Laplace Approximation for Bayesian Logistic Regression: A Derivation

Wei Dai

March 13, 2013

This derivation is based on Bishop (2006). Given data set $\{\boldsymbol{\phi}_n, t_n\}_{n=1}^N$ where $\boldsymbol{\phi}_n$ are the feature vectors and $t_n \in \{0, 1\}$ are the labels, we can write the likelihood function for logistic regression as

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n} \tag{1}$$

where $\boldsymbol{t} = (t_1, ..., t_N)^T$ and $y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n) = \sigma(\boldsymbol{w}^T\boldsymbol{\phi}_n)$ and $\sigma(s) = \frac{1}{1+e^{-s}}$. Using Bayes rule, the posterior distribution over $\boldsymbol{w}$ is

$$p(\boldsymbol{w}|\boldsymbol{t}) = \frac{p(\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})}{p(\boldsymbol{t})} \tag{2}$$

where $p(\boldsymbol{t}) = \int p(\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})d\boldsymbol{w}$ involves logistic sigmoid functions and is intractable. Laplace approximation approximate this posterior with a multivariate Guassian:

$$\begin{aligned} q(\boldsymbol{w}) &= \frac{1}{(2\pi)^{M/2}|\boldsymbol{S}_N|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MAP})^T\boldsymbol{S}_N^{-1}(\boldsymbol{w} - \boldsymbol{w}_{MAP})\right\} \\ &= \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{MAP}, \boldsymbol{S}_N) \end{aligned} \tag{3}$$

where $\boldsymbol{w}_{MAP}$ is the maximum *a posteriori* and thus a mode of the posterior and $\boldsymbol{S}_N^{-1} = -\nabla^2_{\boldsymbol{w}} \ln p(\boldsymbol{w}|\boldsymbol{t})|_{\boldsymbol{w}=\boldsymbol{w}_{MAP}}$ is the Hessian at $\boldsymbol{w}_{MAP}$. Since we are approximating the posterior with Gaussian, it is convenient to use conjugate prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{m}_0, \boldsymbol{S}_0)$. Thus we have

$$\ln p(\boldsymbol{w}|\boldsymbol{t}) = -\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)+\sum_{n=1}^{N}\{t_n \ln y_n+(1-y_n)\ln(1-y_n)\}+const \tag{4}$$

Since eq. 4 is convex in $\boldsymbol{w}$ (?), we can use gradient descent to find $\boldsymbol{w}_{MAP} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{t}) = \arg\max_{\boldsymbol{w}} \ln p(\boldsymbol{w}|\boldsymbol{t})$. Using the following facts:

$$\nabla_{\boldsymbol{x}} \boldsymbol{m}^T \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A}^T \boldsymbol{x} \tag{5}$$

$$\nabla_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{m} = \boldsymbol{A} \boldsymbol{x} \tag{6}$$

$$\nabla_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A} \boldsymbol{x} + \boldsymbol{A}^T \boldsymbol{x} \tag{7}$$

$$\nabla_{\boldsymbol{x}} \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A} \tag{8}$$

$$\frac{\partial}{\partial s} \sigma(s) = \sigma(s)(1 - \sigma(s)) \tag{9}$$

where $\boldsymbol{x}, \boldsymbol{m} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{d \times d}$. Also note that ${\boldsymbol{S}_0^{-1}} = {\boldsymbol{S}_0^{-T}}$, we arrive at

$$\nabla_{\boldsymbol{w}} \ln p(\boldsymbol{w}|\boldsymbol{t}) = -{\boldsymbol{S}_0^{-1}}(\boldsymbol{w} - \boldsymbol{m}_0) + \sum_{n=1}^{N} (t_n - y_n) \boldsymbol{\phi}_n \tag{10}$$

Since we generally want small $||\boldsymbol{w}||$ to avoid overfitting, we use $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{S}_0 = \sigma^2 \boldsymbol{I}$. We have the following update rule:

$$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} + \eta \left(\sum_{n=1}^{N} (t_n - y_{n,(t-1)}) \boldsymbol{\phi}_n - \frac{1}{\sigma^2} \boldsymbol{w}_{t-1} \right) \tag{11}$$

where η is the learning rate constant. We can also get

$$\boldsymbol{S}_N^{-1} = -\nabla^2_{\boldsymbol{w}} \ln p(\boldsymbol{w}|\boldsymbol{t}) = {\boldsymbol{S}_0^{-1}} + \sum_{n=1}^{N} y_n (1 - y_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \tag{12}$$

Thus, we have the approximated posterior $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{\boldsymbol{MAP}}, \boldsymbol{S}_N)$.