CSCI 590: Machine Learning

Lecture 12:

Laplace approximation, model comparison and BIC, Bayesian logistic regression

Instructor: Murat Dundar

Acknowledgement:
These slides are prepared using the course textbook
http://research.microsoft.com/~cmbishop/prml/

Laplace Approximation (1)

If p(z) is a posterior distribution in the form of

$$p(z) = \frac{1}{Z} f(z)$$

and if no analytical solution exists for $Z = \int f(z)dz,$ a closed form solution for p(z) is not possible.

Laplace approximation aims to find a Gaussian approximation q(z) to a probability density p(z) centered on the mode of p(z).

The first step is to find a mode of p(z), i.e., a point z_0 such that $p'(z_0) = 0.$

Laplace Approximation (2)

$$p'(z_0) = 0 \rightarrow f'(z_0) = 0$$

The logarithm of a Gaussian distribution is a quadratic function of its variables. Therefore a second order Taylor expansion of $\ln f(z)$ centered on the mode z_0 may offer a good approximation.

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

where $A = -\frac{d^2}{dz^2} lnf(z)|_{z=z_0}$.

The first order term in the Taylor expansion vanishes as $f'(z_0) = 0$

Laplace Approximation (3)

Taking the exponential

$\exp(\ln f(z)) \cong \exp\left(\ln f(z_0) - \frac{1}{2}A(z-z_0)^2\right)$ we obtain

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

Now that f(z) is approximated a by a Gaussian shaped function to obtain a standard Gaussian distribution we normalize f(z) to get q(z)

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2}\exp\left\{-\frac{1}{2}A(z-z_0)^2\right\}$$

Laplace Approximation (4)

When **z** is multivariate **A** becomes a matrix

$$q(z) = \left(\frac{|\boldsymbol{A}|}{(2\pi)^d}\right)^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)^T \boldsymbol{A}(\boldsymbol{z}-\boldsymbol{z}_0)\right\} = N(\boldsymbol{z}|\,\boldsymbol{z}_0,\boldsymbol{A}^{-1})$$

When the distribution is multimodal there will be different Laplace approximations according to which mode is being considered.

Laplace approximation is based purely on the aspects of the true distribution at a specific value of the variable, and so can fail to capture important global properties.

Model Comparison and BIC (1)

Recall Bayes Factor from Lecture 8.

Bayes factor: $\frac{p(D|M_i)}{p(D|M_j)}$

Earlier we approximated model evidence $p(D|M_i)$ by

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}.$$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \underbrace{\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)}_{\text{Negative}}.$$

Model Comparison and BIC (2)

$$p(D) = \int \underbrace{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}_{f(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}$$

$$f(\boldsymbol{\theta}) \cong f(\boldsymbol{\theta}_{MAP})\exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})\right\}$$

If we replace the approximation for $f(\boldsymbol{\theta})$ with the integrand in p(D) then we get

$$p(D) \cong f(\boldsymbol{\theta}_{MAP})\left(\frac{|\boldsymbol{A}|}{(2\pi)^d}\right)^{-\frac{1}{2}}$$

$$\ln p(D) = \ln p(D|\boldsymbol{\theta}_{MAP}) + \underbrace{\ln p(\boldsymbol{\theta}_{MAP}) - \frac{1}{2}\ln\frac{|\boldsymbol{A}|}{(2\pi)^d}}_{\boldsymbol{Occam\ factor}}$$

Model Comparison and BIC (3)

The first term is log likelihood evaluated using the MAP parameters while the remaining two terms comprise the "Occam factor" which penalizes model complexity.

Assume: a broad Gaussian prior over $p(\boldsymbol{\theta})$, and a full rank Hessian $\ln p(D)$ can be further approximated by

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2} M \ln N$$

where *N* is the number of data points, *M* is the number of parameters. This is known as the *Bayesian Information Criterion* (BIC) or the *Schwarz criterion* (Schwarz, 1978).

Bayesian Logistic Regression (1)

Exact Bayesian inference for logistic regression is intractable.

Evaluation of the posterior distribution would require normalization of the product of a prior distribution and a likelihood function that itself comprises a product of logistic sigmoid functions, one for every data point. Evaluation of the predictive distribution is similarly intractable.

We can consider Laplace approximation to the posterior.

Bayesian Logistic Regression (3)

$$p(\boldsymbol{w}|t) \propto p(\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})$$

$$p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \left\{1 - y_n\right\}^{1-t_n}$$

Taking the log of both sides

$$\begin{aligned}\ln p(\mathbf{w}|\mathbf{t}) \quad \propto \quad & -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ & + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}\end{aligned}$$

We maximize this function with respect to **w** to obtain $\boldsymbol{w}_{MAP}$

Bayesian Logistic Regression (3)

The Hessian matrix is equal to:

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

The Gaussian approximation to the posterior takes the form:

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)$$

To make predictions about future data we will integrate out **w** to obtain the predictive distribution

$$p(\mathcal{C}_1|\phi, \mathbf{t}) = \int p(\mathcal{C}_1|\phi, \mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\phi)q(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

Bayesian Logistic Regression (4)

This integral represents the convolution of a Gaussian with a logistic sigmoid, and cannot be evaluated analytically.

We can approximate the logistic function by the inverse probit function, i.e., we approximate $\sigma(a)$ by $\Phi(\lambda a)$

The advantage of using an inverse probit function is that its convolution with a Gaussian can be expressed analytically in terms of another inverse probit function.