theta) \mid p(\Theta \mid \mathbf{y})]
>\end{align}
>```
This is especially useful when the number of parameters is not obvious, e.g.
when there are lots of covarying parameters.
"
```

WAIC

WAIC

Unlike the AIC, BIC and DIC, the WAIC uses the full lppd as a measure of predictive accuracy, as oppoesed to some optimised log-likelihood.

There are two penalty terms, to be consistent with DIC and LOO-CV, the variance based measure is preferred.

$$p_{waic} = \sum_{i=1}^{n} \mathrm{V}_{s=1}^{S}[log(y_i \mid \Theta_s)]$$

This is the summed pointwise posterior variance of the log predictive density.

WAIC

Then, the WAIC is defined as:

$$\begin{aligned}
\text{WAIC} &= lppd - p_{waic} \\
&= 2\bigg[\sum_{i=1}^{n} log\bigg(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta_s)\bigg) \\
&- \sum_{i=1}^{n} \text{V}_{s=1}^{S}[log(y_i \mid \Theta_s)]\bigg]
\end{aligned}$$

```
md"
## WAIC

Then, the WAIC is defined as:

>```math
>\begin{align}
>\text{WAIC} &= lppd - p_{waic} \\
>&= 2 \bigg[\sum_{i = 1}^{n} log \bigg( \frac{1}{S} \sum_{s=1}^{S} p(y_i \mid
\theta_s) \bigg)  \\
> &- \sum_{i=1}^n \text{V}_{s=1}^{S}[log(y_i \mid \Theta_s)]\bigg]
>\end{align}
>```
"
```