Bayesian Model Selection - Information Criteria

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Model & Variable Selection

- Model selection is the superset of variable selection
- “The two best are not the best two” - Cover
- Multi-objective - Bias vs. Variance
- Approaches
  - Within-sample - e.g., penalty approaches and information criteria, bootstrapping, Bayes factor
  - Out-of-sample - e.g., test sets and cross-validation

Common Performance Measures

Let $\mathbf{y}$ be the vector of actual N outcomes, $\hat{\mathbf{y}}$ be the predicted vector given the data, $\mathcal{D}$, and $\boldsymbol{\theta}$ be the parameters of the likelihood.

- Mean Square Error (MSE): $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- Mean Absolute Deviation (MAD):$\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$
- Log likehood: $\sum_{i=1}^{N} log(f(y_i|\mathcal{D}, \boldsymbol{\theta}))$
- Deviance: $-2\sum_{i=1}^{N} log(f(y_i|\mathcal{D}, \boldsymbol{\theta}))$

Information Criteria

Let k be the number of parameters

- Akaike's Information Criterion (AIC) (Akaike, 1973):
  $-2\sum_{i=1}^{N} log(f(y_i|\mathcal{D},\boldsymbol{\theta})) + 2k$
- Bayesian Information Criterion (BIC) (Schwarz, 1978) :
  $-2\sum_{i=1}^{N} log(f(y_i|\mathcal{D},\boldsymbol{\theta})) + k\ log(N)$

Deviance Information Criterion

Let $\hat{\theta}_{\text{Bayes}} = E[\theta|y]$ be the posterior mean and k_{DIC} be the effective number of parameters defined as

$$k_{\text{DIC}} = 2\left(log\ p(y|\hat{\theta}_{\text{Bayes}}) - E_{\text{post}}[log\ p(y|\theta)]\right)$$

where the expectation in the second term is an average of θ over its posterior distribution. This is calculated with sampling, $\theta^s, s = 1, \ldots, S$, using

$$k_{\text{DIC}} = 2\left(log\ p(y|\hat{\theta}_{\text{Bayes}}) - \frac{1}{S}\sum_{s=1}^{S} log\ p(y|\theta^s)\right)$$

Deviance Information Criterion (DIC) (Spiegelhalter et al., 2001): $-2\sum_{i=1}^{N} log\ p(y_i|\hat{\theta}_{\text{Bayes}}) + 2k_{\text{DIC}}$

Log Pointwise Predictive Density

Let the log pointwise predictive density (LPPD) be

$$\text{LPPD} = log \prod_{i=1}^{N} p_{post}(y_i) = \sum_{i=1}^{N} log \int p(y_i|\theta)p_{post}(\theta)d\theta$$

which we compute using samples from the posterior, $p_{post}(\theta)$ and call them $\theta^s, s = 1, \ldots, S$ to obtain

$$\text{LPPD} = \sum_{i=1}^{N} log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right)$$

Widely Applicable Information Criterion

Let k_{WAIC} be the effective number of parameters defined as

$$k_{\text{WAIC}} = \sum_{i=1}^{N} \text{Var}_{post}(log\ p(y_i|\theta))$$

This is computed using the sample posterior variance, Var^S for each data point, y_i and summed over all data points:

$$k_{\text{WAIC}} = \sum_{i=1}^{N} \text{Var}^S(log\ p(y_i|\theta))$$

Widely Applicable Information Criterion (WAIC) (Watanabe, 2013): $-2LPPD + 2k_{\text{WAIC}}$

Comments in WAIC

- Shows good regularization in practice
- Easier to compute than CV
- Can estimate leave-one-out CV (LOO-CV)
- Useful for Bayesian model averging