1 Variational Bayes Logistic Regression

1.1 Logistic Regression Model

In the standnard logistic regression model, the log transformation of likelihood function $l(\beta)$ is:

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} \log p(y_i|X_i) \\ &= \sum_{i=1}^{n} y_i \log p(X_i, \beta) + (1 - y_i) \log(1 - p(X_i, \beta)), \end{aligned} \tag{1}$$

where

$$p(x, \beta) = \frac{\exp(\beta^T x)}{1 + exp(\beta^T x)}, \text{and } Y = [y_1, \cdots, y_n], X = [X_1, \cdots, X_n].$$

From the frequentist inference perspective, we can find use maximum likelihood estimation by numerical root-finding algorithms, such as gradient descent method and Newton–Raphson method.

Moreover, if we have more information about parameter β, from Bayesian perspective, we can put those information into priors $p(\beta)$. Then, the estimation can be computed by computing posterior distribution $p(\beta|Y) \propto l(Y)p(\beta)$.

However, finding posterior $p(\beta|Y)$ for logistic regression model here is analytic intractable, because the sigmoid data likelihood does not admit a conjugate exponential prior. Thus, approximations are needed in order to obtain Bayes estimation for logistic regression. There are many techniques, including stochastic and deterministic methods, for approximating a probability density function (pdf). In recent years, stochastic sampling techniques such as the Markov chain Monte Carlo (MCMC) method have been developed. The MCMC method uses stochastic sampling inference algorithms to side-step the problem of multiple integration, so it can approximate the Bayesian posterior efficiently. Although it can yield a approximate pdf to an arbitrary level of accuracy, the MCMC method is time-consuming and cannot ensure convergence [2].

Unlike the stochastic method, deterministic approximation methods have no requirements for stochastic sampling. It is known that the Laplace approximation, the maximum entropy approximation and the spline approximation are tractable approximations of pdfs. Besides, in this paper, we will use another deterministic approximation method, variational Bayes method, to get posterior distribution of $p(\beta|Y)$.

1.2 Variational Ideas for Approximating Bayesian Inference

In many statistical problems, we desire to make statistical inference on the process that produces the data collected based on our prior knowledge of this

process. Let θ be the multivariate random variable containing the unknown parameters in the population pdf and latent variables. Our prior knowledge of θ is encapsulated by the pdf $p(\theta)$. When the observation Y is obtained, using the likelihood function $p(Y|\theta)$ and Bayes' rule, the prior $p(\theta)$ can be updated to the posterior pdf

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}.$$

The difficulties of Bayesian inference lie in the normalizing constant $p(Y)$, known as the model evidence. In the following, we adopt the variational Bayes method to approximate the Bayesian inference.

1.2.1 The lower bound of evidence

The main idea of the variational Bayes method is to find a tractable approximate pdf $q(\theta)$ for the true posterior pdf $p(\theta|Y)$ as well as for the model evidence $p(Y)$. Using Jensen's inequality, we have

$$\begin{aligned}
\ln p(Y) &= \ln \int q(\theta) \frac{p(Y,\theta)}{q(\theta)} d\theta \\
&\geq \int q(\theta) \ln \left(\frac{p(Y,\theta)}{q(\theta)} \right) d\theta \\
&= \int q(\theta) \ln p(Y) d\theta - \int q(\theta) \ln \left(\frac{q(\theta)}{p(\theta|Y)} \right) d\theta \\
&= \ln p(Y) - \mathrm{KL}\big(q(\theta)||p(\theta|Y)\big) \equiv L\big(q(\theta)\big), \qquad (2)
\end{aligned}$$

where $\mathrm{KL}(q(\theta)||p(\theta|Y)) \geq 0$ is the Kullback-Leibler (KL) divergence from $q(\theta)$ to $p(\theta|Y)$, and $L(q(\theta))$ can be treated as a lower bound on $\ln p(Y)$. To find an approximate pdf $q(\theta)$ close enough to the true posterior pdf $p(\theta|Y)$, we need to minimize the KL divergence $\mathrm{KL}(q(\theta)||p(\theta|Y))$, or equivalently maximize the lower bound $L(q(\theta))$. Therefore, the statistical inference problem for the posterior pdf $p(\theta|Y)$ is converted to optimizing a variational approximate pdf $q(\theta)$.

1.2.2 Variational Bayes theorem

Next, we seek the optimized pdf $q(\theta)$. One tractable way for the variational Bayes method is to employ the mean-field family [1]. In this family, random vectors that form θ can be divided into several independent groups. Suppose that $q(\theta)$ can be divided into M independent groups:

$$q(\theta) = \prod_{m=1}^{M} q(\theta_m). \qquad (3)$$

Let

$$L_m(\cdot) = L(q(\theta_1, \ldots, \theta_{m-1}, \cdot, \theta_{m+1}, \ldots, \theta_M)).$$

Denote the set of all components of θ not in the mth group by $\theta_{-m} = \{\theta_n, n \neq m\}$. From

$$\begin{aligned} L(q(\theta)) &= \mathbb{E}_{q(\theta)}\left[\ln\left(\frac{p(Y,\theta_m,\theta_{-m})}{q(\theta_m,\theta_{-m})}\right)\right] \\ &= \mathbb{E}_{q(\theta)}\left[\ln\left(\frac{p(\theta_m|Y,\theta_{-m})p(Y,\theta_{-m})}{q(\theta_m,\theta_{-m})}\right)\right], \end{aligned}$$

we have

$$L_m(q(\theta_m)) \equiv \mathbb{E}_{q(\theta)}[\ln p(\theta_m|Y,\theta_{-m})] - \mathbb{E}_{q(\theta)}[\ln q(\theta_m)] + C_1, \tag{4}$$

where

$$C_1 = \mathbb{E}_{q(\theta)}[\ln p(Y,\theta_{-m})] - \mathbb{E}_{q(\theta)}[\ln q(\theta_{-m})]$$

is independent of $q(\theta_m)$.

Taking the functional (Gateaux) derivative of (4) with respect to $q(\theta_m)$ and equating to zero, we could maximum $L_m(q(\theta_m))$ in (4) when $q(\theta_m)$ satisfies [1, 2]

$$q(\theta_m) \propto \exp\left(\mathbb{E}_{q(\theta_{-m})}[\ln p(Y,\theta)]\right). \tag{5}$$

Theorem 1 *Let $p(\theta|Y)$ be the posterior pdf of the random vector θ, and the pdf $q(\theta)$ is restricted to be conditionally independent for $\theta_1, \ldots, \theta_M$,*

$$q(\theta) = q(\theta_1, \ldots, \theta_M) = \prod_{m=1}^{M} q(\theta_m).$$

The optimized pdfs $q(\theta_1), \ldots, q(\theta_M)$ in the sense of maximizing the lower bound of model evidence are given by (5). We refer to $q(\theta_m)$ as the variational Bayes marginal for $m = 1, \ldots, M$ and $q(\theta)$ as the variational Bayes approximated posterior.

Remark 1 *The optimized pdf $q(\theta_m)$ is given by (5) regardless of the particular observation equation and has several desirable properties including convergence and asymptotic normality as the number of iterations grows to infinity. It was proved in [4] that the lower bound $L(q(\theta))$ will increase gradually with iterations and $q(\theta)$ with converge to a local optimal posterior. The optimal pdf $q(\theta)$ can be achieved with convergence by optimizing each sub-vectors iteratively.*

1.3 Variational Bayes Logistic Regression

Details about variational Bayes method of ordinary logistic regression model can be found in [2] and [3]. Here, we briefly introduce the process.

First, the likelihood function is

$$p(Y|X,\beta) = \prod_{i=1}^{n} p(y_i|x_i,\beta) = \prod_{i=1}^{n} \sigma(y_i\beta^T x_i),$$

where

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

Its could be approximated by

$$\sigma(x) \geq \sigma(\epsilon) \exp((x - \epsilon)/2 - \lambda(\epsilon)(x^2 - \epsilon^2)), \lambda(\epsilon) = \frac{1}{2\epsilon}(\sigma(\epsilon) - \frac{1}{2}).$$

Then, we can find lower-bounder as

$$\begin{aligned} &\log p(Y|X, \beta) \geq \log p(\hat{Y}|X, \beta) \\ &= \beta^T \sum_i \frac{x_i y_i}{2} - \beta^T (\sum_i \lambda(\epsilon_i) x_i x_i^T)\beta + \sum_i (\log(\sigma(\epsilon_i)) - \frac{\epsilon_i}{2} + \lambda(\epsilon_i)\epsilon_i^2), \end{aligned}$$

Next, let's assume priors for parameters $\theta = (\beta, \alpha)^T$. We use conjugate Gaussian prior for parameter β as

$$p(\beta|\alpha) \sim N(\beta|0, \alpha^{-1} I),$$

and conjugate Gamma distribution for hyper-parameter α as

$$p(\alpha) \sim Gamma(\alpha|a_0, b_0).$$

Then, we use variational posterior distribution $q(\beta, \alpha)$ as approximation to true posterior $p(\beta, \alpha|X, Y)$. We also assume the posterior distribution $q(\beta, \alpha)$ belongs to mean-field family, such as $q(\beta, \alpha) = q(\beta)q(\alpha)$. The variational posterior of β is

$$q(\beta) \sim N(\mu_N, \Sigma_N)$$

where

$$\mu_N = \Sigma_N \sum_i \frac{x_i y_i}{2}, \Sigma_N^{-1} = \mathbb{E}[\alpha] I + 2 \sum_i \lambda(\epsilon_i) x_i x_i^T.$$

The variational posterior of α is

$$q(\alpha) \sim Gamma(a_N, b_N)$$

where

$$a_N = a_0 + D/2, b_N = b_0 + \frac{1}{2}\mathbb{E}[\beta^T \beta].$$

Remark 2 *Here we also use the automatic relevance determination (ARD) method to get sparse estimation of β, which is similar to Bayesian Ridge Regression. One of the advantages using ARD is unnecessary parameters in β are automatically forced to zero.*

1.4 Variational Bayes Multinomial Logistic Regression

Multinomial logistic regression model generalizes the normal logistic regression by considering multiclass problems. Here, we continue on developing variational Bayes method for the multinomial logistic regression model.

$$L(\beta) = \sum_{j=1}^{J}\sum_{i=1}^{n} n_{ij} \log p(\tilde{y}_i|X_i, \beta_j) + \mathrm{C}, \tag{6}$$

where C is a constant unrelated with β.

Similar, we use mean-filed family to approximate posterior as

$$q(\beta) = \prod_{j=1}^{k} q(\beta_j|\alpha_j)q(\alpha_j),$$

Apply the idea of variational Bayes method, first we fixed $q(\beta_{-k}, \alpha_{-k})$ and then estimate $q(\beta_k, \alpha_k)$. This is equal to binary logistic regression model as

$$p(\check{y} = 1|X) = p(y = k|X), p(\check{y} = 0|X) = p(y \neq k|X),$$

where

$$\check{y}_i = \begin{cases} 1, & \text{if } y_i = k, \\ 0, & \text{if } y_i \neq k. \end{cases}$$

So for data set $(\check{y}_i, X_i)$, we can use above variational Bayes logistic regression model to estimate its parameter $\check{\beta}$. Furthermore, we have a beautiful equation that connect both transformed binary logistic regression model and original multinomial logistic regression model as

$$\begin{aligned} \frac{e^{\check{\beta}^T x_i}}{1 + e^{\check{\beta}^T x_i}} &= \frac{e^{\beta_k^T x_i}}{e^{\beta_k^T x_k} + \sum_{j\neq k} \mathbb{E}[e^{\beta_j^T x_i}]}, \\ \frac{1}{1 + e^{\check{\beta}^T x_i}} &= \frac{\sum_{j\neq k} e^{\beta_j^T x_i}}{e^{\beta_k^T x_k} + \sum_{j\neq k} \mathbb{E}[e^{\beta_j^T x_i}]}, \end{aligned} \tag{7}$$

Together, we have

$$e^{\check{\beta}^T x_i} = \frac{e^{\beta_k^T x_i}}{\sum_{j\neq k} \mathbb{E}[e^{\beta_j^T x_i}]}, \tag{8}$$

which equals to

$$\check{\beta}^T x_i = \beta_k^T x_i - \log(\sum_{j\neq k} \mathbb{E}[e^{\beta_j^T x_i}]). \tag{9}$$

We set $\check{y}_i = \check{\beta}^T x_i + \log(\sum_{j\neq k} \mathbb{E}[e^{\beta_j^T x_i}])$, then we have

$$\check{y}_i = \beta_k^T x_i, \ i = 1, \cdots, N$$

Next, we develop equation for $\mathbb{E}[e^{\beta_j^T x_i}]$. If we assume the variational posterior $q(\beta_j) \sim N(\mu_j, \Sigma_j)$, and for m-th component $q(\beta_j^m) \sim N(\mu_j^m, (\sigma_j^m)^2)$, if components of β_j are independent, then

$$q(\beta_j^m x_i^m) \sim N(x_i^m \mu_j^m, (x_i^m \sigma_j^m)^2)$$

Therefore, $q(e^{\beta_j^m x_i^m}) \sim \log N(x_i^m \mu_j^m, (x_i^m \sigma_j^m)^2)$, which is log-normal distribution,

$$\mathbb{E}[e^{\beta_j^m x_i^m}] = \exp(x_i^m \mu_j^m + \frac{1}{2}(x_i^m \sigma_j^m)^2).$$

Combine all dimensions,

$$\begin{aligned}\mathbb{E}[e^{\beta_j^T x_i}] &= \prod_m \exp(x_i^m \mu_j^m + \frac{1}{2}(x_i^m \sigma_j^m)^2) \\ &= \exp(\mu_j^T x_i + \frac{1}{2} x_i \Sigma_j x_i^T)\end{aligned}$$

Combine equation (9) for all data, we will obtain a nonlinear equation system. The solution to this nonlinear equation system will give variational Bayes estimation for multinomial logistic regression. And many numerical algorithms could be used for finding the root of this nonlinear equation system.

As we can see above, by introducing variational Bayes idea, we can get a relatively simple algorithm for multinomial logistic regression model compare to other estimation methods, which usually require iterative computing steps.

References

[1] Matthew J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[2] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Jan Drugowitsch. Variational Bayesian inference for linear and logistic regression. October 2013.

[4] B. Wang and DM Titterington. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 577–584. 2004.