Model Selection, Comparison, Averaging

Jeffrey Arnold

May 3, 2018

Model Averaging

Combining Models

Three methods to combine models

1. Continuous Model Expansion
2. Discrete Model Expansion
3. Bayesian Model Averaging

Continuous Model Exapansion

- Write a larger model that nests the model you are using.
- Can add either
  - more data, e.g. hierarchical model
  - more parameters
- Upside: more flexible, can use shrinkage to avoid overfitting
- Downside: increased computation

Continuous Model Expansion: Student-t

Normal distribution is Student-t with degrees of freedom ∞.

$$\mathsf{Normal}(y|\sigma, \mu) = \mathsf{StudentT}(y|\nu = \infty, \mu, \sigma)$$

Continuous Model Expansion: Regression

Special case:

$$\text{Normal}(y|\mu, \sigma)$$

General case:

$$\text{Normal}(y_i|\mu_i, \sigma_i)$$

- Model μ_i with regression
- Heteroskedasticity model for σ_i

Continuous Model Expansion: Regression

Special case: Observation i in group $k \in 1:K$,

$$\text{Normal}(y_{i,k} | \alpha + X_i \beta, \sigma)$$

General case: Different intercepts and slopes for each group.

$$\text{Normal}(y_{i,k} | \alpha_{i,k} + X \beta_{i,k}, \sigma)$$

Discrete Model Expansion (Mixture Models)

Suppose we have $\mathcal{M} = \{M_1, \ldots, M_K\}$.

$$p(y) = \sum_{k=1}^{K} \pi_k p(y|M_k)$$

- Mixture models: π_k is a parameter
- Bayesian Model Averaging: plug-in a value for π_k

Discrete Model Expansion

- Like continuous model expansion: directly estimate a meta-model.
- Unless truly "discrete" models, usually a second-best approximation to a continuous model expansion
- Can be computationally difficult, which is why BMA/model selection are used.

Bayes factors

Posterior Probability for a Model

Think of a model, M, as just another discrete parameter.

What is the posterior probability of M given data y?

$$p(M|y) = \frac{p(y|M)p(M)}{p(y)}$$

Bayes Factor

Evidence for M_2 over model M_1 is the ratio of their posterior distributions.

$$\frac{p(M_2|y)}{p(M_1|y)} = \underbrace{\frac{p(y|M_2)}{p(y|M_1)}}_{\text{Bayes Factor}} \times \frac{p(M_2)}{p(M_1)}$$

Problem: Bayes Factors Depend on Priors

$$\text{Bayes Factor}(M_2; M_1) = \frac{p(y|M_2)}{p(y|M_1)}$$

where

$$p(y|M_k) = \int p(\theta_k|M_k)p(y|\theta_k, M_k)d\theta_k$$

- **Problem:** Marginal likelihood integrates over θ!
- **Implications**:
  - Model comparison extremely sensitive to priors, in ways that posterior calculation is not.
  - Cannot use improper priors (or make adjustments)
  - Marginal likelihood hard to compute.

Bayes Factors

- Intuitive way to compare models
- Not that useful in practice; rarely used in practice
- Marginal likelihoods hard to compute
- Sensitivity to priors is **major issue**

Bayesian Model Averaging

Bayesian Model Averaging

- Given $\mathcal{M} = \{M_1, \dots, M_K\}$ models:

$$p(\theta|y) = \sum_{k=1}^{K} \left(\underbrace{p(\theta|M_k, y)}_{\text{posterior of } M_k} \times \underbrace{p(M_k|y)}_{\text{model prior}} \right)$$

- Weighted average of θ estimated for each model
- Unlike mixture model, models estimated separately, and averaging is post-hoc
- **Problem:** $p(M_k|y)$ require marginal likelihoods.

BMA in practice

- Several good implementations in R packages: BMA, BMS
- Generally focus on linear models where some shortcuts available for calculating Bayes Factors
- In linear models big problem is (intelligently) sampling the large space (2^p) of models
- Regularization, shrinkage, and sparse shrinkage models can often handle regression case better
- Calculating marginal likelihood in general case hard, use of approximation like BIC common
- Use pseudo-BMA weights based on prediction
- Theory based on $\mathcal{M}$-complete world, but that's not the case

Spaces of Models

- models being compared: $\mathcal{M} = \{M_1, \ldots, M_K\}$
- true model: $\mathcal{M}_t$
- reference model: $\mathcal{M}_r$

View	Description
$\mathcal{M}$-closed	M_t in $\mathcal{M}$
$\mathcal{M}$-open	M_t not in $\mathcal{M}$
$\mathcal{M}$-completed	M_t not in $\mathcal{M}$, but M_r is.

- prediction methods: $\mathcal{M}$-open or $\mathcal{M}$-completed
- Bayesian model averaging, Bayes factors, BIC and methods using marginal likelihoods: $\mathcal{M}$-closed

PSIS-LOO

What does PSIS-LOO do?

PSIS-LOO $=$ Pareto smoothed importance sample leave-one-out (cross-validation)

- **leave-one-out cross-validation**: that's what it's doing. LOO-CV where model trained on $n-1$ observations, and predicts the one held-out obs.
- **importance sampling**: running LOO-CV requires running the model n times. But $p(theta|y) \approx p(\theta|y_{-i})$, so use importance sampling to avoid that.
- **Pareto smoothed**: IS on it's own won't work, so we need to regularize it

What should you use?

- Use PSIS-LOO (Vehtari, Gelman, and Gabry 2015) implemented in the loo package:
  - computationally efficient
  - fully Bayesian, unlike AIC and DIC
  - perform better than WAIC
  - indicators for when it is a poor approximation (unlike AIC, DIC, and WAIC)
- if still too slow use WAIC, it's next best approximation
- No reason to use AIC or DIC ever; BIC does something different
- For observations which the PSIS-LOO has $k > 0.7$ use LOO-CV
- If too many observations fail PSIS-LOO, use k-fold CV
- If the likelihood doesn't easily partition into observations or LOO is not an appropriate prediction task, use the appropriate CV method.