MAST90125: Bayesian Statistical learning

Lecture 16: Generalised linear models in the Bayesian view

Feng Liu and Guoqi Qian

What have we just covered

In the last two lectures, we have discussed various types of linear models

- linear (fixed-effect) regression
- random effect regression
- linear mixed models
- LASSO

and how we can construct Gibbs samplers to estimate model parameters. For all these problems, we assumed that the residuals are drawn from a normal distribution.

In this lecture, we will consider how to extend regression models to cases when the (conditional) variation in the responses are not normally distributed. You may already know of this as generalised linear models.

Generalised linear models

We have met a specific generalised linear model before:

$$\begin{aligned} y_i &\sim \text{Pois}(\lambda_i) \\ \log(\lambda_i) &= \mathbf{x}_i'\boldsymbol{\beta}, \end{aligned}$$

where $\log(\cdot)$ is called the link function. Consider the following two questions:

- What is the benefit to introduce the link function?
- How to choose the link function?

In the following, we will introduce the origin regarding the link function.

Exponential family of distributions

- In lecture 4, we briefly discussed exponential family in the context of conjugate priors. Recall that the only distributions that can be conjugate priors are members of the exponential family.

- The pdf of the exponential family can be written as

$$p(y|\boldsymbol{\theta}) = f(y)g(\boldsymbol{\theta})e^{\eta(\boldsymbol{\theta})'u(y)}, \tag{1}$$

  where $\eta(\boldsymbol{\theta})$ is a vector of natural parameters. Furthermore $\eta(\boldsymbol{\theta})$ and $u(y)$ have the same length.

- We will now determine the natural parameter, $\eta(\boldsymbol{\theta})$ for some distributions. In GLMs, $\eta(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ is known as the canonical link function.

Determining $\eta(\boldsymbol{\theta})$

- Remember, to determine $\eta(\boldsymbol{\theta})$ we need to write the likelihood in the form of (1).
  - Binomial distribution: $\Pr(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$ (the trick: for a proper $f(x)$, we know $f(x) = e^{\log(f(x))}$).

Determining $\eta(\boldsymbol{\theta})$

- Remember, to determine $\eta(\boldsymbol{\theta})$ we need to write the likelihood in the form of (1).
  - Binomial distribution: $\Pr(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$ (the trick: for a proper $f(x)$, we know $f(x) = e^{\log(f(x))}$).

$$\binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} \left(\frac{p}{1-p}\right)^y (1-p)^n = \binom{n}{y} e^{\phi y} (1+e^{\phi})^{-n},$$

where $\eta(p) = \phi = \text{logit}(p) = \log(p/(1-p))$, $f(y) = \binom{n}{y}$, and $g(p) = (1+e^{\phi})^{-n}$.
The trick: $\left(\frac{p}{1-p}\right)^y = \exp(\log(\left(\frac{p}{1-p}\right)^y)) = \exp(y \log(\left(\frac{p}{1-p}\right))) = \exp(y\phi)$.

Determining $\eta(\boldsymbol{\theta})$

- Remember, to determine $\eta(\boldsymbol{\theta})$ we need to write the likelihood in the form of (1).
  - Multinomial distribution: $\Pr(\mathbf{y}|\mathbf{p}) = n! \prod_{i=1}^{K} \frac{p_i^{y_i}}{y_i!}$, with $\mathbf{p} = (p_1, \ldots, p_K)$, $\sum_{i=1}^{K} p_i = 1$, $\mathbf{y} = (y_1, \ldots, y_K)$, $n = \sum_{i=1}^{K} y_i$.

$$n! \prod_{i=1}^{K} \frac{p_i^{y_i}}{y_i!} = \frac{n!}{\prod_{i=1}^{K} y_i!} \prod_{i=1}^{K} p_i^{y_i} = \frac{n!}{\prod_{i=1}^{K} y_i!} \prod_{i=1}^{K} e^{y_i \log(p_i)} = \frac{n!}{\prod_{i=1}^{K} y_i!} e^{\sum_{i=1}^{K} y_i \log(p_i)}$$

where $\eta_i(\mathbf{p}) = \log(p_i)$, $f(\mathbf{y}) = \frac{n!}{\prod_{i=1}^{K} y_i!}$, and $g(\mathbf{p}) = 1$. This could have identifiability issues, as $\sum_{i=1}^{K} p_i = 1$, $n = \sum_{i=1}^{K} y_i$. To get around this, re-write $\Pr(\mathbf{y}|\mathbf{p})$ as,

$$\Pr(\mathbf{y}|\mathbf{p}) = \frac{n!}{\prod_{i=1}^{K} y_i!} e^{\sum_{i=1}^{K-1} y_i \log(p_i)} e^{\log(p_K)(n - \sum_{i=1}^{K-1} y_i)} = \frac{n!}{\prod_{i=1}^{K} y_i!} e^{\sum_{i=1}^{K-1} y_i \log(\frac{p_i}{p_K})} e^{n \log(p_K)}$$

Determining $\eta(\boldsymbol{\theta})$

- Remember, to determine $\eta(\boldsymbol{\theta})$ we need to write the likelihood in the form of (1).
  - Poisson distribution: $\Pr(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$.

$$\frac{\lambda^y e^{-\lambda}}{y!} = \frac{1}{y!} e^{y \log(\lambda)} e^{-\lambda}$$

where $\eta(\lambda) = \log(\lambda)$, $f(y) = \frac{1}{y!}$, and $g(\lambda) = e^{-\lambda}$.

  - Negative binomial distribution: $\Pr(y|r,p) = \binom{y+r-1}{y} p^y (1-p)^r$.

$$\binom{y+r-1}{y} p^y (1-p)^r = \binom{y+r-1}{y} e^{y \log(p)} (1-p)^r$$

where $\eta(p) = \log(p)$, $f(y) = \binom{y+r-1}{y}$, and $g(p) = (1-p)^r$.

What properties should $\eta(\boldsymbol{\theta})$ have?

- In this lecture, we want to extend regression to situations where the errors are not normally distributed.
- Firstly what can we say about the range of $\mathbf{x_i}'\boldsymbol{\beta}$ in a regression problem?

What properties should $\eta(\boldsymbol{\theta})$ have?

- In this lecture, we want to extend regression to situations where the errors are not normally distributed.
- Firstly what can we say about the range of $\mathbf{x_i}'\boldsymbol{\beta}$ in a regression problem?
  - $(-\infty, \infty)$
- Secondly, which of the 4 examples considered above have the range of the associated natural parameter $\eta(\theta)$ to be $(-\infty, \infty)$?

What properties should $\eta(\boldsymbol{\theta})$ have?

- In this lecture, we want to extend regression to situations where the errors are not normally distributed.
- Firstly what can we say about the range of $\mathbf{x_i}'\boldsymbol{\beta}$ in a regression problem?
  - $(-\infty, \infty)$
- Secondly, which of the 4 examples considered above have the range of the associated natural parameter $\eta(\theta)$ to be $(-\infty, \infty)$?
  - Binomial, Multinomial and Poisson.
- But was the range of θ or $\boldsymbol{\theta}$ still $(-\infty, \infty)$ in these cases?

What properties should $\eta(\boldsymbol{\theta})$ have?

- In this lecture, we want to extend regression to situations where the errors are not normally distributed.
- Firstly what can we say about the range of $\mathbf{x_i}'\boldsymbol{\beta}$ in a regression problem?
  - $(-\infty, \infty)$
- Secondly, which of the 4 examples considered above have the range of the associated natural parameter $\eta(\theta)$ to be $(-\infty, \infty)$?
  - Binomial, Multinomial and Poisson.
- But was the range of θ or $\boldsymbol{\theta}$ still $(-\infty, \infty)$ in these cases?
  - No. A proportion is constrained to $(0, 1)$, and λ is constrained to $(0, \infty)$.

Bayesian fitting of Generalised linear models

- If we choose the canonical link, $\eta(\theta)$, the likelihood for the binomial, multinomial and Poisson distributed data (assumed N independent observations) will be
  - Binomial (assume $\eta(\theta) = \mathbf{x}_j'\boldsymbol{\beta}$):

$$\Pr(\mathbf{y}|\boldsymbol{\beta}) = \prod_{j=1}^{N} \binom{n_j}{y_j} e^{(\mathbf{x}_j'\boldsymbol{\beta})y_j} (1 + e^{(\mathbf{x}_j'\boldsymbol{\beta})})^{-n_j}$$

  - Multinomial (assume $\eta(\theta) = \mathbf{x}_j'\boldsymbol{\beta}$):

$$\Pr(\mathbf{y}|\boldsymbol{\beta}_1 \dots \boldsymbol{\beta}_{K-1}) = \prod_{j=1}^{N} \frac{n!}{\prod_{i=1}^{K} y_{ij}!} e^{\sum_{i=1}^{K-1} y_{ij}(\mathbf{x}_j'\boldsymbol{\beta}_i)} e^{-n_j \log(1+\sum_{i=1}^{K-1} e^{\mathbf{x}_j'\boldsymbol{\beta}_i})}$$

  - Poisson (assume $\eta(\theta) = \mathbf{x}_j'\boldsymbol{\beta}$):

$$\Pr(\mathbf{y}|\boldsymbol{\beta}) = \prod_{j=1}^{N} \frac{1}{y_j!} e^{y_j(\mathbf{x}_j'\boldsymbol{\beta})} e^{-e^{\mathbf{x}_j'\boldsymbol{\beta}}}$$

Bayesian fitting of Generalised linear models

- The likelihoods given on the previous slide are functions of $\boldsymbol{\beta}$. The involvement of $\boldsymbol{\beta}$ is always through $e^{\mathbf{x}_j'\boldsymbol{\beta}}$.
- This suggests the conditional posterior pdf of $\boldsymbol{\beta}$ is unlikely to be a (multivariate) normal, which is different from what we see in normal regression models where the conditional posterior pdf of $\boldsymbol{\beta}$ is (multivariate) normal given a normal prior of $\boldsymbol{\beta}$.
- This further implies it may be difficult to apply Gibbs sampling for simulating the posterior pdf of $\boldsymbol{\beta}$. Refer to the Poisson regression problem studied in Lab 6.

What do we use if Gibbs sampling is not possible?

- We may be able to use Metropolis-Hastings algorithms to approximate the posterior of $\boldsymbol{\beta}$ in GLM. The question becomes how to choose an appropriate proposal distribution for an efficient implementation of MH algorithm.
- In Lab 6, we have seen the use of a normal proposal distribution for simulating the posterior pdf of $\boldsymbol{\beta}$ by the MH algorithm in Poisson regression. Two different normal proposal distributions are used there.
- In lecture 6, we showed that as $n \to \infty$, $p(\boldsymbol{\theta}|y)$ converges in distribution to normal with mean vector $= \hat{\boldsymbol{\theta}}_{MAP}$ and variance-covariance matrix equal to the inverse of the information matrix, $I(\hat{\boldsymbol{\theta}}_{MAP})^{-1}$.
- Applying this asymptotic result to the simulated posterior samples of $\boldsymbol{\beta}$ may provide a better estimation of the posterior pdf of $\boldsymbol{\beta}$. We will look into the detail of this future lectures.

Do we always need to use the canonical link?

- In practice, you are not constrained to the canonical link. Let's go back to GLM:

$$y_i \sim \text{Pois}(\lambda_i), \quad \log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta},$$

where the log link function is used.

Do we always need to use the canonical link?

- In practice, you are not constrained to the canonical link. Let's go back to GLM:

$$y_i \sim \text{Pois}(\lambda_i), \quad \log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta},$$

  where the log link function is used.

- Sometimes an identity link can also be used for Poisson regression model.
  - Consider a homogeneous Poisson point process. This is an example where the identity link is used, as it assumes $\eta(\lambda_i) = \lambda_i = \beta t_i$, leading to the likelihood,

$$p(y_1, \dots y_n | \beta, t_1, \dots t_n) = \prod_{i=1}^{n} \frac{(\beta t_i)^{y_i} e^{-\beta t_i}}{y_i!} = \frac{\beta^{n\bar{y}} e^{-\beta n \bar{t}} \prod_{i=1}^{n} t_i^{y_i}}{\prod_{i=1}^{n} y_i!}.$$

    If we combine this with a $\text{Ga}(\alpha, \gamma)$ prior for β, we can see the posterior for β would be $\text{Ga}(\alpha + n\bar{y}, \gamma + n\bar{t})$.

Special case: regression for binary responses

- A particular case having considerable diversity in choosing the link function is when the response is binary (zero or one). In this lecture, we have already encountered the logit link,
  - $\eta(p) = \log(p/(1-p))$.
- Another commonly used link function is the probit:

$$\eta(p) = \Phi^{-1}(p),$$

  where Φ^{-1} is the inverse cdf of the standard normal.
- In the following, we will focus on the probit link.

Probit regression

- If the link chosen is probit, we can view the observed data y_i as an dichotomisation of a continuous latent variable z_i such that

$$y_i = \begin{cases} 1 & \text{if } z_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } z_i \sim \mathcal{N}(\mu_i, 1).$$

  - But why is the variance of z_i fixed to 1?
    - Because the variance is not identifiable. If we let $z_i \sim \mathcal{N}(\mu_i, \sigma^2)$ the probability that $y_i = 1$ is
      $\Pr(z_i \geq 0) = \Pr(Z \geq -\mu_i/\sigma) = \Pr(Z \leq \mu_i/\sigma) = \Phi(\mu_i/\sigma) = \Phi(c\mu_i/c\sigma) \quad \forall c \in \mathbb{R}$,
      where $Z \sim \mathcal{N}(0, 1)$
  - The motivation for considering a probit link is to enable easy sampling.

Probit regression

- Since z_i depends on μ_i, given the definitions of the previous slide, the likelihood of a probit regression (note that $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ and $\eta(p_i) = \mu_i$) is,

$$\Pr(y_1, \ldots y_n | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i} \times (1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i},$$

where $\Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}$ corresponds to $p_i^{y_i}$ in the binomial case, because

$$p_i = \eta^{-1}(\mu_i) = \Phi(\mu_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}).$$

However, this does not appear to be much of an improvement.

Probit regression

- Based on the definition of Φ and the range of y_i, we can re-write the likelihood as follows,

$$\begin{aligned}\prod_{i=1}^{n} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i} \times (1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i} &= \prod_{i=1}^{n} \int \mathbb{1}_{\text{sign}(z_i)=\text{sign}(y_i-1/2)} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2}} dz_i \\ &= \prod_{i=1}^{n} \int \Pr(y_i|z_i) p(z_i|\mathbf{X}, \boldsymbol{\beta}) dz_i,\end{aligned}$$

- What does this formula look like?

Probit regression

- So what do we gain by augmenting y_i with z_i?
  - $\boldsymbol{\beta}$ only appears as a parameter in the distribution of z_i.
  - As the distribution of z_i is normal, this implies that if we know $z_1, \ldots z_n$, we can use the Gibbs sampler outlined in lectures 13 and 14 to estimate $\boldsymbol{\beta}$, depending on what prior we use for $\boldsymbol{\beta}$.
- But how can we learn z_i? By the rules of probability, the posterior for z_i is,

$$p(z_i|y_i, \mathbf{X}, \boldsymbol{\beta}) = \frac{p(y_i, z_i|\mathbf{X}, \boldsymbol{\beta})}{\Pr(y_i|\mathbf{X}, \boldsymbol{\beta})} = \begin{cases} \frac{1}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})\sqrt{2\pi}} e^{-\frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2}} & \text{if } y_i = 1 \text{ and } z_i \geq 0. \\ \frac{1}{(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))\sqrt{2\pi}} e^{-\frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2}} & \text{if } y_i = 0 \text{ and } z_i \leq 0. \\ 0 & \text{otherwise} \end{cases}$$

- This posterior of z_i corresponds to a truncated normal distribution defined on $(0, \infty)$ if $y_i = 1$ and $(-\infty, 0)$ if $y_i = 0$.

Gibbs sampling for probit regression

- To further help understanding, let's determine a Gibbs sampler for probit regression. We know $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and will assume $p(\boldsymbol{\beta}) \propto 1$.
- The joint pdf $p(\mathbf{y}, \mathbf{z}, \boldsymbol{\beta})$ is

$$\prod_{i=1}^{n} \mathbb{1}_{\text{sign}(z_i)=\text{sign}(y_i-1/2)} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2}}.$$

- The kernel of $\boldsymbol{\beta}$ is

$$\prod_{i=1}^{n} e^{-\frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2}} = e^{-\frac{\sum_{i=1}^{n}(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2}} = e^{-\frac{(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})}{2}} \propto e^{-\frac{\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}}{2}},$$

  which implies that the conditional posterior $p(\boldsymbol{\beta}|\mathbf{z})$ is $\mathcal{N}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}, (\mathbf{X}'\mathbf{X})^{-1})$.
- Then we just cycle between sampling from $p(\boldsymbol{\beta}|\mathbf{z})$ and $p(z_i|\boldsymbol{\beta}, y_i)$; $i = 1, \ldots, n$.

Model checking reminders

- As with all models studied, remember to incorporate appropriate model checking diagnostics when fitting your model.
- This is particularly important in generalised linear models where residual variation is constrained by the choice of likelihood.
  - In previous lectures we have fitted normal regression models, and included posterior predictive checking. For example, we used posterior predictive checking to look for evidence on whether coefficients were correctly specified or whether residual variance was associated with grouping.
  - What might you do if you find the model is not good?