STATS305B – Lecture 9

Jonathon Taylor
Scribed by Michael Howes

02/02/22

Contents

1 Gibbs sampler

Here is a follow-up on the Gibb's sampler from last lecture. The procedure is as follows:

1. Initialize $\beta^{(0)}$.
2. At time set $t \geq 1$, pick a coordinate $1 \leq j \leq p$.
3. Draw u^{new} from the density $g(u|\beta_{-j}^{(t-1)}, Y)$.
4. Set $\beta_j^{(t)} = u^{\text{new}}$ and $\beta_{-j}^{(t)} = \beta_{-j}^{(t-1)}$.
5. Increment t to $t+1$ and return to step 2.

2 Multinomial model

We will now consider some models for a multinomial response Y.

2.1 Baseline model

An issue with the multinomial model is that $\sum_{j=1}^{k} \pi_j = 1$, and so we only have $k-1$ free parameters. We thus work with the following re-parametrized density

$$f(y_1, \ldots, y_{k-1}|\pi) = \binom{N}{y_1, \ldots, y_{k-1}, n - \sum_{j=1}^{k-1} y_j} \left(\prod_{j=1}^{k-1} \pi_j^{y_j} \right) \left(1 - \sum_{j=1}^{k-1} \pi_j \right)^{N - \sum_{j=1}^{k-1} y_j},$$

which has support

$$\{y \in \mathbb{Z}_+^{k-1} : y_1 + \ldots + y_{k-1} \leq N\},$$

and parameter space

$$\{\pi \in [0,1]^{k-1} | \pi_1 + \ldots + \pi_{k-1} \leq 1\}.$$

The log-likelihood is

$$\log(L(\pi|Y)) = \sum_{j=1}^{k-1} y_j \log \left(\frac{\pi_j}{1 - \sum_{l=1}^{k-1} \pi_l} \right) + N \log \left(1 - \sum_{l=1}^{k-1} \pi_l \right).$$

We will thus use the natural parameters

$$\eta_j = \log \left(\frac{\pi_j}{1 - \sum_{l=1}^{k-1} \pi_l} \right), \text{ for } j = 1, \ldots, k-1.$$

To perform regressions with features X we parametrize η as a linear combination of the features. That is, we suppose that

$$\eta_{ij} = X[i, \cdot]^T \beta[\cdot, j],$$

or simply $\eta = X\beta$ where $\eta \in \mathbb{R}^{n \times (k-1)}$, $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^{p \times (k-1)}$. In this model we thus have $p \times (k-1)$ parameters since there are p features and $k-1$ classes for each observation. The response Y is also in $\mathbb{R}^{n \times (k-1)}$. Suppose that for each observation i, $Y[i, \cdot]$ has exactly one 1. So that $Y[i, \cdot] \sim \mathsf{Multinomial}(1, \pi_i)$. To recover the probabilities π from the parameters β we have

$$\mathbb{P}_\beta(Y_{i,j} = 1 | X_i) = \pi_{ij}(\beta) = \frac{\exp(X[i, \cdot]^T \beta[\cdot, j])}{1 + \sum_{l=1}^{k-1} \exp(X[i, \cdot]^T \beta[\cdot, l])},$$

for $1 \leq j \leq k-1$. As a function of β, the log-likelihood is

$$\log L(\beta|Y) = \sum_{i=1}^{n} \left(\sum_{j=1}^{k-1} Y_{i,j} X[i, \cdot]^T \beta[, \cdot j] - \log \left(1 + \sum_{l=1}^{k-1} \exp(X[i, \cdot]^T \beta[\cdot, l]) \right) \right).$$

And

$$\nabla \log L(\beta|Y) = X^T (Y - \mathbb{E}_\beta[Y]) \in \mathbb{R}^{p \times (k-1)}.$$

This is because the term

$$\log \left(1 + \sum_{l=1}^{k-1} \exp(X[i, \cdot]^T \beta[\cdot, l]) \right),$$

is the cumulant generating function of Y. Furthermore,

$$\frac{\partial^2}{\partial \beta_{i,j} \beta_{l,m}} \log L(\beta|X) = -\sum_{c=1}^{n} X_{ci} X_{cl} \mathrm{Cov}_\beta(Y_c)_{jm},$$

where $\mathrm{Cov}_\beta(Y_c) = \mathrm{diag}(\pi_c) - \pi_c \pi_c^T$, still assuming that $N_c = 1$ for all $1 \leq c \leq n$. Since our likelihood is a function of $p \times (k-1)$ parameters, the gradient is also a $p \times (k-1)$ matrix and the Hessian is a $(p \times (k-1)) \times (p \times (k-1))$ tensor. This complicates notation but the ideas of gradient descent and quasi-Newton methods still work. Importantly baseline models can be easily fit in R by using `nnet::multinom()`.

2.1.1 Interpretation of parameters in baseline model

Like in logistic regression the parameters β tell us something about the odds-ratios for different covariates. In particular for a fixed $1 \leq i \leq k-1$ and $1 \leq j \leq p$, we have

$$\begin{aligned} OR_{i,j} &= \frac{\frac{\mathbb{P}(Y_i=1|X_{-j}=x_{-j},X_j=x_j+1)}{\mathbb{P}(Y_k=1|X_{-j}=X_{-j},x_j+1)}}{\frac{\mathbb{P}(Y_i=1|X_{-j}=x_{-j},X_j=x_j)}{\mathbb{P}(Y_k=1|X_{-j}=X_{-j},x_j)}} \\ &= e^{\beta_{i,j}}. \end{aligned}$$

So $\beta_{i,j}$ tells us, conditioned on the other features, how much feature j increases the probability of being in class i relative to the probability of being in class k.

2.2 Latent variable model

Suppose $U_{i,j} = X[i,\cdot]^T\beta[\cdot,j] + \varepsilon_{i,j}$ for some noise distribution $\varepsilon_{i,j}$ that is independent and identically distributed across i,j. We could turn this into a multinomial model Y_i belonging to the class l that satisfies,

$$l = \underset{1\leq j\leq k}{\operatorname{argmax}} U_{i,j}.$$

Thus,

$$\pi_j(X) = \mathbb{P}_\beta(Y_j = 1|X) = \mathbb{P}_\beta\left(X[i,\cdot]^T\beta[\cdot,j] + \varepsilon_{i,j} \geq \max_{l\neq j} X[i,\cdot]^T Y[\cdot,l] + \varepsilon_{i,l}\right).$$

If we assume that ε_{ij} follow the `cloglog` or Gumble distribution. Then we recover the baseline model with parameters $\beta[\cdot,j] - \beta[\cdot,k]$. A more natural model would be $\varepsilon_{i,j} \sim \mathsf{N}(0,1)$. In this case the likelihood become rather complicated, but data augmented sampling is relatively straight forward. Thus, using normal noise is convenient for Bayesian analysis.

2.3 Cumulative logit model

It is common for the levels of Y to be ordinal. This allows for a different model. The cumulative logit model uses,

$$\mathbb{P}(Y \leq j|X) = \sum_{l=1}^{j} \pi_l(X).$$

To make this a glm we would model

$$\texttt{logit}(\mathbb{P}(Y \leq j|X)) = \alpha_j + \beta^T X,$$

where $1 \leq j \leq k-1$ and $\beta \in \mathbb{R}^p$. This gives a model with fewer parameters, but this model is trickier to fit since the link is not canonical. Like the baseline model, we can think of the cumulative logit model as a latent variable model. In this case we have $\varepsilon_i \sim F$ i.i.d. for every i. We then define $U = X^T\beta + \varepsilon$ and we have

$$Y_j = 1 \iff U \in [\alpha_{j-1}, \alpha_j].$$

An example of such ordinal data would be a questionnaire about going to graduate school, where each student has covariates such as parents education, test scores, income and the response is a rating of how likely they are to go to graduate school. A cumulative logit model could be fit in R using the function `polr` from the library `MASS`. The name `polr` is short for "proportional odds logistic regression" which is another name for the cumulative logit model. But a warning: the function `MASS::polr` actually fits the model

$$\texttt{logit}(\mathbb{P}(Y \leq j|X)) = \alpha_j - \beta^T X.$$

And so the interpretation of the coefficients is reversed.

2.4 Adjacent category model

Consider the parameters

$$\log \frac{\pi_j(X)}{\pi_{j+1}(X)}, \quad \text{for } 1 \le j \le k-1.$$

We could make the following model for an ordinal response Y,

$$\log \frac{\pi_j(X)}{\pi_{j+1}(X)} = \alpha_j + X^T\beta.$$

Like the cumulative model, this model also has $K-1+p$ parameters. Note that, for $1 \le j \le k-1$

$$\begin{aligned}\log \frac{\pi_j(X)}{\pi_k(X)} &= \sum_{j=1}^{k-1} \log \frac{\pi_j(X)}{\pi_{j+1}(X)} \\ &= \alpha_j^* + (k-j)X^T\beta,\end{aligned}$$

for where $\alpha_j^* = \sum_{l=j}^{k-1} \alpha_l$. Thus, the adjacent category model is a baseline multinomial model subject to a certain constraint. In the baseline multinomial model we have parameters $B \in \mathbb{R}^{p\times(k-1)}$ (what we previously called β). In the adjacent category model we require that $B = \beta v^T$ for some $\beta \in \mathbb{R}^p$ where $v \in \mathbb{R}^{k-1}$ has entries $v_j = k - j$.

2.5 Bayesian multinomial models

One can use STAN to generate posterior samples in the baseline and the cumulative logit models. As mentioned before, data augmented sampling can be used for the cumulative logit model. For the baseline model, STAN uses the over-parametrized model

$$P(Y = j|X) = \frac{\exp(X^T\beta[\cdot, j])}{\sum_{l=1}^k \exp(X^T\beta[\cdot, l])}.$$

3 Regularization for glms

The objective function for a glm can be written as a minimization problem

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} - \log L(\beta|Y) = \underset{\beta}{\operatorname{argmin}}\, \Lambda(X\beta) - \beta^T(X^TY),$$

where $\Lambda(\eta) = \sum_{i=1}^n \Lambda(\eta_i)$ is the cumulant generation function for an independent sample of size n. Recall that for the most familiar types of glms we have

- Gaussian: $\Lambda(\eta) = \frac{\eta^2}{2}$,
- Logistic: $\Lambda(\eta) = \log(1 + e^\eta)$,
- Poisson: $\Lambda(\eta) = e^\eta$.

The above minimization problem is convex in β and so it is natural to add a penalty to get the new objective

$$\widehat{\beta}_{\mathcal{P}} = \underset{\beta}{\operatorname{argmin}}\, \Lambda(X\beta) - \beta^T(X^TY) + \mathcal{P}(\beta),$$

where $\mathcal{P}(\beta)$ is a convex penalty on β. The two main penalties we will consider are the *ridge penalty* $\mathcal{P}(\beta) = \frac{\lambda}{2}\sum_{j=1}^p \beta_j^2$ and the *LASSO penalty* $\mathcal{P}(\beta) = \lambda \sum_{j=1}^p |\beta_j|$.

3.1 Ridge regression

As stated before, in ridge regression we have

$$\mathcal{P}(\beta) = \frac{\lambda}{2} \|\beta\|_2^2 = \frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda > 0$ is a hyperparameter. The ridge estimate is thus

$$\widehat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \, \Lambda(X\beta) - \beta^T (X^T Y) + \frac{\lambda}{2} \|\beta\|_2^2 .$$

When do ridge regression, we normally scale each feature $X[\cdot, j]$ so that the coefficients β are unit-less. The penalized objective can be minimized by Newton–Raphson or Fisher scoring. But we will also see methods such as coordinate descent and proximal gradient descent which work better for regularized models.

3.2 "Bayesian" interpretation

If we put a $\mathsf{N}(0, \lambda^{-1} I)$ prior on β, then the ridge estimator is the *MAP* or *maximum a posteriori* estimator. While this involves a prior, it isn't truly Bayesian. A Bayesian would work with the full posterior distribution note just the MAP which is the mode of the distribution. We could also consider a general quadratic penalty,

$$\mathcal{P}(\beta) = \frac{1}{2} \beta^T Q \beta,$$

where Q is positive definite and symmetric. The resulting regularized estimator is the MAP estimator if we assume a prior $\beta \sim \mathsf{N}(0, Q^{-1/2})$.