Bayesian Logistic Regression

April 26, 2022

1 Model Definition

We observe a set of pairs $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, ..., N\}$, with $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in [0, 1]$ (a binary classification response). we assume that the class-conditional probability of belonging to the "1" class is given by a nonlinear transformation of a linear function of $\boldsymbol{x}$:

$$P(y = 1|\boldsymbol{x}, \boldsymbol{w}) = \sigma(\boldsymbol{x}^T \boldsymbol{w}). \tag{1}$$

The most commonly used function σ is the logistic function:

$$\sigma(z) = \frac{1}{1 + exp(-z)}. \tag{2}$$

The class-conditional probability can be formulated as Bernoulli distribution with probability of success $p = \sigma(\boldsymbol{x}^T \boldsymbol{w})$:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \left(\frac{1}{1 + exp(-\boldsymbol{x}^T \boldsymbol{w})}\right)^{y_i} \left(\frac{exp(-\boldsymbol{x}^T \boldsymbol{w})}{1 + exp(-\boldsymbol{x}^T \boldsymbol{w})}\right)^{1-y_i} \tag{3}$$

We further assume that that predictions for different $\boldsymbol{x}$ are independent given $\boldsymbol{w}$, then the likelihood can be written

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{w}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{N} \left(\frac{1}{1 + exp(-\boldsymbol{x}_i^T \boldsymbol{w})}\right)^{y_i} \left(\frac{exp(-\boldsymbol{x}_i^T \boldsymbol{w})}{1 + exp(-\boldsymbol{x}_i^T \boldsymbol{w})}\right)^{1-y_i} \tag{4}$$

Since we are constructing a Bayesian model, we must assign a prior distribution on the unknown variables in the model. We choose zero mean normal priors with variance s^2 for the weights $\boldsymbol{w}$ which corresponds to weak information regarding the true parameters values. I.e.,

$$p(\boldsymbol{w}) = \mathcal{N}(w|\boldsymbol{0}, \boldsymbol{\Sigma}_p). \tag{5}$$

Using Bayes theorem, the posterior distribution is given by

$$p(\boldsymbol{w}|\mathcal{D}) = p(\boldsymbol{w}|\mathcal{Y},\mathcal{X}) = \frac{p(\mathcal{Y}|\mathcal{X},\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{Y}|\mathcal{X})} = \frac{p(\mathcal{Y}|\mathcal{X},\boldsymbol{w})p(\boldsymbol{w})}{\int p(\mathcal{Y}|\mathcal{X},\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}}. \tag{6}$$

To make predictions based the training data $\mathcal{D}$ for a test point $\boldsymbol{x}_*$we have

$$p(y_*|\boldsymbol{x}_*,\mathcal{D}) = \int p(y_*|\boldsymbol{w},\boldsymbol{x}_*,\mathcal{D})p(\boldsymbol{w}|\mathcal{D}) \tag{7}$$

We note that the posterior distribution, $p(\boldsymbol{w}|\mathcal{D})$, does not belong to a nice parametric family. Furthermore, the integral $p(y_*|\boldsymbol{x}_*,\mathcal{D})$ is intractable as well.

2 MAP Estimation

The MAP estimates, $\hat{\boldsymbol{w}}$ is defined as:

$$\begin{aligned}\hat{\boldsymbol{w}} &= \underset{w}{argmax}[p(\boldsymbol{w}|\mathcal{D})] = \underset{w}{argmax}[p(\mathcal{Y}|\mathcal{X},\boldsymbol{w})p(\boldsymbol{w})] = \\ &= \underset{w}{argmax}\{\sum_{i=1}^{N} y_i log[\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] + \sum_{i=1}^{N}(1-y_i)log[1-\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] - 0.5\boldsymbol{w}^T\Sigma_p\boldsymbol{w}\}\end{aligned} \tag{8}$$

we note that $\hat{\boldsymbol{w}}$ has no simple analytic form. However, it is easy to show that for some sigmoid functions, such as logistic and cumulative Gaussian , the log likelihood (4) is a concave function of $\boldsymbol{w}$ for fixed $\mathcal{D}$. As the quadratic term, $0.5\boldsymbol{w}^T\Sigma_p\boldsymbol{w}$, is also concave then the log posterior is a concave function, which means that it is relatively easy to find its unique maximum.

Although finding the MAP is a fast and easy way of obtaining estimates of the unknown model parameters, it is limited because there is no associated estimate of uncertainty produced with the MAP estimates.

3 Laplace Approximation

The idea of Laplace approximation is to approximate the posterior density with a Gaussian:

$$p(\boldsymbol{w}|\mathcal{D}) \approx \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{9}$$

The Laplace approximation is based on a second order Taylor expansion of the negative log of the unnormalized posterior

$$\Psi(\boldsymbol{w}) = -\log p(D|\boldsymbol{w}) - \log p(\boldsymbol{w}) \tag{10}$$

around the MAP estimate $\hat{\boldsymbol{w}}$. This results in the following Gaussian distribution

$$p(\boldsymbol{w}|\mathcal{D}) \approx \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}, H^{-1}) \tag{11}$$

where H is the Hessian of $\Psi(\boldsymbol{w})$ evaluated at $\hat{\boldsymbol{w}}$:

$$H = \nabla\nabla\Psi(\boldsymbol{w})|_{\boldsymbol{w}=\hat{\boldsymbol{w}}} \tag{12}$$

In the case of logistic regression, the Hessian is given by

$$H = \Sigma_p^{-1} + \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T \sigma(\boldsymbol{x}_i^T \hat{\boldsymbol{w}})[1 - \sigma(\boldsymbol{x}_i^T \hat{\boldsymbol{w}})] \tag{13}$$

To make a prediction, one needs the predictive distribution (7). However, the integral can not be computed analytically even with the Gaussian approximation of the posterior (11). A numerical approximation can however be easily obtained by Monte Carlo sampling (SLLN)

$$p(y_*|\boldsymbol{x}_*, \mathcal{D}) \approx \frac{1}{S}\sum_{i=1}^{S} \sigma(\boldsymbol{x}_{new}^T \boldsymbol{w}_s)$$

where $\boldsymbol{w}_s$ are independently sampled from $\mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}, H^{-1})$

4 Multiclass logistic regression

In the multiclass logistic regression formulation, the responses are consdired to be the set of 1-of-K encoded random vectors $\boldsymbol{y}$ of dimension K having the property that exectly one element has the the value 1 and the others have the value 0. In this formulation, the response vector, $\boldsymbol{y}$, can be modeled as categorial varaible,

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K) = \prod_{k=1}^{K} p_k^{y_k}$$

where p_k are the class conditional densities given by the softmax function,

$$p_k = P(C_k|\boldsymbol{x}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K) = \frac{\exp(\boldsymbol{w}_k^T \boldsymbol{x})}{\sum_{k'} \exp(\boldsymbol{w}_{k'}^T \boldsymbol{x})}$$

Assuming the data samples $\boldsymbol{x}_n$, $n = 1, ...N$ are statisticaly independent given , the likelihood can be written

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K) = \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{w}_1, ..., \boldsymbol{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} \left(\frac{\exp(\boldsymbol{w}_k^T \boldsymbol{x}_n)}{\sum_{k'} \exp(\boldsymbol{w}_{k'}^T \boldsymbol{x}_n)}\right)^{y_{nk}}$$

since we are constructing a Bayesian model, we must assign a prior distribution on the unknown variables in the model. We model each weights vector $w_{\boldsymbol{k}}$with multivariate normal prior with zero mean vector and covaraince matrix variance $\boldsymbol{I}_p s^2$ which corresponds to weak information regarding the true parameters values. I.e.,

$$p(\boldsymbol{w}_k) = \mathcal{N}(\mathbf{0}, \Sigma_p) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p s^2)$$

Since $\boldsymbol{w}_k, k = 1, .., K$ are assumed statisticaly independent we obtain

$$p(\boldsymbol{W}) = p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K) = \prod_{k=1}^{K} p(\boldsymbol{w}_k) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p s^2)$$

Using Bayes theorem, the posterior distribution is given by

$$\begin{aligned} p(\boldsymbol{W}|D) = p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K|\mathcal{D}) = p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K|\mathcal{Y}, \mathcal{X}) = \\ &= \frac{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K)p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K)}{p(\mathcal{Y}|\mathcal{X})} = \\ &= \frac{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K)p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K)}{\int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K)p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K)d\boldsymbol{w}_1...d\boldsymbol{w}_K} \end{aligned}$$

To make predictions based on the training data $\mathcal{D}$ for a test point $\boldsymbol{x}_*$we have

$$p(\boldsymbol{y}_*|\boldsymbol{x}_*, \mathcal{D}) = \int p(\boldsymbol{y}_*|\boldsymbol{x}_*, \mathcal{D}, \boldsymbol{W})p(\boldsymbol{W}|D)$$

The MAP estimates, $\hat{\boldsymbol{w}}_k$ is defined as:

$$\begin{aligned} \hat{\boldsymbol{w}}_k &= \arg\max_{\boldsymbol{w}_k} p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K|\mathcal{D}) = \arg\max_{\boldsymbol{w}_k} \log p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K|\mathcal{D}) = \\ &= \arg\max_{\boldsymbol{w}_k} [\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K) + \log p(\boldsymbol{w}_1, ..., \boldsymbol{w}_K)] \\ &= \arg\max_{\boldsymbol{w}_k} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk} \log \frac{\exp(\boldsymbol{w}_k^T \boldsymbol{x}_n)}{\sum_{k'} \exp(\boldsymbol{w}_{k'}^T \boldsymbol{x}_n)} - 0.5 s^{-2} \sum_{k=1}^{K} \boldsymbol{w}_k^T \boldsymbol{w}_k \right] \end{aligned}$$

The unnormalized log posteriror is given by

$$\Psi(\boldsymbol{w}_1, ..., \boldsymbol{w}_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk} \log \frac{\exp(\boldsymbol{w}_k^T \boldsymbol{x}_n)}{\sum_{k'} \exp(\boldsymbol{w}_{k'}^T \boldsymbol{x}_n)} + 0.5 s^{-2} \sum_{k=1}^{K} \boldsymbol{w}_k^T \boldsymbol{w}_k$$

The Hessian matrix is given by (see Section 5)

$$H = \mathcal{X} D \mathcal{X}^T + \Sigma_p^{-1}$$

where

$$\mathcal{D} \in \mathbb{R}^{N \times N}, \mathcal{D} = diag(d_1, d_2, ..., d_N),\ d_i = [1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})]\sigma(\boldsymbol{x}_i^T \boldsymbol{w})]$$

$$\mathcal{X} \in \mathbb{R}^{d \times N}, \mathcal{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]$$

5 Appendix A - Derivation of the Hessian Matrix

The Hessian Matrix is defined as follows

$$H = \nabla\nabla\Psi(\boldsymbol{w}) = \frac{\partial\Psi(\boldsymbol{w})}{\partial w \partial w^T} \tag{14}$$

for selecting $\Psi(\boldsymbol{w})$ to be the unnormalized log posterior for the logistic regression model we obtain

$$H = \frac{\partial\Psi(\boldsymbol{w})}{\partial\boldsymbol{w}\partial\boldsymbol{w}^T} = \frac{\partial}{\partial\boldsymbol{w}\partial\boldsymbol{w}^T}\left\{-\sum_{i=1}^{N} y_i \log[\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] - \sum_{i=1}^{N}(1-y_i)\log[1-\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] + 0.5\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right\} \tag{15}$$

We note the following properties of logistic function

$$1-\sigma(z) = 1 - \frac{1}{1+exp(-z)} = \frac{exp(-z)}{1+exp(-z)} = \frac{1}{1+exp(z)} = \sigma(-z) \tag{16}$$

$$\frac{\partial\sigma(z)}{\partial z} = \frac{exp(-z)}{[1+exp(-z)]^2} = \frac{exp(-z)}{1+exp(-z)}\frac{1}{1+exp(-z)} = [1-\sigma(z)]\sigma(z) \tag{17}$$

We start by taking the derivative of $\Psi(\boldsymbol{w})$ with respect to w^T

$$\begin{aligned}
\frac{\partial\Psi(\boldsymbol{w})}{\partial\boldsymbol{w}^T} &= \frac{\partial}{\partial\boldsymbol{w}^T}\left\{-\sum_{i=1}^{N} y_i \log[\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] - \sum_{i=1}^{N}(1-y_i)\log[1-\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] + 0.5\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right\} = \\
&\underset{(16)}{=} \frac{\partial}{\partial\boldsymbol{w}^T}\left\{-\sum_{i=1}^{N} y_i \log[\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] - \sum_{i=1}^{N}(1-y_i)\log[\sigma(-\boldsymbol{x}_i^T\boldsymbol{w})] + 0.5\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right\} \\
&= -\sum_{i=1}^{N} y_i\sigma(\boldsymbol{x}_i^T\boldsymbol{w})^{-1}\frac{\partial}{\partial\boldsymbol{w}^T}\sigma(\boldsymbol{x}_i^T\boldsymbol{w}) - \sum_{i=1}^{N}(1-y_i)\sigma(-\boldsymbol{x}_i^T\boldsymbol{w})^{-1}\frac{\partial}{\partial\boldsymbol{w}^T}\sigma(-\boldsymbol{x}_i^T\boldsymbol{w}) + \Sigma_p^{-1}\boldsymbol{w} = \qquad (18) \\
&= -\sum_{i=1}^{N} y_i[1-\sigma(\boldsymbol{x}_i^T\boldsymbol{w})]\boldsymbol{x}_i + \sum_{i=1}^{N}(1-y_i)\underbrace{[1-\sigma(-\boldsymbol{x}_i^T\boldsymbol{w})]}_{=\sigma(\boldsymbol{x}_i^T\boldsymbol{w})}\boldsymbol{x}_i + \Sigma_p^{-1}\boldsymbol{w} = \\
&= \sum_{i=1}^{N}\left\{-y_i[1-\sigma(\boldsymbol{x}_i^T\boldsymbol{w})] + (1-y_i)\sigma(\boldsymbol{x}_i^T\boldsymbol{w})\right\}\boldsymbol{x}_i + \Sigma_p^{-1}\boldsymbol{w} = \\
&= \sum_{i=1}^{N}[-y_i + \sigma(\boldsymbol{x}_i^T\boldsymbol{w})]\boldsymbol{x}_i + \Sigma_p^{-1}\boldsymbol{w} \qquad (19)
\end{aligned}$$

To obtain the Hessian matrix, we take the derivative of (18) with respect to $\boldsymbol{w}$

$$
\begin{aligned}
H = \frac{\partial \Psi(\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} &= \frac{\partial}{\partial \boldsymbol{w}} \left\{ \sum_{i=1}^{N} [-y_i + \sigma(\boldsymbol{x}_i^T \boldsymbol{w})] \boldsymbol{x}_i + \Sigma_p^{-1} \boldsymbol{w} \right\} = \\
&= \sum_{i=1}^{N} \boldsymbol{x}_i \frac{\partial}{\partial \boldsymbol{w}} \sigma(\boldsymbol{x}_i^T \boldsymbol{w}) + \Sigma_p^{-1} = \\
&= \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T [1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})] \sigma(\boldsymbol{x}_i^T \boldsymbol{w}) + \Sigma_p^{-1}
\end{aligned} \tag{20}
$$

Denote

$$
\mathcal{X} \in \mathbb{R}^{d \times N}, \mathcal{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N] \tag{21}
$$

$$
\mathcal{D} \in \mathbb{R}^{N \times N}, \mathcal{D} = diag(d_1, d_2, ..., d_N),\ d_i = [1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})] \sigma(\boldsymbol{x}_i^T \boldsymbol{w})] \tag{22}
$$

Thus (20) can be expressed as

$$
H = \mathcal{X} D \mathcal{X}^T + \Sigma_p^{-1}
$$