Model Averaging

David Fletcher

Department of Mathematics and Statistics
University of Otago

Overview

- Why model averaging?
- Bayesian approach
- Frequentist approaches
- Model weights (frequentist)
  - Information-criterion-based
  - Bagging and stacking
  - Focussed
- Future work

Why model averaging?

- Parameter estimation often based on a single (best) model
- Model selection process ignored

- Estimates biased and uncertainty under-estimated?
- Breiman 1992: "quiet scandal"

- Model averaging can help allow for model uncertainty
- Relevant for prediction (rather than structural understanding)

Model averaging

Bayesian approach

- Focus on the probability that a model is true
- Priors and posteriors for different models
- Fully allow for all uncertainty (models and parameters)

Frequentist approaches

- Focus on improved prediction (largest model assumed true)
- Weighted mean of estimates from different models
- Confidence intervals with good coverage and width

- Stone 1974: "model mixing"
- Primarily developed in ecology, econometrics, machine learning

Frogs and Toads

Risk of a stroke (Volinsky et al 1997)

- Survival data: 4502 individuals, 23 potential risk factors
- Cox proportional hazards models

- Data randomly split into two halves
- One half for model selection and model averaging (using BIC)
- Other half classified as low, medium or high risk

Frogs and Toads

Risk of a stroke (Volinsky et al 1997)

- Survival data: 4502 individuals, 23 potential risk factors
- Cox proportional hazards models

Assigned risk group versus stroke occurrence

Risk	Model averaging		Best PMP model		Stepwise selection	
Group	Stroke	Stroke-free	Stroke	Stroke-free	Stroke	Stroke-free
Low	7	751	8	750	10	724
Medium	24	770	27	799	28	801
High	55	645	51	617	48	641

Risk of a Stroke

2^3 factorial experiment (Mead 1988)

- Eight frogs and eight toads
- Moist or dry conditions
- With or without water-balance hormone injection

- Response: % weight increase after immersion in water
- Factors: species (S), condition (C), hormone (H)

Risk of a Stroke

Source	d.f.	Sum of squares	Mean square	F-ratio	p
S	1	514.2	514.2	15.0	0.005
C	1	469.8	469.8	13.7	0.006
H	1	218.3	218.3	6.4	0.036
SC	1	39.4	39.4	1.2	0.315
SH	1	165.8	165.8	4.8	0.059
CH	1	58.1	58.1	1.7	0.229
SCH	1	43.9	43.9	1.3	0.291
Error	8	274.4	34.3		
Total	15	1783.9			

- Estimate and 95% CI for each treatment-combination
- Full model: potentially inefficient
- Best model (e.g. main effects): potentially over-optimistic

Risk of a Stroke

Model	AIC Weights
Null	0.000
S	0.000
C	0.000
H	0.000
S+C	0.006
S+H	0.001
C+H	0.001
S+C+H	0.030
S+C+SC	0.003
S+H+SH	0.001
C+H+CH	0.000
S+C+H+SC	0.019
S+C+H+SH	0.161
S+C+H+CH	0.025
S+C+H+SC+SH	0.131
S+C+H+SC+CH	0.018
S+C+H+SH+CH	0.197
S+C+H+SC+SH+CH	0.184
S+C+H+SC+SH+CH+SCH	0.222

Risk of a Stroke

- Simulations using AIC weights (Fletcher and Dillingham 2011)
- Main effects & interactions: low, med, high (relative to σ^2)

Coverage and width of 95% CI for treatment combination mean

	Mean coverage		Mean width	
Scenario	Best model	Model-averaged	Best model	Model-averaged
Low	0.82	0.95	0.52	0.70
Medium	0.91	0.94	0.81	0.89
High	0.94	0.94	0.95	0.98

- Best model: can give poor coverage
- Full model: perfect coverage but can be too wide
- Model averaging: good coverage and narrower than full model

When is model averaging useful?

- Focus on prediction, rather than description/understanding
- Add discussion of sensitivity of predictions to choice of model?

- Interpretation of parameter of interest same for all models

  Expected value of response variable
  Not usually appropriate for regression parameters

- Need to assess lack-of-fit of largest model
- Model weights might be estimated poorly

- Links with:

  Shrinkage: reduce variance by allowing small increase in bias
  Combining time series forecasts from different models
  Ensemble forecasting (e.g. earthquakes)

Bayesian approach

- Unified approach to allowing for uncertainty
- Model averaging natural

- Priors for both parameters and models
- Weighted combination of posteriors from different models
- Weight: posterior probability of that model being true

Bayesian approach

M candidate models

y = response variable

θ = parameter of interest

$$p(\theta|y) = \sum_{m=1}^{M} p(m|y)p(\theta|y,m)$$

$p(m|y)$ = posterior probability for model m

$p(\theta|y,m)$ = posterior for θ under model m

Bayesian approach

$$p(\theta|y) = \sum_{m=1}^{M} p(m|y)p(\theta|y,m)$$

$$p(m|y) \propto p(m)p(y|m)$$

$p(m)$ = prior probability for m

$p(y|m)$ = marginal likelihood for m

Bayesian approach

- Conceptual simplicity
- Natural means of allowing for uncertainty
- Summarize uncertainty via a distribution

- Ideal frequentist properties under the following assumptions:

  1. Data-generating model selected at random from our set of models, using prior model probabilities

  2. Parameter values then generated using specified prior distributions

  3. Data generated from selected model and parameter values

Bayesian approach

- Posterior model probabilities sensitive to priors for parameters
- Computational challenges (RJMCMC)
- Uniform prior for models?

- Focus on identification of true model, rather than prediction
- For $\theta = E(Y|x)$, need priors to depend on n (Yang 2005)

Frequentist approaches

- Model-averaged estimate

$$\bar{\theta} = \sum_{m=1}^{M} w_m \hat{\theta}_m$$

- $\hat{\theta}_m =$ estimate from model m
- Sampling distribution difficult to assess analytically

Information-criterion-based weights

- Most common: AIC, AICc, BIC weights

$$w_m \propto \exp\left(-\frac{1}{2}\mathrm{IC}_m\right)$$

- Motivated by approximation to posterior model probability when model prior uniform

Bagging

- Bootstrap-aggregating (bootstrap-smoothing; Efron 2014)
- Use bootstrap to mimic model selection (Breiman 1996)

- Generate B bootstrap samples (e.g. from largest model)
- Select best model each time (by AIC, for example)

$$\bar{\theta} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_{(b)}$$

- $\hat{\theta}_{(b)}$ = estimate from best model for bootstrap sample b
- Bootstrap-analogue of mean of model-averaged posterior

Stacking

- Model averaging equivalent of leave-one-out cross-validation

  Statistics: Stone 1974 (model-mixing)

  Machine Learning: Wolpert 1992 (stacking)

  Phylogenetics: van der Laan 2007, Polley 2010 (super learner)

  Econometrics: Hansen 2012 (jackknife model-averaging)

Stacking

- Model averaging equivalent of leave-one-out cross-validation
- Find weights that minimize

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \qquad \text{where} \quad \bar{y}_i = \sum_{m=1}^{M} w_m \hat{y}_{mi}$$

$\hat{y}_{mi}$ = prediction from fitting m to all data except y_i

- Constrained optimization
- Linear model for y_i in terms of predictors $\hat{y}_{1i}, \ldots, \hat{y}_{Mi}$
- Modify objective function for binomial or count data

Focussed weights

- Replace AIC by Focussed IC (Hjort et al 2003)
  - Not focussed on $\bar{\theta}$
  - Requires tuning constant
- Minimize asymptotic MSE $\left(\bar{\theta}\right)$ (e.g. Liang 2011)
  - Calculations not straightforward
  - Depends on asymptotic approximation
- Minimize bootstrap-based estimate of MSE $\left(\bar{\theta}\right)$
  - Focussed version of stacking (Fletcher 2015)
  - Does not rely on asymptotics
  - Computationally intensive

Future work

- Focussed model weights
- Bayesian version of stacking? (Monteith 2011, Kim 2012)
- Comparison of frequentist model weights
- Weights that optimize interval estimation?

Model Averaging

David Fletcher

Department of Mathematics and Statistics
University of Otago