STA 360/602L: Module 6.5

Bayesian model selection

Dr. Olanrewaju Michael Akande

Bayesian model selection

- Now that we have a general sense of how Bayesian hypothesis works, let's get into model selection, and use some of the same ideas.
- General setting:

1. Define a list of models. That is, let Γ be a "finite" set of different possible models.
2. Each model γ is in Γ, including the "true" model. Also, let θ_γ represent the parameters in model γ.
3. Put a prior over the set Γ. Let $\Pi_\gamma = p[\gamma] = \Pr[\gamma \text{ is true}]$, for all $\gamma \in \Gamma$.

   Most common choice is the uniform prior, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, where $\#\Gamma$ is the total number of models in Γ.
4. Put a prior on the parameters in each model, that is, each $\pi(\theta_\gamma)$.
5. Compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model, and select a model based on the posterior probabilities

Bayesian model selection

- For each model $\gamma \in \Gamma$, we need to compute $\Pr[\gamma|Y]$.
- Let $p_\gamma(Y)$ denote the marginal likelihood of the data under model γ, that is, $p[Y|\gamma]$. As before,

$$\hat{\Pi}_\gamma = \Pr[\gamma|Y] = \frac{p[Y|\gamma] \cdot p[\gamma]}{\sum_{\gamma^\star \in \Gamma} p[Y|\gamma^\star] \cdot p[\gamma^\star]} = \frac{p_\gamma(Y)\Pi_\gamma}{\sum_{\gamma^\star \in \Gamma} p_{\gamma^\star}(Y)\Pi_{\gamma^\star}}$$

$$= \frac{\Pi_\gamma \cdot \left[\int_{\Theta_\gamma} p_\gamma(Y|\theta_\gamma) \cdot \pi(\theta_\gamma) \mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star \in \Gamma} \Pi_{\gamma^\star} \cdot \left[\int_{\Theta_{\gamma^\star}} p_{\gamma^\star}(Y|\theta_{\gamma^\star}) \cdot \pi(\theta_{\gamma^\star}) \mathrm{d}\theta_{\gamma^\star}\right]}.$$

- If we assume a uniform prior on Γ, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, then

$$\hat{\Pi}_\gamma = \frac{p_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} p_{\gamma^\star}(Y)}$$

$$= \frac{\left[\int_{\Theta_\gamma} p_\gamma(Y|\theta_\gamma) \cdot \pi(\theta_\gamma) \mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star \in \Gamma} \left[\int_{\Theta_{\gamma^\star}} p_{\gamma^\star}(Y|\theta_{\gamma^\star}) \cdot \pi(\theta_{\gamma^\star}) \mathrm{d}\theta_{\gamma^\star}\right]}.$$

Bayesian model selection

- How should we choose the Bayes optimal model?
- We can specify a loss function. The most natural is

$$L(\hat{\gamma}, \gamma) = \mathbf{1}(\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}),$$

that is,

1. Loss equals zero if the correct model is chosen; and
2. Loss equals one if incorrect model is chosen.

- Next, select $\hat{\gamma}$ to minimize Bayes risk. Here, Bayes risk (expected loss over posterior) is

$$R(\hat{\gamma}) = \sum_{\gamma \in \Gamma} \mathbf{1}(\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}) \cdot \hat{\Pi}_\gamma = 0 \cdot \hat{\Pi}_{\gamma_{\text{true}}} + \sum_{\gamma \neq \gamma_{\text{true}}} \hat{\Pi}_\gamma = \sum_{\gamma \neq \hat{\gamma}} \hat{\Pi}_\gamma = 1 - \hat{\Pi}_{\hat{\gamma}}$$

- To minimize $R(\hat{\gamma})$, choose $\hat{\gamma}$ such that $\hat{\Pi}_{\hat{\gamma}}$ is the largest! That is, select the model with the largest posterior probability.

Inference vs prediction

- What if the goal is prediction? Then maybe we should care more about predictive accuracy, rather than selecting specific variables.
- For predictions, we care about the posterior predictive distribution, that is

$$\begin{aligned}
p(y_{n+1}|Y=(y_1,\ldots,y_n)) &= \int_\Gamma \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\gamma,\theta_\gamma|Y)\,\mathrm{d}\theta_\gamma \mathrm{d}\gamma \\
&= \int_\Gamma \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\theta_\gamma|Y,\gamma)\cdot\Pr[\gamma|Y]\,\mathrm{d}\theta_\gamma \mathrm{d}\gamma \\
&= \sum_{\gamma\in\Gamma} \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\theta_\gamma|Y,\gamma)\cdot\hat{\Pi}_\gamma\,\mathrm{d}\theta_\gamma \\
&= \sum_{\gamma\in\Gamma} \hat{\Pi}_\gamma \cdot \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\theta_\gamma|Y,\gamma)\,\mathrm{d}\theta_\gamma \\
&= \sum_{\gamma\in\Gamma} \hat{\Pi}_\gamma \cdot p(y_{n+1}|Y,\gamma),
\end{aligned}$$

which is just averaging out the predictions from each model, over all possible models in Γ, with the posterior probability of each model, and this is known as Bayesian model averaging (BMA).

Back to Bayesian linear regression

- So what does this mean specifically in the context of linear regression?
- First, recall that for model γ, the posterior probability that the model is the right model is

$$\hat{\Pi}_\gamma = \frac{\Pi_\gamma p_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} \Pi_{\gamma^\star} p_{\gamma^\star}(Y)}.$$

- *Practical issues*
  - We need to calculate marginal likelihoods for ALL models in Γ.
  - In general for, we cannot calculate the marginal likelihoods unless we have a proper or conjugate priors.
  - For linear regression, that would mean looking to priors like Zellner's g-prior, the horseshoe prior you were introduced to in the lab, and so on.

Bayesian variable selection

- To explore Bayesian variable selection, rewrite each model $\gamma \in \Gamma$ as

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_{n \times n}).$$

- γ represents the set of predictors we want to throw into our model.
- Using the notation as before, each $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1}) \in \{0, 1\}^p$, so that the cardinality of Γ is 2^p, that is, the number of models in Γ.
- That is,
  - $\gamma_j = 1$ means the j'th predictor is included in the model, but $\gamma_j = 0$ means it is not;
  - $\boldsymbol{X}_\gamma$ is the matrix of predictors with $\gamma_j = 1$;
  - $\boldsymbol{\beta}_\gamma$ is the corresponding vector of predictors with $\gamma_j = 1$.
- Set $p_\gamma = \sum_{j=1}^{p} \gamma_j$, so that p_γ is the number of predictors included in model γ, then $\boldsymbol{X}_\gamma$ is $n \times p_\gamma$ and $\boldsymbol{\beta}_\gamma$ is $p_\gamma \times 1$.

Bayesian variable selection

- Recall that we can also write each model as

$$Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

- As an example, suppose we had data with 6 potential predictors including the intercept, so that each $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

- Then for model with $\gamma = (1, 1, 0, 0, 0, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\implies Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

  with $p_\gamma = 2$.

- Whereas for model with $\gamma = (1, 0, 0, 1, 1, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\implies Y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

  with $p_\gamma = 3$.

Bayesian variable selection

- The outline for variable selection would be as follows:

1. Write down likelihood under model γ. That is,

$$p(\boldsymbol{y}|\boldsymbol{X},\gamma,\boldsymbol{\beta}_\gamma,\sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\boldsymbol{y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)\right\}$$

2. Define a prior for γ, $\Pi_\gamma = \Pr[\gamma]$. For example, (i) uniform over all 2^p possible models, or even (ii) beta prior (since each $\gamma_j \in \{0,1\}$).

3. Put a prior on the parameters in each model. Using the g-prior, we have

$$\pi(\boldsymbol{\beta}_\gamma|\sigma^2) = \mathcal{N}_p\left(\boldsymbol{\beta}_{0\gamma} = \mathbf{0}, \Sigma_{0\gamma} = g\sigma^2\left[\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right]^{-1}\right)$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

Bayesian variable selection

- With those pieces, the conditional posteriors are straightforward.
- We can then compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model and select model with the highest posterior probability.
- We can also compute posterior $\Pr[\gamma_j = 1|Y]$, the posterior probability of including the j'the predictor, often called marginal inclusion probability (MIP), allowing for uncertainty in the other predictors.
- Also straightforward to do model averaging once we all have posterior samples.
- The Hoff book works through one example and you can find the Gibbs sampler for doing inference there. I strongly recommend you go through it carefully!
- In this course however, we will focus on using R packages for doing the same.

What's next?

Move on to the readings for the next module!