Bayesian Machine Learning
ORIE 6741

Fall 2016
Tu/Th 11:40 am - 12:55 pm
Room: Rhodes Hall 571
https://people.orie.cornell.edu/andrew/orie6741

Lecture 2 Supplement
August 25, 2016

Marginal Likelihood of Bayesian Basis Regression Model

Let

$$y = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) + \epsilon\,, \tag{1}$$

$$\epsilon|\sigma^2 \sim \mathcal{N}(0, \sigma^2)\,, \tag{2}$$

$$\mathbf{w}|\alpha \sim \mathcal{N}(\mathbf{w}; 0, \alpha^2 I)\,. \tag{3}$$

Then, using the sum and product rules of probability,

$$p(\mathbf{y}|\alpha, \sigma) = \int p(\mathbf{y}|\mathbf{w}, \sigma^2), p(\mathbf{w}|\alpha) d\mathbf{w}\,. \tag{4}$$

We can solve this integral, (through standard Gaussian identities, or by completing the square and integrating a quadratic form in $\mathbf{w}$), to find the *marginal likelihood*

$$\log p(\mathbf{y}|\alpha, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{m}{2}\log\alpha^2 - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|A| - E(\mathbf{m}_N)\,, \tag{5}$$

$$E(\mathbf{m}_N) = \frac{1}{2\sigma^2}||\mathbf{y} - \Phi\mathbf{m}_N||^2 + \frac{1}{2\sigma^2}\mathbf{m}_N^\top\mathbf{m}_N\,, \tag{6}$$

$$A = \frac{I}{\alpha^2} + \frac{1}{\sigma^2}\Phi^\top\Phi = \nabla\nabla E(\mathbf{w})\,, \tag{7}$$

$$\mathbf{m}_N = \frac{1}{\sigma^2}A^{-1}\Phi^\top\mathbf{y}\,. \tag{8}$$

The $\log|A|$ term can be thought of as a *complexity penalty* and the $E(\mathbf{m}_N)$ term as a *model fit* term. Because we have marginalised the parameters $\mathbf{w}$, there will be an automatic calibration between model fit and model complexity when learning the parameters α, σ through optimisation. A fully Bayesian approach would place a prior distribution over these parameters and also marginalise them away.

Exponential Family Distributions

The exponential family of distributions over $\mathbf{x}$ given a set of parameters $\boldsymbol{\theta}$ is defined to be the set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^\top\mathbf{u}(\mathbf{x})\right]\,, \tag{9}$$

where $\mathbf{x}$ may be a scalar or vector, and may be discrete or continuous. Here $\boldsymbol{\theta}$ are called the natural parameters of the distribution, and $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$. The function $g(\boldsymbol{\theta})$ can be interpreted as a coefficient that ensures that the distribution is normalised and therefore satifies

$$g(\boldsymbol{\theta}) \int h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^\top \mathbf{u}(\mathbf{x})\right] d\mathbf{x} = 1\,. \tag{10}$$

Exercise: Can you show that the Bernoulli distribution is a member of the exponential family?

Conjugate priors

We seek a prior $p(\boldsymbol{\theta})$ which is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior. That is,

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\,, \tag{11}$$

and we want to find a $p(\boldsymbol{\theta})$ with the same functional form as $p(\boldsymbol{\theta}|\mathbf{x})$ when we multiply against the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$.

For any member of the exponential family, there exists a conjugate prior of the form

$$p(\boldsymbol{\theta}|\mathbf{z}, \nu) = f(\mathbf{z}, \nu) g(\mathbf{x})^\nu \exp\left[\nu \boldsymbol{\theta}^\top \mathbf{z}\right]\,, \tag{12}$$

where $f(\mathbf{z}, \nu)$ is a normalisation coefficient, and $g(\boldsymbol{\theta})$ is the same as in Eq. (9).

To see that this prior function is conjugate, multiply Eq. (12) with Eq. (9), to find

$$p(\boldsymbol{\theta}|X, \mathbf{z}, \nu) \propto g(\boldsymbol{\theta})^{\nu+N} \exp\left[\boldsymbol{\theta}^\top \left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu \mathbf{z}\right)\right]\,. \tag{13}$$

Here N are the number of datapoints and $\mathbf{x}_n$ represents each data point.