STA 414/2104

Statistical Methods for Machine Learning and Data Mining

Radford M. Neal, University of Toronto, 2012

Week 5

More on Bayesian Linear Basis Function Models

Comparison with Regularized Estimates

In a Bayesian linear basis function model, the 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 Markov chain Monte Carlo (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:

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

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":

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

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:

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

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

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

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 then re-estimate as follows:

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

Details are in David MacKay's thesis (Section 2.4).

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, which can be done by MCMC methods, using the marginal likelihood for σ, ω, etc. (integrating over β).

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, ω, 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, ω, etc. 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 $\omega \to 0$, 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'll later see that there's a computational trick that allows this.