Bayesian Learning - Lecture 5

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Normal model with conjugate prior
- The linear regression model
- Non-linear regression
- Regularization priors

Normal model - normal 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 normal prior

- 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:

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

- Parameters $\theta = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)$.
- Assumptions:
  - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ (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

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

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

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

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

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

Linear regression - 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:

$$\begin{aligned} \beta|\sigma^2, \mathbf{y} &\sim N\left[\hat{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right] \\ \sigma^2|\mathbf{y} &\sim \mathit{Inv}\text{-}\chi^2(n-k, s^2) \end{aligned}$$

where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$.

- Simulate from the joint posterior by iteratively simulating from
  - $p(\sigma^2|y)$
  - $p(\beta|\sigma^2, y)$
- Marginal posterior of β :

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

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

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

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

Polynomial regression

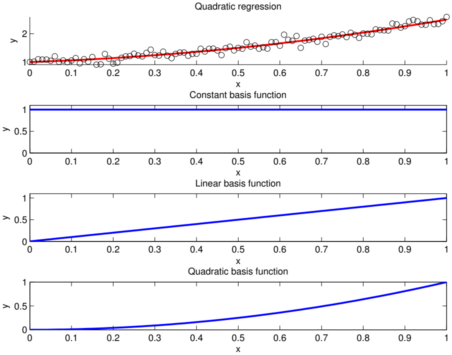
- **Polynomial regression**

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k.$$

$$\mathbf{y} = \mathbf{X}_P \beta + \varepsilon,$$

where

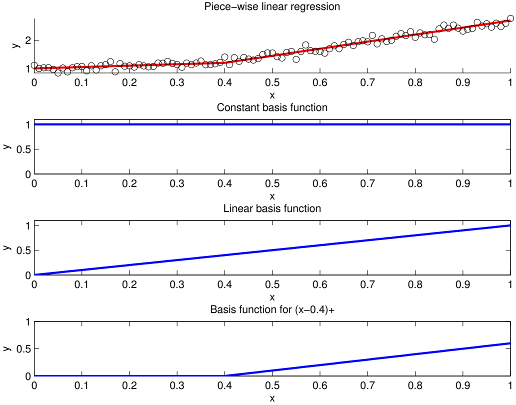
$$\mathbf{X}_P = (1, x, x^2, \ldots, x^k).$$



Spline regression

- Polynomials are too global. Need more local basis functions.
- ***Truncated power splines* given *knot* locations** $k_1, ..., k_m$

$$b_{ij} = \begin{cases} (x_i - k_j)^p & \text{if } x_i > k_j \\ 0 \text{ otherwise} \end{cases}$$



Splines, cont.

- Note: given the knots, the non-parametric spline regression model is a linear regression of y on the m 'dummy variables' b_j

$$y = \mathbf{X}_b \beta + \varepsilon,$$

where X_b is the basis regression matrix

$$\mathbf{X}_b = (b_1, ..., b_m).$$

- It is also common to include an intercept and the linear part of the model separately. In this case we have

$$\mathbf{X}_b = (1, x, b_1, ..., b_m).$$

Smoothness prior for splines

- Problem: too many knots leads to **over-fitting**.
- Solution: **smoothness/shrinkage/regularization prior**

$$\beta_i|\sigma^2 \overset{iid}{\sim} N\left(0, \frac{\sigma^2}{\lambda}\right)$$

- Larger λ gives smoother fit. Note: here we have $\Omega_0 = \lambda I$.
- Equivalent to a penalized likelihood:

$$-2 \cdot \log p(\beta|\sigma^2, \mathbf{y}, \mathbf{X}) \propto RSS(\beta) + \lambda\beta'\beta$$

- Posterior mean gives **ridge regression** estimator

$$\tilde{\beta} = \left(\mathbf{X}'\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}'\mathbf{y}$$

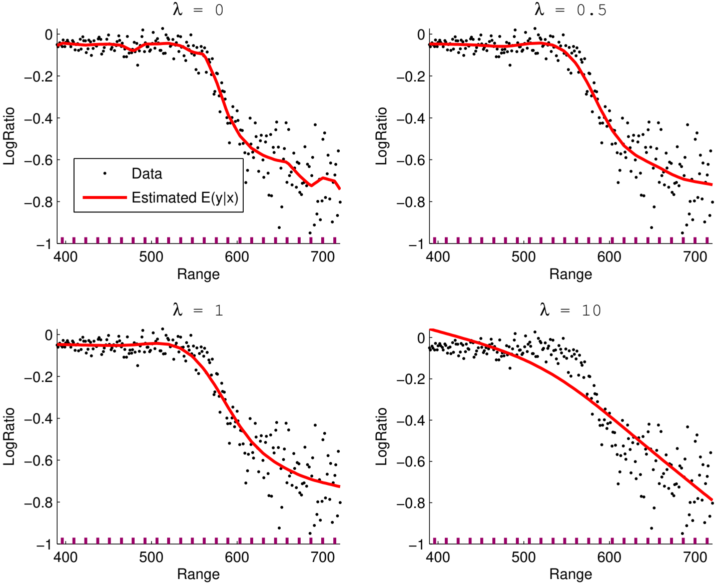
- **Shrinkage** toward zero

$$\text{As } \lambda \to \infty,\ \tilde{\beta} \to 0$$

- When $\mathbf{X}'\mathbf{X} = I$

$$\tilde{\beta} = \frac{1}{1+\lambda}\hat{\beta}_{OLS}$$

Bayesian spline with smoothness prior



Smoothness prior for splines, cont.

- The famous **Lasso** variable selection method is equivalent to using the posterior mode estimate under the prior:

$$\beta_i|\sigma^2 \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right)$$

with density

$$p(\beta_i) = \frac{\lambda}{2\sigma^2} \exp\left(-\frac{\lambda\,|\beta_i|}{\sigma^2}\right)$$

- The Bayesian shrinkage prior is **interpretable**. **Not ad hoc**.
- Laplace distribution have heavy tails.
- Laplace: many β_i are close to zero, but some β_i may be very large.
- Normal distribution have light tails.
- Normal prior: most β_i are fairly equal in size, and no single β_i can be very much larger than the other ones.

Estimating the shrinkage

- How do we determine the degree of smoothness, λ? Cross-validation is one possible approach.
- Bayesian: λ is unknown $\Rightarrow$ use a prior for λ.
- One possibility: $\lambda \sim Inv-\chi^2(\eta_0, \lambda_0)$. The user specifies η_0 and λ_0.
- Alternative approach: specify the prior on the *degrees of freedom*.
- Hierarchical setup:

$$\begin{aligned}
\mathbf{y}|\beta, \mathbf{X} &\sim N(\mathbf{X}\beta, \sigma^2 I_n) \\
\beta|\sigma^2, \lambda &\sim N\left(0, \sigma^2\lambda^{-1} I_m\right) \\
\sigma^2 &\sim Inv-\chi^2(\nu_0, \sigma_0^2) \\
\lambda &\sim Inv-\chi^2(\eta_0, \lambda_0)
\end{aligned}$$

Regression with estimated shrinkage

- The joint posterior of β, σ^2 and λ is

$$
\begin{aligned}
\beta|\sigma^2,\lambda,y &\sim N\left(\mu_n,\Omega_n^{-1}\right) \\
\sigma^2|\lambda,y &\sim Inv-\chi^2\left(\nu_n,\sigma_n^2\right) \\
p(\lambda|y) &\propto \sqrt{\frac{|\Omega_0|}{|X'X+\Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2}\cdot p(\lambda)
\end{aligned}
$$

where $p(\lambda)$ is the prior for λ, and

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

More complexity

- The **location of the knots** can be treated as unknown, and estimated from the data. Joint posterior

$$p(\beta, \sigma^2, \lambda, k_1, ..., k_m | \mathbf{y}, \mathbf{X})$$

- The marginal posterior for $\lambda, k_1, ..., k_m$ is a nightmare.
- MCMC can be used to simulate from the joint posterior. Li and Villani (2013, SJS).
- The basic spline model can be extended with:
  - **Heteroscedastic errors** (also modelled with a spline)
  - **Non-normal errors** (student-t or mixture distributions)
  - **Autocorrelated/dependent errors** (AR process for the error term)
- MCMC can again be used to simulate from the joint posterior.