Vehtari et al 2016
Practical Bayesian model evaluation using leave-one-out cross-validation and

Daniele Zago

April 11, 2021

Summary

The authors discuss a novel method for computing LOOCV and the *Widely Applicable Information Criterion* (WAIC), which uses a Pareto-smoothed importance sampling technique in order to avoid simulating MCMC samples from the posterior distribution $\vartheta|y_{-i}$ for each held-out data point y_i.

Idea: Since LOOCV can be computed from importance sampling, and the estimate is noisy, they fit a Pareto distribution to the upper tail of the distribution of the importance weights and get a more reliable estimate. This computation turns out to be very fast compared to the time required to fit the model.

1 Background

Given a posterior distribution $p(\vartheta|y)$ and posterior predictive distribution $p(\tilde{y}|y)$, the ***predictive accuracy*** is

$$\text{elpd} = \sum_{i=1}^{n} \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y)\, d\tilde{y}_i,$$

where $p_t(\cdot)$ is the true data-generating process. The Bayesian LOOCV estimate of elpd is

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^{n} \log p(y_i|y_{-i}) = \sum_{i=1}^{n} \int p(y_i|\vartheta)p(\vartheta|y_{-i}).$$

We can evaluate this estimate via *raw* importance sampling from the draws $\vartheta^1, \vartheta^2, \ldots, \vartheta^S \sim p(\vartheta|y)$ using weights $r_i^s = 1/p(y_i|\vartheta^s) \propto p(\vartheta^s|y_{-i})/p(\vartheta^s|y)$.

Proof.

$$\begin{aligned} p(\vartheta|y) &\propto p(y|\vartheta)p(\vartheta) \\ &= p(y_{-i}|\vartheta)p(y_i|\vartheta)p(\vartheta) \\ &= p(\vartheta|y_{-i})p(y_i|\vartheta), \end{aligned}$$

therefore $p(\vartheta|y)/p(y_i|\vartheta) = p(\vartheta|y_{-i})$.

$\square$

Then, the raw estimate becomes

$$\begin{aligned} p(\tilde{y}_i|y_{-i}) &\approx \frac{\sum_{s=1}^{S} r_i^s p(\tilde{y}_i|\vartheta^s)}{\sum_{s=1}^{S} r_i^s} \\ &\approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\vartheta^s)}}. \end{aligned}$$

which is prone to have a very high variance, since the importance weights can have high or infinite variance.

2 Proposal

The authors propose the following scheme called *Pareto-Smoothed Importance Sampling* (PSIS):

1. Fit a generalized Pareto distribution to the largest $M = 0.2S$ importance weights separately for each held-out y_i.
2. Replace the M largest ratios by the expected values of the order statistics

$$\tilde{w}_i^s = F^{-1}\left(\frac{z - 1/2}{M}\right), \quad z = 1, \dots, M.$$

   where F^{-1} is the inverse-cdf of the generalized Pareto distribution.

3. Truncate each vector of weights at $S^{3/4}\overline{w}_i$, where $\overline{w}_i$ is the mean of the weights, to guarantee finite variance.

The resulting weights should be better behaved than the raw importance ratios, and the estimated shape $\hat{k}$ of the Pareto can be used to assess reliability:

- $k < 1/2$: estimate converges quickly.
- $1/2 < k < 1$: estimated variance is finite bu may be large.
- $k > 1$: estimated variance is again finite but may be very large.

In general, $\hat{k} > 0.7$ for a specific y_{-i} is considered problematic and should be sampled directly, or use a more robust model.