The Laplace Approximation

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

Topics in Linear Models for Classification

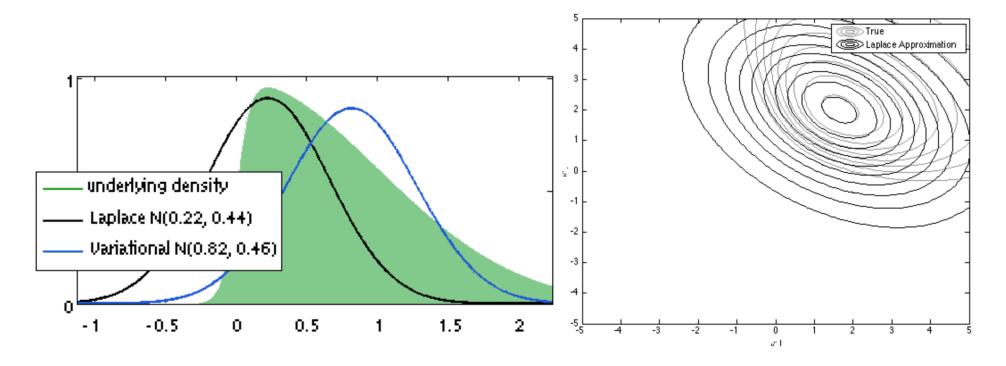
- Overview
- 1. Discriminant Functions
- 2. Probabilistic Generative Models
- 3. Probabilistic Discriminative Models
- 4. The Laplace Approximation
- 5. Bayesian Logistic Regression

Topics in Laplace Approximation

- Motivation
- Finding a Gaussian approximation: 1-D case
- Approximation in M-dimensional space
- Weakness of Laplace approximation
- Model Comparison using BIC

What is Laplace Approximation?

 The Laplace approximation framework aims to find a Gaussian approximation to a continuous distribution



Why study Laplace Approximation?

- We shall discuss Bayesian treatment of logistic regression
 - It is more complex than the Bayesian treatment of linear regression
 - In particular we cannot integrate exactly
- In ML we need to predict a distribution of the output that may involve integration

Bayesian Linear Regression

- Recapitulate:
 - in Bayesian *linear* regression we integrate exactly:

Gaussian noise:

$$p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = N(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1})$$

Gaussian prior:

$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_{\theta}, S_{\theta})$$

 $S_{\theta} = \alpha^{I} I$

Gaussian posterior:

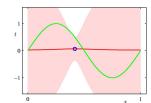
$$p(\boldsymbol{w}|\mathbf{t}) = N(\boldsymbol{w}|\boldsymbol{m}_N, S_N)$$

$$m_N = \beta S_N \Phi^T t$$

 $S_N^{-1} = \alpha I + \beta \Phi^T \Phi$

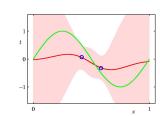
Predictive Distribution::

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \mathbf{w}, \beta) \cdot p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) d\mathbf{w}$$

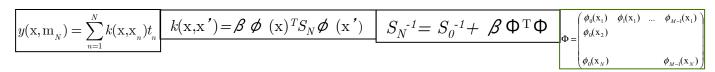


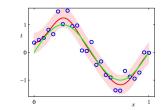
Result of integration:

$$\left| p(t \mid \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = N(t \mid \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})) \quad \text{where} \quad \sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T S_N \boldsymbol{\phi}(\boldsymbol{x}) \right|$$



Equivalent Kernel





Bayesian Logistic Regression

- In Bayesian treatment of Logistic regression
 - we cannot directly integrate over the parameter vector \boldsymbol{w} since the posterior is not Gaussian
- It is therefore necessary to introduce some form of approximation
- Later we will consider analytical approximations and numerical sampling

Approximation due to intractability

- Bayesian logistic regression
 - Prediction $p(C_k|\mathbf{x})$ involves integrating over w

$$p(C_1 | \phi, \mathbf{t}) \simeq \int \sigma(\mathbf{w}^T \phi) p(\mathbf{w}) d\mathbf{w}$$

- Convolution of Sigmoid-Gaussian is intractable
 - It is not Gaussian-Gaussian as in linear regression
- Need to introduce methods of approximation
- Approaches
 - Analytical Approximations
 - Laplace Approximation
 - Numerical Sampling

Laplace approximation framework

- Simple but widely used framework
- Aims to find a Gaussian approximation to a probability density defined over a set of continuous variables
- Method aims specifically at problems in which the distribution is unimodal
- Consider first the case of single continuous variable

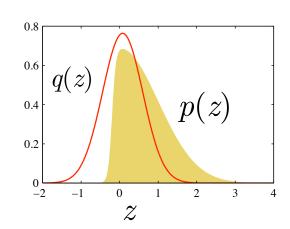
Laplace Approximation: 1D case

Single continuous variable z with distribution

p(z) defined by

$$p(z) = \frac{1}{Z} f(z)$$
 where $Z = \int f(z) dz$ is a normalization coefficient

Value of Z is unknown
 f(z) is a scaled version of p(z)



- Goal is to find Gaussian approximation q(z) which is centered on the mode of the distribution p(z)
- First step is to find mode of p(z)
 - i.e., a point z_{θ} such that $p'(z_0)=0$

Srihar

Quadratic using Taylor's series

• A Taylor's series expansion of f(x) centered on the mode x_0 has the power series

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!} \Big(x - x_0 \Big) + \frac{f''(x_0)}{2!} \Big(x - x_0 \Big)^2 + \frac{f^{(3)}(x_0)}{3!} \Big(x - x_0 \Big)^3 + \dots$$

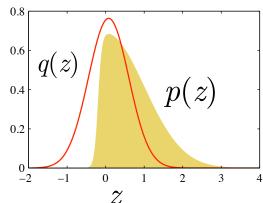
- $-f\left(x\right)$ is assumed infinitely differentiable at x_{0}
- Note that when x_0 is the maximum $f'(x_0)=0$ and that term disappears
- A Gaussian f(x) has the property that $\ln f(x)$ is a quadratic function of x
 - We will use a Taylor's series approximation of the function $\ln f(z)$ and use only the quadratic term

Animation

Finding mode of a distribution

- Find point z_0 such that $p'(z_0)=0$,

- Or equivalently
$$\left| \frac{df(z)}{dz} \right|_{z=z_0} = 0$$



- Approximate f(z) using 2nd derivative $p''(z_0)$
 - Logarithm of Gaussian is a quadratic.
 - Use Taylor expansion of $\ln f(z)$ centered at mode z_0

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}A(z - z_0)^2$$
where $A = -\frac{d^2}{dz^2} \ln f(z) \Big|_{z=z_0}$

A is the second derivative of logarithm of scaled p(z)

First order term does not appear since z_0 is a local maximum and $f'(z_0) = 0$

Final form of Laplacian (one dim)

• Approximation of f(z):
Assuming $\ln f(z)$ to be quadratic

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}A(z - z_0)^2$$
where $A = -\frac{d^2}{dz^2} \ln f(z) \Big|_{z=z_0}$

Taking exponential

$$f(z) \approx f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

• To Normalize f(z) we need Z

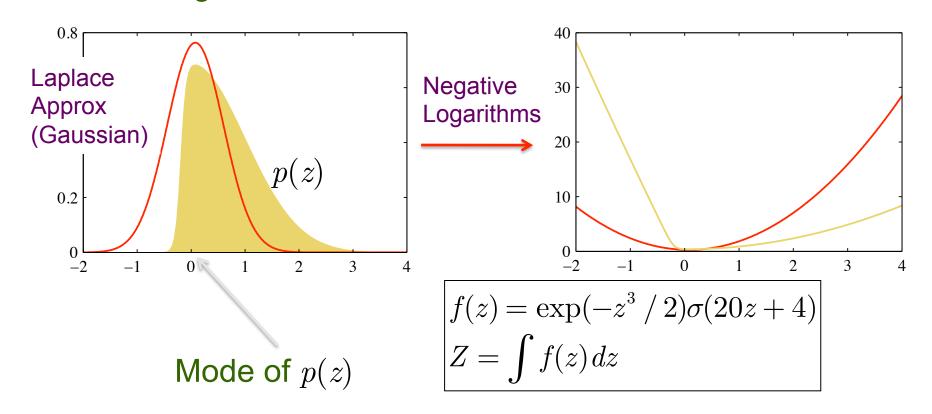
$$Z = \int f(z) dz \approx f(z_0) \int \exp\left\{-\frac{A}{2}(z - z_0)^2\right\} dz = f(z_0) \frac{(2\pi)^{1/2}}{A^{1/2}}$$

•The normalized version of f(z) is

$$q(z) = \frac{1}{Z}f(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z - z_0)^2\right\} \sim N(z \mid z_0, A^{-1})$$

Laplace Approximation Example

Applied to distribution $p(z)\alpha \exp(-z^2/2)\sigma(20z+4)$ where σ is sigmoid



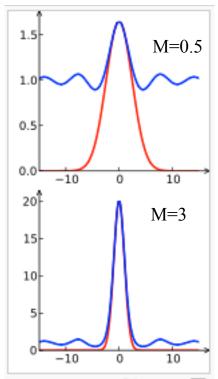
Gaussian approximation will only be well-defined if its precision A>0, or second derivative of f(z) at point z_0 is negative

Laplace's Method

- Approximate integrals of the form $\int_{a}^{b} e^{Mf(x)} dx$
 - Assume f(x) has global maximum at x_0
 - Then $f(x_0) >>$ other values of f(x)
 - with $e^{Mf(x)}$ growing exponentially with M
 - So enough to focus on f(x) at x_0
- As M increases, integral is wellapproximated by a Gaussian

$$\int_a^b e^{Mf(x)} dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}} e^{Mf(x_0)} \text{ as } M \to \infty.$$

- Second derivative appears in denominator
- Proof involves Taylor's series expansion of f(x) at x_0



The function $e^{Mf(x)}$, in blue, is shown on top for M = 0.5, and at the bottom for M = 3. Here, $f(x) = \sin x/x$, with a global maximum at $x_0 = 0$. It is seen that as M grows larger, the approximation of this function by a Gaussian function (shown in red) is getting better. This observation underlies Laplace's method.

Laplace Approx: M-dimensions

- Task: approximate p(z)=f(z)/Z defined over M-dim space z
- At stationary point z₀ the gradient ∇f(z) vanishes
- Expanding around this point

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} (z - z_0)^T A (z - z_0)$$

– where A is the $M \times M$ Hessian matrix

$$A = -\nabla\nabla \ln f(\mathbf{z}) \, |_{\mathbf{z} = \mathbf{z}_0}$$

Taking exponentials

$$f(z) \simeq f(z_0) \exp \left\{ -\frac{1}{2} (z - z_0)^T A (z - z_0) \right\}$$

Normalized Multivariate Laplacian

$$Z = \int f(z) dz$$

$$\approx f(z_0) \int \exp\left\{-\frac{1}{2}(z - z_0)^T A(z - z_0)\right\} dz$$

$$= f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}}$$

• Distribution q(z) is proportional to f(z) as

$$q(z) = \frac{1}{Z} f(z) = \frac{|A|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2} (z - z_0)^T A (z - z_0)\right\}$$
$$= N(z|z_0, A^{-1})$$

Steps in Applying Laplace Approx.

- 1. Find the mode z_0
 - Run a numerical optimization algorithm
- 2. Evaluate Hessian matrix A at that mode

- Many distributions encountered in practice are multimodal
 - There will be different approximations according to which mode considered
 - Z of the true distribution need not be known to apply Laplace method

Weakness of Laplace Approx.

- Directly applicable only to real variables
 - Based on Gaussian distribution
- May be applicable to transformed variable
 - If $0 \le \tau < \infty$ then consider Laplace approx of $\ln \tau$
- Based purely on a specific value of the variable
- Variational methods have a more global perspective

Approximating Z

• As well as approximating the distribution p(z) we can also obtain an approximation to the normalizing constant Z

$$Z = \int f(z) dz$$

$$\simeq f(z_0) \int \exp\left\{-\frac{1}{2}(z - z_0)^T A(z - z_0)\right\} dz$$

$$= f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}}$$

 Can use this result to obtain an approximation to model evidence that plays a central role in Bayesian model comparison

Model Comparison and BIC

- Consider data set D and models $\{M_i\}$ having parameters $\{\theta_i\}$
- For each model define likelihood $p(D|\theta_i, M_i)$
- Introduce prior $p(\theta_i|M_i)$ over parameters
- Need model evidence $p(D|M_i)$ for various models

Model Comparison and BIC

For a given model from sum rule

$$p(D) = \int p(D \mid \theta) p(\theta) d\theta$$

Identifying $f(\theta) = p(D|\theta)p(\theta)$ and Z = p(D) and using

$$Z = f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}}$$

$$\ln p(D) = \ln \int f(\theta) d\theta$$

$$= \ln \left[f(\theta_{map}) \frac{(2\pi)^{M/2}}{|A|^{1/2}} \right]$$

$$= \ln p(D|\theta_{map}) + \ln p(\theta_{map}) + \frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |A|$$

Occam factor that penalizes model complexity

where θ_{map} is the value of θ at the mode of the posterior A is Hessian of second derivatives of negative log posterior

Bayes Information Criterion (BIC)

Assuming broad Gaussian prior over parameters & Hessian is of full rank, approximate model evidence is

$$\ln p(D) \approx \ln p(D|\theta_{\text{map}}) - \frac{1}{2}M \ln N$$

- N is the number of data points
- M is the no of parameters in θ
- Compared to AIC given by $\ln p(D|\boldsymbol{w}_{ML})$ -M BIC penalizes model complexity more heavily

Weakness of AIC, BIC

- AIC and BIC are easy to evaluate
- But can give misleading results since
 - Hessian matrix may not have full rank since many parameters not well-determined
- Can obtain more accurate estimate from

$$\ln p(D) = \ln p(D|\theta_{\text{map}}) + \ln p(\theta_{\text{map}}) + \frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |A|$$

Used in the context of neural networks

Summary

- Bayesian approach for logistic regression is more complex than for linear regression
 - Since posterior over w is not Gaussian
- Predictive Distribution needs an integral over parameters
 - Simplified when Gaussian
- Laplace approximation fits the best Gaussian
 - Defined for both univariate and multivariate
- Normalization term is useful as BIC criterion
- AIC and BIC are simple but not accurate