Bayesian leave-one-out cross-validation for large data

Måns Magnusson (Aalto University)

Michael Riis Andersen (Technical University of Denmark)

Johan Jonasson (University of Gothenburg)

Aki Vehtari (Aalto University)

Motivation: Model selection for large data

- Bigger data sets and more complex models

- We still need to evaluate and compare models

- elpd_M quantifies how model M *generalizes* to unseen data $\tilde{y}_i$

$$\overline{\text{elpd}}_M = \int \log p_M(\tilde{y}_i|y) p_t(\tilde{y}_i) d\tilde{y}_i \,,$$

True data generating process

Posterior predictive distribution

Expected log predictive density

Leave-one-out cross-validation

- Basic idea: Hold out observation i and predict y_i based on y_{-i}

- Estimate elpd$_M$ using leave-one-out cross-validation (loo)

$$\overline{\text{elpd}}_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n} \log p_M(y_i|y_{-i}) = \frac{1}{n}\sum_{i=1}^{n} \log \int p_M(y_i|\theta) p_M(\theta|y_{-i}) d\theta$$

- Desirable properties
  - \+ almost unbiased for large n
  - \+ straight-forward handling of hierarchical structures

- Two major problems
  - \- Need to fit the model n times
  - \- Need to evaluate predictive densities n times

Our contributions: Method

$$\overline{\text{elpd}}_{\text{loo}} = \frac{1}{n} \sum_{i=1}^{n} \log p_M(y_i | y_{-i})$$

- We propose a fast approximation for $\overline{\text{elpd}}_{\text{loo}}$
  1. Approximate full data posterior $q_M(\theta|y)$ using Variational Bayes/Laplace
  2. Compute $p_M(y_i|y_{-i})$ using importance sampling with q_M as proposal
  3. Subsample the sum over n using the Hansen-Hurwitz estimator

- Solves both problems with leave-one-out CV
  1. Only need to fit the model once on the full data set
  2. Predictive distributions $p_M(y_i|y_{-i})$ are only needed for a small subset

Our contributions: Results

- Theoretical results (under regularity conditions)

$$\widehat{\text{elpd}}_{\text{loo}} \xrightarrow{p} \overline{\text{elpd}}_{\text{loo}} \quad \text{for} \quad n \to \infty$$

- Extensive empirical results
  1. Variational Bayes, Laplace approx., MCMC
  2. Bayesian linear regression
  3. Hierarchical models
- For more details, come see us at poster #231
- Thank you for listening!