Bayesian Learning - Lecture 5

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Lecture overview

- Normal model with conjugate prior
- The linear regression model
- Regression with binary response

Normal model - conjugate prior

- Model

$$y_1, ..., y_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- Conjugate prior

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

Normal model with conjugate prior, cont.

- Posterior

$$\theta|y,\sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|y \sim Inv\text{-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

- Marginal posterior

$$\theta \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$$

The Linear Regression Model

- The ordinary linear regression model:

$$
\begin{aligned}
y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \\
& \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).
\end{aligned}
$$

- Parameters $\theta = (\beta_1, \beta_2, \ldots, \beta_k, \sigma^2)$.
- Assumptions:
  - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ (linear function)
  - $Var(y_i) = \sigma^2$ (homoscedasticity)
  - $Corr(y_i, y_j | X) = 0$, $i \neq j$.
  - Normality of ε_i.

Linear regression in matrix form

- The linear regression model in matrix form

$$\underset{(n\times 1)}{y} = \underset{(n\times k)(k\times 1)}{X\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Usually $x_{i1} = 1$, for all i. β_1 is the intercept.
- Likelihood for the full sample

$$y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I_n)$$

Posterior for the uniform prior

- Standard non-informative prior: uniform on $(\beta, \log \sigma^2)$

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

- Joint posterior of β and σ^2:

$$p(\beta, \sigma^2 | y) = p(\beta | \sigma^2, y) p(\sigma^2 | y).$$

- Conditional posterior of β :

$$\begin{aligned} \beta | \sigma^2, y &\sim N\left[\hat{\beta}, \sigma^2 (X'X)^{-1}\right] \\ \hat{\beta} &= (X'X)^{-1} X'y \end{aligned}$$

- Marginal posterior of σ^2 :

$$\begin{aligned} \sigma^2 | y &\sim Inv\text{-}\chi^2(n-k, s^2) \\ s^2 &= \frac{1}{n-k} (y - X\hat{\beta})'(y - X\hat{\beta}). \end{aligned}$$

Posterior for the uniform prior, cont.

- Marginal posterior of β :

$$\beta|y \sim t_{n-k}\left[\hat{\beta}, s^2(X'X)^{-1}\right]$$

  which is proper if $n > k$ and X has full column rank.

- Simulate from the joint posterior by iteratively simulating from $p(\sigma^2|y)$ and $p(\beta|\sigma^2, y)$.
- Predictive distribution of response $\tilde{y}$ with known predictors $\tilde{x}$

$$\tilde{y}|y, \tilde{x} \sim t_{n-k}\left[\tilde{x}'\hat{\beta}, s^2(1 + \tilde{x}'(X'X)\tilde{x})^{-1}\right]$$

$$\begin{aligned} \text{Predictive Precision} &= s^{-2} + \tilde{x}'(s^{-2}X'X)\tilde{x} \\ &= \varepsilon\text{-Precision} + \tilde{x}'(\text{Posterior Precision of } \beta)\tilde{x}. \end{aligned}$$

Linear regression - conjugate prior

- Joint prior for β and σ^2

$$
\begin{aligned}
\beta|\sigma^2 &\sim N\left(\mu_0, \sigma^2\Omega_0^{-1}\right) \\
\sigma^2 &\sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)
\end{aligned}
$$

- Posterior

$$
\begin{aligned}
\beta|\sigma^2 &\sim N\left[\mu_n, \sigma^2\Omega_n^{-1}\right] \\
\sigma^2 &\sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_n &= \left(X'X + \Omega_0\right)^{-1}\left(X'X\hat{\beta} + \Omega_0\mu_0\right) \\
\Omega_n &= X'X + \Omega_0 \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + \left(y'y + \mu_0'\Omega_0\mu_0 - \mu_n'\Omega_n\mu_n\right)
\end{aligned}
$$

Regression with binary response

- Response is assumed to be **binary** (0-1).
- Example: Predicting whether or not an e-mail is good ($y = 1$) or spam ($y = 0$). Covariates: mean word length, proportion of \$-symbols.
- **Logistic regression**

  $$\Pr(y_i = 1 \mid x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

  Likelihood:

  $$p(y|X, \beta) = \prod_{i=1}^{n} \frac{[\exp(x_i'\beta)]^{y_i}}{1 + \exp(x_i'\beta)}.$$

  Posterior is non-standard, but in most situation can be approximated well by a normal distribution. Optimization.
- Alternative: **Probit regression**

  $$Pr(y_i = 1|x_i) = \Phi(x_i^{'}\beta)$$

Asymptotic posterior - Heuristics

- Taylor expansion of log-posterior around the posterior mode $\theta = \tilde{\theta}$:

$$\ln p(\theta|y) = \ln p(\tilde{\theta}|y) + \frac{\partial \ln p(\theta|y)}{\partial \theta}|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta})$$
$$+ \frac{1}{2!}\frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2}|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta})^2 + \ldots$$

- From the definition of the posterior mode:

$$\frac{\partial \ln p(\theta|y)}{\partial \theta}|_{\theta=\tilde{\theta}} = 0$$

- So, in large samples (where we can ignore higher order terms):

$$\ln p(\theta|y) \approx \ln p(\tilde{\theta}|y) - \frac{1}{2}H_{\mathbf{y}}(\tilde{\theta})(\theta - \tilde{\theta})^2$$

where $H_{\mathbf{y}}(\tilde{\theta}) = -\frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2}|_{\theta=\tilde{\theta}}$.

- Approximate posterior

$$\theta|y \sim N\left[\tilde{\theta}, H_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$$

Normal approximation of posterior

- If posterior is approximately normal, sufficient to find the posterior mode and (inverse) information matrix.
- Standard (e.g. gradient-based) optimization routines may be used. (optim.r). Input: an expression proportional to $p(\theta|y)$ and initial values. Output: optimum (posterior) mode and Hessian matrix (minus observed information).
- Joint posterior $p(\theta_1, \theta_2|y)$ may not be be close to normal, but perhaps $p(\theta_2|\theta_1, y)$ and $p(\theta_2|y)$ are.
- Even if the posterior of θ is approx normal, interesting functions of θ may not be (e.g. predictions). Still need to resort to numerical methods.
- Re-parametrization $\phi = g(\theta)$ may improve normal approximation. If $\theta \geq 0$ use logs. If $0 \leq \theta \leq 1$, use $\mathrm{Logit}(\theta) = \ln[\theta/(1-\theta)]$.
- Posterior mode and inverse Hessian can be used to approximate the posterior with a student-t density.