CSC 2541: Bayesian Methods for Machine Learning

Radford M. Neal, University of Toronto, 2011

Lecture 6

A Bayesian Linear Basis Function Model

Let's set up a Bayesian linear basis function model by giving β a Gaussian prior:

$$\begin{aligned} y_i \,|\, x_i,\, \beta &\sim N(\phi(x_i)^T\beta,\, \sigma^2) \\ \beta &\sim N(m_0, S_0) \end{aligned}$$

This Gaussian prior will turn out to be conjugate.

For the moment, we regard σ^2, m_0, and S_0 as known.

Often, we will let $m_0 = 0$ and let S_0 be diagonal, so that the β_j are independent. We might let β_0 have a large variance, and all the other β_j have the same variance.

The symbol y will sometime denote a single, generic response value, and other times denote the vector $[y_1, \ldots, y_n]^T$ of responses for training cases. We use Φ for the matrix of basis function values for the n training cases.

Multivariate Gaussian Model with Multivariate Gaussian Prior

To warm up... Suppose we model an observed vector b as having a multivariate Gaussian distribution with known covariance matrix B and unknown mean x. We give x a multivariate Gaussian prior with known covariance matrix A and known mean a.

The posterior distribution of x will be Gaussian, since the product of the prior density and the likelihood is proportional to the exponential of a quadratic function of x:

$$\text{Prior} \times \text{Likelihood} \quad \propto \quad \exp(-(x-a)^T A^{-1}(x-a)/2)\ \exp(-(b-x)^T B^{-1}(b-x)/2)$$

The log posterior density is this quadratic function ($\cdots$ is parts not involving x):

$$\begin{aligned} -\tfrac{1}{2}\Big[(x-a)^T A^{-1}(x-a) \ +\ (b-x)^T B^{-1}(b-x)\Big] \ +\ \cdots \\ =\ -\tfrac{1}{2}\Big[x^T(A^{-1}+B^{-1})x \ -\ 2x^T(A^{-1}a+B^{-1}b)\Big] \ +\ \cdots \\ =\ -\tfrac{1}{2}\Big[(x-c)^T(A^{-1}+B^{-1})(x-c)\Big] \ +\ \cdots \end{aligned}$$

where $c = (A^{-1}+B^{-1})^{-1}\,(A^{-1}a+B^{-1}b)$. This is the density for a Gaussian distribution with mean c and variance $(A^{-1}+B^{-1})^{-1}$.

Posterior for Linear Basis Function Model

Both the log prior and the log likelihood are quadratic functions of β. The log likelihood for β is

$$-\frac{1}{2}\Big[(y-\Phi\beta)^T(\sigma^2 I)^{-1}(y-\Phi\beta)\Big] \;+\; \cdots \;\;=\;\; -\frac{1}{2}\frac{1}{\sigma^2}\Big[\beta^T\Phi^T\Phi\beta \;-\; 2\beta^T\Phi^T y\Big] \;+\; \cdots$$

which is the same quadratic function of β as for a Gaussian log density with covariance $\sigma^2(\Phi^T\Phi)^{-1}$ and mean $(\Phi^T\Phi)^{-1}\Phi^T y$.

This combines with the prior for β in the same way on the previous slide, with the result that the posterior distribution for β is Gaussian with covariance

$$S_n \;=\; \Big[S_0^{-1} \,+\, (\sigma^2(\Phi^T\Phi)^{-1})^{-1}\Big]^{-1} \;=\; \Big[S_0^{-1} \,+\, (1/\sigma^2)\Phi^T\Phi\Big]^{-1}$$

and mean

$$\begin{aligned} m_n \;&=\; (S_n^{-1})^{-1}\Big[S_0^{-1}m_0 \,+\, (1/\sigma^2)\Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^T y\Big] \\ &=\; S_n\Big[S_0^{-1}m_0 \,+\, (1/\sigma^2)\Phi^T y\Big] \end{aligned}$$

Predictive Distribution for a Test Case

We can write the response, y, for some new case with inputs x as

$$y \;=\; \phi(x)^T\beta \,+\, e$$

where the "noise" e has the $N(0,\sigma^2)$ distribution, independently of β.

Since the posterior distribution for β is $N(m_n,\, S_n)$, the posterior distribution for $\phi(x)^T\beta$ will be $N(\phi(x)^T m_n,\, \phi(x)^T S_n \phi(x))$.

Hence the predictive distribution for y will be $N(\phi(x)^T m_n,\, \phi(x)^T S_n \phi(x) + \sigma^2)$.

Comparison with Regularized Estimates

The Bayesian predictive mean for a test case is what we would get using the posterior mean value for the regression coefficients — a consequence of the model being linear in the parameters.

We can compare the Bayesian mean prediction with the prediction using the regularized (maximum penalized likelihood) estimate for β, which is

$$\hat{\beta} \;=\; (\lambda I^* + \Phi^T\Phi)^{-1}\Phi^T y$$

where I^* is like the identity matrix except that $I^*_{1,1} = 0$.

Compare with the posterior mean, if we set the prior mean, m_0, to zero:

$$\begin{aligned} m_n &= S_n(1/\sigma^2)\Phi^T y \\ &= (S_0^{-1} + (1/\sigma^2)\Phi^T\Phi)^{-1}(1/\sigma^2)\Phi^T y \\ &= (\sigma^2 S_0^{-1} + \Phi^T\Phi)^{-1}\Phi^T y \end{aligned}$$

If $S_0^{-1} = (1/\omega^2)I^*$, then these are the same, with $\lambda = \sigma^2/\omega^2$. This corresponds to a prior for β in which the β_j are independent, all with variance ω^2, except that β_0 has an infinite variance.

A Semi-Bayesian Way to Estimate σ^2 and ω^2

We see that σ^2 (the noise variance) and ω^2 (the variance of regression coefficients, other than β_0) together (as σ^2/ω^2) play a role similar to the penalty magnitude, λ, in the maximum penalized likelihood approach.

We can find values for σ^2 and ω^2 in a semi-Bayesian way by maximizing the *marginal likelihood* — the probability of the data (y) given values for σ^2 and ω^2. [We need to set the prior variance of β_0 to some finite ω_0^2 (which could be very large), else the probability of the observed data will be zero.]

We can also select basis function parameters (eg, s) by maximizing the marginal likelihood.

Such maximization is somewhat easier than the full Bayesian approach, in which we define some prior distribution for σ^2 and ω^2 (and any basis function parameters we haven't fixed), and then average predictions over their posterior distribution. [One would probably use some MCMC method to do this averaging.]

Finding the Marginal Likelihood for σ^2 and ω^2

The marginal likelihood for σ^2 and ω^2 given a vector of observed responses, y, is found by integrating over β with respect to its prior:

$$P(y \,|\, \sigma^2,\, \omega^2) \;\; = \;\; \int P(y \,|\, \beta, \sigma^2)\, P(\beta \,|\, \omega^2)\, d\beta$$

This is the denominator in Bayes' Rule, that normalizes the posterior.

Here, the basis function values for the training cases, based on the inputs for those cases, are considered fixed.

Both factors in this integrand are exponentials of quadratic functions of β, so this turns into the same sort of integral as that for the normalizing constant of a Gaussian density function, for which we know the answer.

Details of Computing the Marginal Likelihood

We go back to the computation of the posterior for β, but we now need to pay attention to some factors we ignored before. I'll fix the prior mean of β to $m_0 = 0$.

The log of the probability density of the data is

$$-\frac{N}{2}\log(2\pi) \ - \ \frac{N}{2}\log(\sigma^2) \ - \ \frac{1}{2}(y-\Phi\beta)^T(y-\Phi\beta)/\sigma^2$$

The log prior density for β is

$$-\frac{M}{2}\log(2\pi) \ - \ \frac{1}{2}\log(|S_0|) \ - \ \frac{1}{2}\beta^T S_0^{-1}\beta$$

expanding and then adding these together, we see the following terms that don't involve β:

$$-\frac{N+M}{2}\log(2\pi) \ - \ \frac{N}{2}\log(\sigma^2) \ - \ \frac{1}{2}\log(|S_0|) \ - \ \frac{1}{2}y^T y/\sigma^2$$

and these terms that do involve β:

$$-\frac{1}{2}\beta^T\Phi^T\Phi\beta/\sigma^2 \ + \ \beta^T\Phi^T y/\sigma^2 \ - \ \frac{1}{2}\beta^T S_0^{-1}\beta$$

More Details. . .

We can combine the quadratic terms that involve β, giving

$$-\frac{1}{2}\Big[\beta^T(S_0^{-1}+\Phi^T\Phi/\sigma^2)\beta \;-\; 2\beta^T\Phi^T y/\sigma^2\Big]$$

We had previously used this to identify the posterior covariance and mean for β. Setting the prior mean to zero, these are

$$S_n \;=\; \Big[S_0^{-1} \;+\; (1/\sigma^2)\Phi^T\Phi\Big]^{-1}, \qquad m_n \;=\; S_n\Phi^T y/\sigma^2$$

We can write the terms involving β using these, then "complete the square":

$$\begin{aligned}
&-\frac{1}{2}\Big[\beta^T S_n^{-1}\beta \;-\; 2\beta^T S_n^{-1}m_n\Big]\\
&\quad=\; -\frac{1}{2}\Big[\beta^T S_n^{-1}\beta \;-\; 2\beta^T S_n^{-1}m_n \;+\; m_n^T S_n^{-1}m_n\Big] \;+\; \frac{1}{2}m_n^T S_n^{-1}m_n\\
&\quad=\; -\frac{1}{2}(\beta-m_n)^T S_n^{-1}(\beta-m_n) \;+\; \frac{1}{2}m_n^T S_n^{-1}m_n
\end{aligned}$$

The second term above doesn't involve β, so we can put it with the other such.

And Yet More Details…

We now see that the log of the prior times the probability of the data has these terms not involving β:

$$-\frac{N+M}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log(|S_0|) \;-\; \frac{1}{2}y^T y/\sigma^2 \;+\; \frac{1}{2}m_n^T S_n^{-1} m_n$$

and this term that does involve β:

$$-\frac{1}{2}(\beta - m_n)^T S_n^{-1}(\beta - m_n)$$

When we exponentiate this and then integrate over β, we see that

$$\int \exp\Big(-\frac{1}{2}(\beta - m_n)^T S_n^{-1}(\beta - m_n)\Big)\, d\beta \;\;=\;\; (2\pi)^{M/2}\,|S_n|^{1/2}$$

since this is just the integral defining the Gaussian normalizing constant.

The final result is that the log of the marginal likelihood is

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\Big(\frac{|S_0|}{|S_n|}\Big) \;-\; \frac{1}{2}y^T y/\sigma^2 \;+\; \frac{1}{2}m_n^T S_n^{-1} m_n$$

Another Formula for the Marginal Likelihood

The last two terms in the formula on the previous slide seem a bit mysterious.

They can be rewritten as follows:

$$\begin{aligned}
&-\frac{1}{2}y^Ty/\sigma^2 \;+\; \frac{1}{2}m_n^TS_n^{-1}m_n \\
&\qquad = \;\; -\frac{1}{2}y^Ty/\sigma^2 \;+\; m_n^TS_n^{-1}m_n \;-\; \frac{1}{2}m_n^TS_n^{-1}m_n \\
&\qquad = \;\; -\frac{1}{2}y^Ty/\sigma^2 \;+\; m_n^T\Phi^Ty/\sigma^2 \;-\; \frac{1}{2}m_n^T\Phi^T\Phi m_n/\sigma^2 \;-\; \frac{1}{2}m_n^TS_0^{-1}m_n \\
&\qquad = \;\; -\frac{1}{2}||y-\Phi m_n)||^2/\sigma^2 \;-\; \frac{1}{2}m_n^TS_0^{-1}m_n
\end{aligned}$$

This gives another formula for the log marginal likelihood, which is more intuitive and also better numerically (avoids large roundoff in computing y^Ty):

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\left(\frac{|S_0|}{|S_n|}\right) \;-\; \frac{1}{2}||y-\Phi m_n)||^2/\sigma^2 \;-\; \frac{1}{2}m_n^TS_0^{-1}m_n$$

Here, $(1/2)\log(|S_0|/|S_n|)$ is the log of the factor by which the prior contracts to the posterior, the next term is the data fit with the posterior mean, and the last term is the prior density at the posterior mean.

Computations for the Semi-Bayesian Approach

Maximizing the marginal likelihood with respect to σ^2, ω^2, and parameters of the basis functions could be done by many standard optimization methods.

For maximizing with respect to σ^2 and ω^2, there's also an iterative re-estimation procedure (see the next slide).

We can then use the posterior mean, m_n, to predict the response in a test case with inputs x, as $\phi(x)^T m_n$. The posterior covariance, S_n, is used in producing a predictive variance for the response, which is $\phi(x)^T S_N \phi(x) + \sigma^2$.

Note that these semi-Bayesian predictions are all based on a *single* set of values for σ^2, ω^2, etc., although they do integrate over β.

Re-estimating σ^2 and ω^2

Naively, one might iterate finding the posterior mean and covariance of β, based on the current estimates for σ^2 and ω^2, with the following re-estimation of σ^2 and ω^2:

$$\widehat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \Phi(x_i)^T m_n)^2, \qquad \widehat{\omega}^2 \;=\; \frac{1}{M}\sum_{j=0}^{M-1}[m_n]_j$$

This assumes $S_0 = \omega^2 I$. But this isn't quite right: consider that some data points could be fitted nearly exactly when the model is flexible, and some coefficients in m_n could be nearly zero if they aren't relevant to any data point.

Instead, we find the "effective number of parameters", γ, as

$$\gamma \;=\; \sum_{j=1}^{M}\frac{\lambda_i}{\lambda_i + 1/\omega^2}$$

where the λ_i are the eigenvalues of $\Phi^T\Phi/\sigma^2$, and re-estimate as follows:

$$\widehat{\sigma}^2 \;=\; \frac{1}{n-\gamma}\sum_{i=1}^{n}(y_i - \Phi(x_i)^T m_n)^2, \qquad \widehat{\omega}^2 \;=\; \frac{1}{\gamma}\sum_{j=0}^{M-1}([m_n]_j)^2$$

See David MacKay's thesis (Section 2.4) for more explanation.

Computations for the Fully-Bayesian Approach

The full Bayesian approach is to integrate over the posterior distribution for σ, ω, etc. as well as β. This can be done by MCMC (eg, Metropolis or slice sampling), using the marginal likelihod times the prior density for σ, etc.

We then make a prediction for the response in a test case by averaging the posterior mean for β based on a sample of values for σ, etc. The standard deviation for the unknown response can be found as well. We could also approximate the whole predictive distribution, which in general is not Gaussian.

Alternatively, we can sample for β as well as σ, etc. (perhaps with Gibbs sampling). This avoids any expensive matrix computations, but fails to take advantage of conjugacy. We'd need to do this if we used a non-conjugate prior for β.

Note: We can't use an improper prior for ω that gives infinite mass to $\omega \rightarrow 0$, since $\omega = 0$ gives only finite misfit to the data. Similarly, if ϕ allows the data to be fit exactly, we may not be able to use an improper prior for σ with infinite mass at zero.

How Feasible are Linear Basis Function Models?

Modeling a general non-linear relationship of y to x with a linear basis function model seems attractive when x is of low dimension, but when there are many inputs, we would seem to need a huge number of local basis functions to "cover" the high dimensional input space. This is at least a computational problem.

One possibility is to use a relatively small number of basis functions, that cover only the actual area where x values are found, which may be the vicinity of a manifold of much lower dimension. We might:

– pick a subset of data points as centres for basis functions

– make the basis functions depend on parameters that adapt to the data.

A neural network with one hidden layer is an example of the latter approach.

Instead, we might go ahead and use a huge number of basis functions, maybe an infinite number. We need a computational trick for this. . .

Prior Distribution of Responses for a Linear Basis Function Model with Gaussian Noise and Gaussian Prior

When M, the number of basis functions, is greater than n, the number of observations in our training set, it is computationally attractive to shift focus from the parameters β_j for $j = 0, \ldots, M-1$ (collectively written β) to the observed responses, y_i for $i = 1, \ldots, n$ (collectively written y).

We need to find the prior distribution of y implied by the prior distribution of β.

If the prior distribution of β is Gaussian, the prior of y will also be Gaussian, since $y = \Phi\beta + e$ is a linear function of jointly Gaussian variables.

If the prior for β has mean zero, so will the prior for y.

If the prior covariance of β is S_0, the prior covariance of y will be $\sigma^2 I + \Phi S_0 \Phi^T$. If the β_j are independent in the prior, with the variance of β_j being ω_j^2, then

$$\text{Cov}(y_i, y_{i'}) \;=\; \sigma^2 \delta_{i,i'} \;+\; \sum_{j=0}^{M-1} \omega_j^2 \phi_j(x_i) \phi_j(x_{i'})$$

where $\delta_{i,i'} = 1$ if $i = i'$ and zero otherwise.

Predicting Directly Using the Prior for Responses

In similar fashion, we can find the prior covariance between responses in any two cases, whether they be training cases or future test cases.

Let C be the $n \times n$ covariance matrix of all the responses, $y_1, \ldots, y_n$, in the training set. For some test case with input x_*, let k be the vector of covariances of the response for the test case, y_*, with the responses for training cases. Finally, let v be the variance of the test response (covariance of y_* with itself).

As before, we assume prior means of zero (from a prior mean of zero for β).

We can now make predictions directly, without further reference to the β parameters, by finding

$$P(y_* \,|\, y_1, \ldots, y_n)$$

Since conditional distributions from multivariate Gaussians are Gaussian, this predictive distribution will be Gaussian, fully specified by mean and variance.

Applying the general formulas for Gaussian conditional distributions, we get

$$E(y_* \,|\, y_1, \ldots, y_n) \;=\; k^T C^{-1} y, \qquad \text{Var}(y_* \,|\, y_1, \ldots, y_n) \;=\; v \,-\, k^T C^{-1} k$$

Takes $O(n^3 + n^2 M)$ time to compute. Compare $O(M^3 + nM^2)$ for previous method.

Marginal Likelihood Directly from the Prior for the Responses

When σ, ω, and perhaps some parameters of the ϕ functions are not known, we may wish to estimate or sample them based on the marginal likelihood given the observed responses, y.

We saw how to do this before, working with the posterior distribution of β, in $O(M^3+nM^2)$ time.

Working directly with the covariances of the responses, the marginal likelihood is just the Gaussian prior probability density for the responses. So the log marginal likelihood is

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{1}{2}\log(|C|) \;-\; \frac{1}{2}y^T C^{-1} y$$

This takes $O(n^3+n^2M)$ time to compute.

For both prediction and marginal likelihood, which method is faster depends on the relative magnitudes of n and M. When M is sufficiently bigger than n, it's better to work directly with the responses, integrating away β.

Letting the Number of Basis Functions go to Infinity

When working directly with the responses, the basis functions and the prior for the β_j are used only to find the covariance between the responses in two cases, which we can write as

$$\text{Cov}(y_i, y_{i'}) \;\;=\;\; \sigma^2 \delta_{i,i'} \;+\; K(x_i, x_{i'})$$

where K is the noise-free covariance function:

$$K(x, x') \;\;=\;\; \sum_{j=0}^{M-1} \omega_j^2 \phi_j(x) \phi_j(x')$$

If our choice of ω_j and ϕ_j for $j = 1, 2, 3, \ldots$ is such that the sum above reaches a finite limit as $M \to \infty$, the model with infinite M makes sense.

If there's a formula to compute this infinite sum, we can implement this model with infinite M. If time to compute $K(x, x')$ is linear in the number of inputs, p, the marginal likelihood or a prediction will take $O(n^3 + n^2 p)$ time to compute.

[If we are predicting for many test cases, each additional test case takes $O(n)$ time for just the predictive mean, and $O(n^2)$ time if we also want the variance.]

An Infinite Basis Function Model with Sines and Cosines

With one input, let's use as basis functions $\phi_0(x) = 1$, and for $h = 1, 2, 3, \ldots$

$$\phi_{2h-1}(x) \;=\; \sin(f_h x), \qquad \phi_{2h}(x) \;=\; \cos(f_h x)$$

where each f_h is independently drawn from the $N(0, \rho^2)$ distribution.

For $j = 1, \ldots, M-1$, we'll let

$$\omega_j^2 \;=\; \frac{\eta^2}{(M-1)/2}$$

We now look at the limit as $M \to \infty$ of

$$\begin{aligned}
K(x, x') \;&=\; \omega_0^2 \;+\; \sum_{j=1}^{M-1} \omega_j^2 \phi_j(x) \phi_j(x') \\
&=\; \omega_0^2 \;+\; \sum_{h=1}^{(M-1)/2} \frac{\eta^2}{(M-1)/2} \Big[\sin(f_h x) \sin(f_h x') + \cos(f_h x) \cos(f_h x') \Big] \\
&=\; \omega_0^2 \;+\; \eta^2 \, \frac{1}{(M-1)/2} \sum_{h=1}^{(M-1)/2} \Big[\sin(f_h x) \sin(f_h x') + \cos(f_h x) \cos(f_h x') \Big]
\end{aligned}$$

The average of $(M-1)/2$ terms above approaches an integral as $M \to \infty$.

Covariance Function for the Model with Sines and Cosines

We can now find the covariance function as $M \to \infty$:

$$
\begin{aligned}
K(x,x') \;&=\; \omega_0^2 \;+\; \eta^2\, \frac{1}{(M-1)/2} \sum_{h=1}^{(M-1)/2} \Big[\sin(f_h x)\sin(f_h x') + \cos(f_h x)\cos(f_h x') \Big] \\
&\to\; \omega_0^2 \;+\; \eta^2 \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\rho} \exp\Big(-\,\frac{f^2}{2\rho^2}\Big) \Big[\sin(fx)\sin(fx') + \cos(fx)\cos(fx') \Big]\, df \\
&=\; \omega_0^2 \;+\; \eta^2\, \frac{1}{\sqrt{2\pi}\rho} \int_{-\infty}^{+\infty} \exp\Big(-\,\frac{f^2}{2\rho^2}\Big)\; \cos(f(x-x'))\, df \\
&=\; \omega_0^2 \;+\; \eta^2\, \frac{1}{\sqrt{2\pi}\rho} \Big[\sqrt{2\pi}\rho\, \exp(-\rho^2(x-x')^2/2) \Big] \\
&=\; \omega_0^2 \;+\; \eta^2\, \exp(-\rho^2(x-x')^2/2)
\end{aligned}
$$

This is simple to compute, so it's easy to use the model with infinite M.