Variational Bayesian Logistic Regression

Sargur N. Srihari
University at Buffalo, State University of New York
USA

Topics in Linear Models for Classification

- Overview
- 1. Discriminant Functions
- 2. Probabilistic Generative Models
- 3. Probabilistic Discriminative Models
- 4. The Laplace Approximation
- 5.1 Bayesian Logistic Regression
- 5.2 Variational Bayesian Logistic Regression

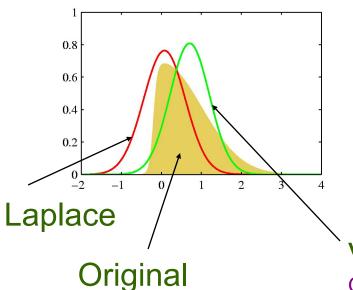
Topics in Variational Bayesian Logistic Regression

- Bayesian Logistic Regression Posterior
- K-L Divergence
- Variational Inference
- Stochastic Variational Inference

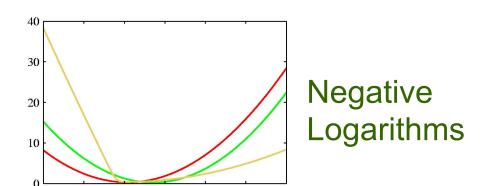
Variational Logistic Regression

- Variational: Based on calculus of variations
 - How a derivative of a functional changes
 - Functional takes function as input and returns a value
- Approximation is more flexible than Laplace with additional variational parameters

Example of Variational Method



distribution



Variational distribution is Gaussian

Optimized with respect to mean and variance

Bayesian Logistic Regression Posterior

- Data: $D = \{(\mathbf{x}^{(1)}, t^1), ..., (\mathbf{x}^{(N)}, t^N)\}$ N i.i.d.samples $t \in \{0, 1\}$
- Parameters: $\boldsymbol{w} = \{w_1,...,w_M\}$
 - Probabilistic model specifies the joint distribution $p(D, \mathbf{w}) = p(D/\mathbf{w})p(\mathbf{w})$
 - Which is a product of sigmoid likelihood and a Gaussian:

$$p(D \mid \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{t_n} \left\{ 1 - y_n \right\}^{1 - t_n} \quad y_n = p(C_1 \mid \boldsymbol{x}^{(n)}) = \sigma(\boldsymbol{w}^T \boldsymbol{x}^{(n)}) \quad p(\boldsymbol{w}) = N(\boldsymbol{w} \mid m_0, S_0)$$

- Goal is to approximate posterior distribution $p(\boldsymbol{w}|D)$
 - which is also a normalized product of sigmoid likelihood and Gaussian $p(w \mid D) = \frac{p(D,w)}{p(D)} = \frac{p(D/w)p(w)}{p(D)}$
- with a Gaussian q(w), so we can use it in prediction

$$p(C_1 \mid \boldsymbol{x}) \simeq \int \boldsymbol{\sigma}(\boldsymbol{w}^T \boldsymbol{x}) q(\boldsymbol{w}) d\boldsymbol{w}$$

Srihari

Decomposition of Log Marginal Probability

$$\ln p(D) = L(q) + KL(q \mid\mid p)$$

$$where$$

$$L(q) = \int q(\boldsymbol{w}) \ln \left\{ \frac{p(D, \boldsymbol{w})}{q(\boldsymbol{w})} \right\} d\boldsymbol{w}$$

$$and$$

$$KL\{q \mid\mid p\} = -\int q(\boldsymbol{w}) \ln \left\{ \frac{p(\boldsymbol{w} \mid D)}{q(\boldsymbol{w})} \right\} d\boldsymbol{w}$$

We wish to maximize $\ln p(D)$ called evidence, by suitable choice of q

We want to minimize K-L Divergence over a family for q(w)

Which finds distribution $q(\mathbf{w})$ that best approximates $p(\mathbf{w}|D)$

Some Observations:

Lower bound on $\ln p(D)$ is L(q)

Maximizing the lower bound L(q) wrt distribution q(w)

is equivalent to *minimizing* KL Divergence

When KL divergence vanishes q(w) equals the posterior p(w|D)

Plan:

We seek that distribution q(w) for which L(q) is largest

We consider restricted family for q(w)

Seek member of this family for which KL divergence is minimized

Variational Inference

• We want to form our Gaussian approximation $q(\mu,\Sigma)=N(\boldsymbol{w};\mu,\Sigma)$ to the posterior $p(\boldsymbol{w}|D)$ by minimizing the KL divergence

$$\left| D_{\!\scriptscriptstyle KL} \! \left(q(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, || \; p(\boldsymbol{w} \, | \, D) \right) = E_{N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})} \! \left[\log \frac{N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{p(\boldsymbol{w} \, | \, D)} \right] \right|$$

- We fit the variational parameters μ and Σ , not the original parameters $\textbf{\textit{w}}$
 - Although there is an interpretation that μ is an estimate of $\textbf{\textit{w}}$, while Σ indicates a credible range of where the parameters could plausibly be around this estimate.

Criterion to Minimize

 As we can only evaluate the posterior up to a constant, we write

$$D_{\mathit{KL}}\left(q\mid\mid p\right) = E_{N(\boldsymbol{w};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\log\frac{N(\boldsymbol{w};\boldsymbol{\mu},\boldsymbol{\Sigma})}{p(\boldsymbol{w})p(D|\boldsymbol{w})}\right] + \log p(D)$$

where we have used $p(\boldsymbol{w}|D)=p(\boldsymbol{w},D)/p(D)$

Srihari

- And minimize J
- Because $D_{\text{KL}} \ge 0$ we obtain a lower bound on the model likelihood $\log p(D) \ge J$,
 - -J is called evidence lower bound (ELBO)
- For logistic regression model and prior

$$\left|J = E_{N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})} \bigg[\log N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{w}) - \sum_{n=1}^{N} \sigma(\boldsymbol{w}^{T} \boldsymbol{x}^{(n)}) \bigg] \right|$$

which we wish to minimize

Method of Optimization

$$\left|J = E_{N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[\log N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{w}) - \sum_{n=1}^{N} \log \sigma(\boldsymbol{w}^T \boldsymbol{x}^{(n)}) \right] \right|$$

- We could evaluate this expression numerically
 - The first two expectations can be computed analytically, and the remaining N terms in the sum can be reduced to a one-dimensional integral
- We could similarly evaluate the derivatives wrt μ and Σ , and fit the variational parameters with a gradient-based optimizer
 - However the inner loop each function evaluation would require N numerical integrations, or further approximation (since it expectation of a summation)

Stochastic Variational Inference

 We can avoid numerical integration by a simple Monte Carlo estimate of the expectation

$$\left| J \approx \frac{1}{S} \sum_{s} \left[\log N(\boldsymbol{w}^{(s)}; \mu, \Sigma) - \log p(\boldsymbol{w}^{(s)}) - \sum_{n=1}^{N} \log \sigma(\boldsymbol{w}^{(s)T} \boldsymbol{x}^{(n)}) \right], \boldsymbol{w}^{(s)} \sim N(\mu, \Sigma) \right|$$

- Really cheap estimate is to use one sample
 - Further form an unbiased estimate of the sum over data points by randomly picking one term

$$\boxed{J \approx \Bigl[\log N(\boldsymbol{w}^{(s)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{w}^{(s)}) - N \log \sigma(\boldsymbol{w}^{(s)T}\boldsymbol{x}^{(n)})\Bigr], n \sim \text{Uniform}[1, ..N], \boldsymbol{w}^{(s)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\Bigr]}$$

Removing dependence on μ and Σ

- The goal is to move the variational parameters μ and Σ , which express which weights are plausible, so that cost J gets smaller (on average)
 - If the variational parameters change, the samples we would draw change, which complicates our reasoning.
- We can remove the variational parameters from the random draws by writing down how a Gaussian random generator works: $\mathbf{w}^{(s)} = (\mu + \Sigma^{1/2} \mathbf{v})$,
 - where $\nu \sim N(0,1)$, $\Sigma^{1/2}$ is a matrix square root, such as Cholesky decomposition. Then our cost function becomes

• Where all dependence on variational parameters μ and Σ are now in the estimator itself

Minimizing cost function J

- We can minimize the cost function J by stochastic gradient descent. We draw a datapoint n at random, some Gaussian white noise ν , and evaluate J and its derivatives wrt μ and Σ
- We then make a small change to μ and Σ to improve the current estimate of J and repeat.

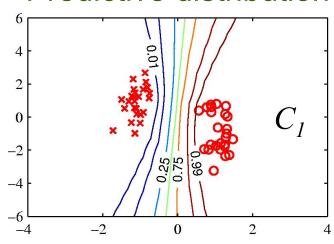
Applicability to neural networks

- While some of the fine details are slightly complicated, none of them depend on the fact we were considering logistic regression.
- We could easily replace the logistic function $\sigma(.)$ with another model likelihood, such as from a neural network. As long as we can differentiate the log-likelihood, we can apply stochastic variational inference.

Example of Variational Bayes Logistic Regression

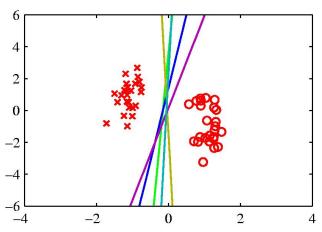
Linearly separable data set

Predictive distribution



 $p(C_I|\phi)$ is plotted from 0.01 to 0.99

Decision boundaries



Corresponding to five samples of parameter vector w

Large margin solution of SVM has qualitatively similar ₁₅ result to the Bayesian solution