Bayesian Learning - Lecture 11

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Overview

- **Bayesian variable selection**.
- **Model checking** using posterior predictive distribution.

Bayesian variable selection

- Linear regression:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon.$$

- Which variables have **non-zero** coefficient? Example of hypotheses:

$$\begin{aligned} H_0 &: \beta_0 = \beta_1 = ... = \beta_p = 0 \\ H_1 &: \beta_1 = 0 \\ H_2 &: \beta_1 = \beta_2 = 0 \end{aligned}$$

- Introduce **variable selection indicators** $\mathcal{I} = (I_1, ..., I_p)$.
- Example: $\mathcal{I} = (1, 1, 0)$ means that $\beta_1 \neq 0$ and $\beta_2 \neq 0$, but $\beta_3 = 0$, so x_3 drops out of the model.

Bayesian variable selection, cont.

- Model inference, just crank the Bayesian machine:

$$p(\mathcal{I}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathcal{I}) \cdot p(\mathcal{I})$$

- The prior $p(\mathcal{I})$ is typically taken to be $I_1, ..., I_p|\theta \overset{iid}{\sim} Bernoulli(\theta)$.
- θ is the **prior inclusion probability**.

Bayesian variable selection, cont.

- Model inference, just crank the Bayesian machine:

$$p(\mathcal{I}|\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}|\mathbf{X},\mathcal{I}) \cdot p(\mathcal{I})$$

- The prior $p(\mathcal{I})$ is typically taken to be $I_1, ..., I_p|\theta \overset{iid}{\sim} Bernoulli(\theta)$.
- θ is the **prior inclusion probability**.

- Challenge: Computing the **marginal likelihood** for each model ($\mathcal{I}$)

$$p(\mathbf{y}|\mathbf{X},\mathcal{I}) = \int p(\mathbf{y}|\mathbf{X},\mathcal{I},\beta)p(\beta|\mathbf{X},\mathcal{I})d\beta$$

Bayesian variable selection, cont.

- Let $\beta_{\mathcal{I}}$ denote the **non-zero** coefficients under $\mathcal{I}$.
- Prior:

$$\beta_{\mathcal{I}}|\sigma^2 \sim N\left(0, \sigma^2\Omega_{\mathcal{I},0}^{-1}\right)$$
$$\sigma^2 \sim Inv - \chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Marginal likelihood**

$$p(\mathbf{y}|\mathbf{X}, \mathcal{I}) \propto \left|\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}^{-1}\right|^{-1/2} \left|\Omega_{\mathcal{I},0}\right|^{1/2} \left(\nu_0\sigma_0^2 + RSS_{\mathcal{I}}\right)^{-(\nu_0+n-1)/2}$$

where $\mathbf{X}_{\mathcal{I}}$ is the covariate matrix for the subset given by $\mathcal{I}$.

- $RSS_{\mathcal{I}}$ is (almost) the residual sum of squares under model implied by $\mathcal{I}$

$$RSS_{\mathcal{I}} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}\right)^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{y}$$

Bayesian variable selection via Gibbs sampling

- But there are 2^p model combinations to go through! Ouch!
- ... but most will have essentially zero posterior probability. Phew!

Bayesian variable selection via Gibbs sampling

- But there are 2^p model combinations to go through! Ouch!
- ... but most will have essentially zero posterior probability. Phew!

- **Simulate** from the joint posterior distribution:

$$p(\beta, \sigma^2, \mathcal{I}|\mathbf{y},\mathbf{X}) = p(\beta, \sigma^2|\mathcal{I}, \mathbf{y},\mathbf{X})p(\mathcal{I}|\mathbf{y}, \mathbf{X}).$$

- Simulate from $p(\mathcal{I}|\mathbf{y})$ using **Gibbs sampling**:
  - Draw $I_1|\mathcal{I}_{-1}, \mathbf{y}, \mathbf{X}$
  - Draw $I_2|\mathcal{I}_{-2},\mathbf{y}, \mathbf{X}$
  - ...
  - Draw $I_p|\mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$

Bayesian variable selection via Gibbs sampling

- But there are 2^p model combinations to go through! Ouch!
- ... but most will have essentially zero posterior probability. Phew!

- **Simulate** from the joint posterior distribution:

$$p(\beta, \sigma^2, \mathcal{I}|\mathbf{y}, \mathbf{X}) = p(\beta, \sigma^2|\mathcal{I}, \mathbf{y}, \mathbf{X})p(\mathcal{I}|\mathbf{y}, \mathbf{X}).$$

- Simulate from $p(\mathcal{I}|\mathbf{y})$ using **Gibbs sampling**:
  - Draw $I_1|\mathcal{I}_{-1}, \mathbf{y}, \mathbf{X}$
  - Draw $I_2|\mathcal{I}_{-2}, \mathbf{y}, \mathbf{X}$
  - ...
  - Draw $I_p|\mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$

- Only need to compute $Pr(I_i = 0|\mathcal{I}_{-i}, \mathbf{y}, \mathbf{X})$ and $Pr(I_i = 1|\mathcal{I}_{-i}, \mathbf{y}, \mathbf{X})$.
- Automatic model averaging, all in one simulation run.
- If needed, simulate from $p(\beta, \sigma^2|\mathcal{I}, \mathbf{y}, \mathbf{X})$ for each draw of $\mathcal{I}$.

Pseudo code for Bayesian variable selection

0 Initialize $\mathcal{I}^{(0)} = (I_1^{(0)}, I_2^{(0)} ..., I_p^{(0)})$

1 Simulate σ^2 and β from [Note: $\nu_n, \sigma_n^2, \mu_n, \Omega_n$ all depend on $\mathcal{I}^{(0)}$]

- $\sigma^2|\mathcal{I}^{(0)}, \mathbf{y}, \mathbf{X} \sim Inv - \chi^2\left(\nu_n, \sigma_n^2\right)$
- $\beta|\sigma^2, \mathcal{I}^{(0)}, \mathbf{y}, \mathbf{X} \sim N\left[\mu_n, \sigma^2\Omega_n^{-1}\right]$

2.1 Simulate $I_1|\mathcal{I}_{-1}, \mathbf{y}, \mathbf{X}$ by [define $\mathcal{I}_{prop}^{(0)} = (1 - I_1^{(0)}, I_2^{(0)} ..., I_p^{(0)})$]

- compute marginal likelihoods: $p(\mathbf{y}|\mathbf{X}, \mathcal{I}^{(0)})$ and $p(\mathbf{y}|\mathbf{X}, \mathcal{I}_{prop}^{(0)})$
- Simulate $I_1^{(1)} \sim Bernoulli(\kappa)$ where

$$\kappa = \frac{p(\mathbf{y}|\mathbf{X}, \mathcal{I}^{(0)}) \cdot p(\mathcal{I}^{(0)})}{p(\mathbf{y}|\mathbf{X}, \mathcal{I}^{(0)}) \cdot p(\mathcal{I}^{(0)}) + p(\mathbf{y}|\mathbf{X}, \mathcal{I}_{prop}^{(0)}) \cdot p(\mathcal{I}_{prop}^{(0)})}$$

2.2 Simulate $I_2|\mathcal{I}_{-2}, \mathbf{y}, \mathbf{X}$ as in Step 2.1, but $\mathcal{I}^{(0)} = (I_1^{(1)}, I_2^{(0)}, ..., I_p^{(0)})$

$\vdots$

2.p Simulate $I_p|\mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$ as in Step 2.1, but $\mathcal{I}^{(0)} = (I_1^{(1)}, I_2^{(1)}, ..., I_p^{(0)})$

3 Repeat Steps 1-2 many times.

Simple general Bayesian variable selection

- The previous algorithm only works when we can integrate out all the model parameters to obtain

$$p(\mathcal{I}|\mathbf{y}, \mathbf{X}) = \int p(\beta, \sigma^2, \mathcal{I}|\mathbf{y}, \mathbf{X}) d\beta d\sigma$$

- **MH** - **propose** β and $\mathcal{I}$ jointly from the proposal distribution

$$q(\beta_p|\beta_c, \mathcal{I}_p)q(\mathcal{I}_p|\mathcal{I}_c)$$

- Main difficulty: how to propose the non-zero elements in β_p?
- Simple approach:
  - Approximate posterior with all variables in the model: $\beta|\mathbf{y}, \mathbf{X} \overset{approx}{\sim} N\left[\hat{\beta}, J_{\mathbf{y}}^{-1}(\hat{\beta})\right]$
  - Propose β_p from $N\left[\hat{\beta}, J_{\mathbf{y}}^{-1}(\hat{\beta})\right]$, conditional on the zero restrictions implied by $\mathcal{I}_p$. Formulas are available.

Posterior predictive analysis

- If $p(y|\theta)$ is a 'good' model, then the data actually observed should not differ 'too much' from simulated data from $p(y|\theta)$.
- Bayesian: simulate data from the **posterior predictive distribution**:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

- Difficult to compare y and y^{rep} because of dimensionality.
- Solution: compare **low-dimensional statistic** $T(y,\theta)$ to $T(y^{rep},\theta)$.
- Evaluates the full probability model consisting of both the likelihood *and* prior distribution.

Posterior predictive analysis, cont.

- **Algorithm** for simulating from the posterior predictive density $p[T(y^{rep})|y]$:

1 Draw a $\theta^{(1)}$ from the posterior $p(\theta|y)$.

2 Simulate a data-replicate $y^{(1)}$ from $p(y^{rep}|\theta^{(1)})$.

3 Compute $T(y^{(1)})$.

4 Repeat steps 1-3 a large number of times to obtain a sample from $T(y^{rep})$.

Posterior predictive analysis, cont.

- **Algorithm** for simulating from the posterior predictive density $p[T(y^{rep})|y]$:

1. Draw a $\theta^{(1)}$ from the posterior $p(\theta|y)$.
2. Simulate a data-replicate $y^{(1)}$ from $p(y^{rep}|\theta^{(1)})$.
3. Compute $T(y^{(1)})$.
4. Repeat steps 1-3 a large number of times to obtain a sample from $T(y^{rep})$.

- We may now compare the observed statistic $T(y)$ with the distribution of $T(y^{rep})$.
- **Posterior predictive p-value**: $\Pr[T(y^{rep}) \geq T(y)]$
- Informal **graphical analysis**.

Posterior predictive analysis - Examples

- Ex. 1. Model: $y_1, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $T(y) = \max_i |y_i|$.

- Ex. 2. Assumption of no reciprocity in networks.
  $y_{ij}|\theta \overset{iid}{\sim} Bernoulli(\theta)$. $T(y) =$proportion of reciprocated node pairs.

- Ex. 3. ARIMA-process. $T(y)$ may be the autocorrelation function.

- Ex. 4. Poisson regression. $T(y)$ frequency distribution of the response counts. Proportions of zero counts.

Posterior predictive analysis - Normal model, max statistic