$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int p(D|\Theta)p(\Theta)d\Theta}$$

Bayesian Learning 732A46: Lecture 11

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

May 2016

Lecture overview

- Bayesian variable selection
- Model checking using posterior predictive distribution

Bayesian variable selection

- Like **Hypothesis testing** (**but fun**!).
- **Linear regression**:

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

- Which variables have **non-zero** coefficient? Examples 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** $\mathcal{I} = (\mathcal{I}_1, ..., \mathcal{I}_p)$.
- **Example** ($p = 3$): $\mathcal{I} = (1, 1, 0)$ means that $\beta_1 \neq 0$ and $\beta_2 \neq 0$, but $\beta_3 = 0$, so covariate x_3 drops out of the model.

Bayesian variable selection, cont.

- Crank the **Bayesian machine**:

$$p(\mathcal{I}|y) \propto p(y|\mathcal{I})p(\mathcal{I}).$$

- **Note**: A **probability distribution over models**. Model inference!
- The prior $p(\mathcal{I})$ is typically taken to be $\mathcal{I}_1, ..., \mathcal{I}_p|\theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$.
- θ is the **prior inclusion probability**.
- **Note**: This prior "shrinks" the number of "active" parameters towards $p\theta$.
- **Challenge**: Compute the **marginal likelihood** for each model $(\mathcal{I})$

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

Bayesian variable selection, cont.

- Let $\beta_{\mathcal{I}}$ denote **the subset of non-zero** coefficients under $\mathcal{I}$.
- **Conjugate prior**:

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

- **Marginal likelihood** (normal regression)

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

- $X_{\mathcal{I}}$ is the **covariate matrix** for the subset given by $\mathcal{I}$.
- $\Omega_{\mathcal{I},0}$ is (almost) the **prior precision** for the subset given by $\mathcal{I}$.
- $\mathrm{RSS}_{\mathcal{I}}$ is (almost) the **residual sum of squares** under model implied by $\mathcal{I}$

$$\mathrm{RSS}_{\mathcal{I}} = y'y - y'X_{\mathcal{I}}\left(X_{\mathcal{I}}'X_{\mathcal{I}} + \Omega_{\mathcal{I},0}\right)^{-1}X_{\mathcal{I}}'y.$$

Bayesian variable selection via Gibbs sampling

- The **posterior** of the indicators $p(\mathcal{I}|y) \propto p(y|\mathcal{I})p(\mathcal{I})$...
  - ... is **independent** of β, σ^2 [nothing but a **marginal likelihood**!]
  - ... $p(\mathcal{I}|y)$ is a **non-standard distribution**...
  - ... includes a sample space with 2^p outcomes...
  - ... but the **full conditional** of a single $\mathcal{I}_j$ has two outcomes - **Bernoulli**!
  - ... how do we simulate $p(\mathcal{I}|y)$?
  - **Gibbs sampling** to the rescue!
- But the **outcome space** is still 2^p (huge!). **Example**:

$$p = 20 \implies 2^{20} = 1,048,576 \quad \text{different models to explore...}$$

- ... Don't I have to run the sampler for a **huge number** of iterations to **converge**?
- Most of the 2^p models have **essentially zero probability**. We are saved!

The Gibbs sampler for $\mathcal{I}$ in linear regression

Gibbs sampling for $\mathcal{I}$ in **normal linear regression**

Obtain N samples from $p(\mathcal{I}|y)$ in the **linear regression** with **normal data** and **conjugate prior**.

- Set an (arbitrary) start point

$$\mathcal{I}^{(0)} = (\mathcal{I}_1^{(0)}, \mathcal{I}_2^{(0)}, \ldots, \mathcal{I}_p^{(0)}).$$

- **For** $i = 1, \ldots, N$,

  For $j = 1, \ldots, p$,

$$\mathcal{I}_j^{(i)} \sim p(\mathcal{I}_j|\mathcal{I}_{-j}, y) = \mathrm{Bernoulli}(\theta_j),$$

$$\theta_j = \frac{p\left(y|\mathcal{I}_1^{(i)}, \ldots, \mathcal{I}_j = 1, \ldots, \mathcal{I}_p^{(i-1)}\right) p(\mathcal{I}_j = 1)}{\sum_{m=0}^{1} p\left(y|\mathcal{I}_1^{(i)}, \ldots, \mathcal{I}_j = m, \ldots, \mathcal{I}_p^{(i-1)}\right) p(\mathcal{I}_j = m)}.$$

$$\mathcal{I}^{(i)} = (\mathcal{I}_1^{(i)}, \mathcal{I}_2^{(i)} \ldots, \mathcal{I}_p^{(i)})$$

The Gibbs sampler for $\mathcal{I}$ in linear regression, cont

- Now we have $\{\mathcal{I}^{(i)}\}_{i=B}^{N}$ (**discard burn-in**, always!).
- **But what about the parameters**? How do we sample β and σ^2?
- **Decompose** the joint posterior as usual

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

Sample β and σ^2 conditional on $\mathcal{I}$

- **For** $i = B, \ldots, N$,
  1. $\sigma^2|\mathcal{I}^{(i)}, y, \sim \text{Inv-}\chi^2\left(\nu_n, \sigma_n^2\right)$
  2. $\beta|\sigma^2, \mathcal{I}^{(i)}, y \sim \mathcal{N}\left(\mu_n, \sigma^2\Omega_n^{-1}\right)$
- **Note**: the **standard updates** for **linear regression with a conjugate prior** from **Lecture 5**, but $\nu_n, \sigma_n^2, \mu_n, \Omega_n$ (and $\beta_0 = 0, \Omega_0$) all depend on $\mathcal{I}^{(i)}$. **For example**:

$$\mu_n = (X'_{\mathcal{I}^{(i)}}X_{\mathcal{I}^{(i)}} + \Omega_{0,\mathcal{I}^{(i)}})^{-1}X'_{\mathcal{I}^{(i)}}y.$$

- **Note**: **Automatic model averaging** by integrating (by simulation) out the indicators!

General Bayesian variable selection

- The previous algorithm worked because the **marginal likelihood**

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

  was **analytically tractable** [normal data and choice of prior].
- **Bayesian variable selection** by **Metropolis-Hastings**: Markov chain in space $(\beta, \mathcal{I})$ to sample $p(\beta, \mathcal{I}|y)$
- **Note**: β contains regression coefficients + other unknowns.
- **Proposal for MH** - **propose** β and $\mathcal{I}$ jointly from

$$q(\beta_p, \mathcal{I}_p|\beta_c, \mathcal{I}_c) = q_2(\beta_p|\beta_c, \mathcal{I}_p) q_1(\mathcal{I}_p|\mathcal{I}_c).$$

- **Main difficulty**: how to propose the **non-zero elements** in β_p?
- **Simple approaches**:
  1. **Approximate posterior** with all variables in the model:

$$\beta|y \overset{approx}{\sim} \mathcal{N}\left(\beta^\star, J^{-1}_{\beta^\star, y}\right).$$

  2. Propose as in 1. but **conditional on the zero restrictions** implied by $\mathcal{I}_p$. Formulas are available (conditional of a multivariate normal is also normal).

Evaluating models by posterior predictive analysis

- **Idea**: 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** (the joy of averaging!): simulate data from

$$p(y^{\mathrm{rep}}|y) = \int p(y^{\mathrm{rep}}|\theta)p(\theta|y)d\theta \quad \text{[\textbf{Posterior predictive}]}.$$

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

Evaluating models by posterior predictive analysis, cont.

Simulate from the posterior predictive density $p\left(T(y^{\text{rep}})|y\right)$

Obtain N samples from $p\left(T(y^{\text{rep}})|y\right)$.

- **For** $i = 1, \ldots, N$,
  1. Simulate a **parameter** $\theta^{(i)} \sim p(\theta|y)$.
  2. Simulate a **data-replicate** $y^{(i)}$ from $p(y^{\text{rep}}|\theta^{(i)})$.
  3. $T^{(i)} = T(y^{(i)})$.

- Compare the **observed statistic** $T(y)$ with the distribution of $T(y^{rep})$ from our **simulation**.
- **Posterior predictive p-value**:

$$\Pr\left(T(y^{rep}) \geq T(y)\right).$$

- Informal **graphical analysis**.

Posterior predictive analysis - Examples

- **Example 1**: **Normal model**: $y_1, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $T(y) = \max_i |y_i|$.
- **Example 2**: **ARIMA-process**. $T(y)$ may be the **autocorrelation function**.
- **Example 3**: **Poisson regression**. $T(y)$ frequency distribution of the **response counts**. Or proportions of **zero counts**.

Posterior predictive analysis - Normal model, max statistic