Building Bayesian Models Through Splitting Data

Qingzhao Yu
Steven N. MacEachern
Mario Peruggia

11/22/2004

OUTLINE

- Data description
- Splitting data
- Model building
- Model updating
- Model combining
- Model comparison
- Conclusion
- Future Research

Data Description

- The Kentucky Derby is a 1.25 mile race held annually at the Churchill Downs race track in Louisville, Kentucky. The data are taken from the website:

  www.kentuckyderby.com

- The variables:

  a. Year — year of race

  b. Winner — name of the winning horse

  c. Condition — track condition

  d. Speed — speed of the winner, in feet per second

  e. Time — time of the winner, in seconds and fractional seconds

Data Description (cont.)

- Task: To provides a Bayesian model which can be used to predict the winning speeds for the races in the validation data set.

- Ground rules: Each model should produce a distribution for the winning speeds which is absolutely continuous with respect to Lebesgue measure.

Splitting Data

- The full data set consists of 108 observations from year 1896 to 2003.
- We held out the most recent 20 years of data for model validation and synthesis. The remaining 88 years were split into two sets of 44 years each, with data set 1 containing even years and data set 2 containing odd years.

Model 1

- Using data 1
- Drawing the scatter plot of year and speed, we see that speed increases at a larger rate with time first and then increases at a smaller rate later. And very obviously, that with better track condition, the speed is faster. So I add a new indicator variable I(year>=1965), and use a indicator variable for fast and good track.


F1.ps

Model 1(cont.)

- For the response variable, using the log(speed). Because speeds are always positive and the speed increases with time not in a linear way.
- The explanatory variables X including log(year), I(year>=1965), I(condition=fast), I(condition=good), and the interaction of log(year) and I(year>=1965).

Model 1(cont.)

- Assume log(speed)~N(XB,σ^2I)

$$h^{-1} = \sigma^2$$

- Use the non-informative prior distribution for B and h

$$(B,h) \sim 1/h$$

- The posterior distribution then is

$$B|h \sim N((X'X)^{-1}X'Y,\ h^{-1}(X'X)^{-1})$$

$$h \sim Gamma((n-p)/2, RSS/2)$$

Model 2

- Using data 2
- Believe that there is a superior speed that cannot be surpassed. The rate of speed increasing is influenced by the condition of the track, but the maximal speeds are the same for all the three conditions.

Model 2(cont.)

- Thus the model is built as the following:

  speed[i] ~ norm(mu[i],1/tau)

  in which

  mu[i]=(sp.m * (year[i]-C)) / (K[i] + (year[i]-C))

  where

  log(K[i]) <- alpha + beta * ind_slow[i] + gamma * ind_good[i];
- Sp.m is the superior speed the horse can reach. C is the year when the speed reaches 0. And K denotes the increasing rate of speed for each track condition.

Model 2(cont.)

- So the parameters we need to estimate here are sp.m, C, alpha, beta, gamma and tau. The variables are conditions of the tracks and years.
- Prior distributions

  sp.m ~ dnorm(55.0,25) I(0.0,);

  alpha ~ dnorm(0.0,16) I(0.0,);

  beta ~ dnorm(0.0,9) I(0.0,);

  gamma ~ dnorm(0.0,9) I(0.0,);

  C ~ dnorm(1850.0,2500) I(0.0,);

  tau ~ dgamma(0.1,0.1).
- Use Winbugs to get the posterior distributions.

Model Updating

- For model 1, update it using data set 2. Since the priors are normal and gamma distribution, we have closed forms for the posterior distributions.
- For model 2, update it sing data set1. Again using winbugs to get the posterior distributions and the Bayesian summary.

Model Combination

- Store $\theta^{(1)}$, $\theta^{(2)}$, …. , $\theta^{(N)}$ for each model.
- $\pi^*_j=1/N*f(x_1|\ \theta^{(j)})$

  $\pi_j= \pi^*_j/\Sigma\ \pi^*_j$
- $m(x_2|x_1)=\Sigma\ \pi^*_j f(x_2|\ \theta^{(2)})$

  …

  $m(x_{20}|x_1,\ldots,x_{19})$
- $m(X)= m(x_1)\ldots\ m(x_{20}|x_1,\ldots,x_{19})$
- The weights for each model is changing as we adding new observations.
- For all the 20 validation data, model1/model2=0.6482574.

Model Comparison

- We want to compare the automatic selected models with the human models.
- Using the combination of data set 1 and 2, we use the AIC and BIC rules to find the best models, comparing all the possible combinations of the variables.
- To compare the performance of the models, we use sum of square errors, absolute errors and the likelihood for the validation data.

Model Comparison (cont.)

- For model 1,we want the regression models to be hierarchical model. Then we calculate AIC and BIC for each model, and find the models with the smallest AIC and BIC. The models will be the automatic selected model.
- For model 2, we find the mles for the parameters for each different combinations of variables and calculate the AIC and BIC, then find the models.

Model Comparison (cont.)

- update one by one

	aic1		bic1		Q
lkhd for v1-20	4E-08	0.65821	1E-07	2.15025	7E-08
square loss	6.30	1.08	5.61	0.96	5.81
absolute loss	9.24	1.02	9.21	1.02	9.06
	AIC2		BIC2		M
lkhd for v1-20	1E-06	1.78E-01	2E-07	2.80064	3E-07
square loss	4.51	0.89	5.47	0.94	5.05
absolute loss	8.49	1.03	8.87	0.98	8.27
	aic1	bic1	aic2	bic2	human
lkhd for v1-20	0.30	0.8	6.80	1.1	2E-07
square loss	1.30	1.16	0.93	1.13	4.83
absolute loss	1.10	1.09	1.01	1.05	8.41

Model Comparison (cont.)

- update once for all

	aic1		bic1		Q
lkhd for v1-20	2.80E-09	0.041755	9.93E-08	1.47987	6.71E-08
square loss	8.027722	0.96	5.858469	0.70	8.35662
absolute loss	10.39115	0.99	9.415078	0.90	10.4892
	AIC2		BIC2		M
lkhd for v1-20	8.71E-07	3.15	5.66E-08	0.84397	2.77E-07
square loss	4.68699	0.75	6.211413	0.74	6.25855
absolute loss	8.281895	0.89	9.415078	0.90	9.26021
	aic1	bic1	aic2	bic2	human
lkhd for v1-20	0.02	0.58	5.07	0.329	1.7E-07
square loss	1.93	1.41	1.13	1.49	4.16
absolute loss	1.28	1.16	1.02	1.16	8.10

Model Comparison (cont.)


Updating.ps

Conclusion

- In our method, we use one of the splitting data set to build models, then using other data sets to update the models. After this, we use Bayesian model average to combine the built models.
- Our combined model is at least as good as the automatic selected model by BIC and AIC. And the performance is stable.

Future Research

- Doing more real data sets analysis to see the performance of this method.
- Using more automatic model selection methods to compare the performance of model building method.
- Build up theoretic system to systematically prove that this method can provide good models.