Bayesian Gaussian / Linear Models

Read Sections 2.3.3 and 3.3 in the text by Bishop

Multivariate Gaussian Model with Multivariate Gaussian Prior

Suppose we model the 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):

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

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 $(A^{-1}+B^{-1})^{-1}$.

Bayesian Linear Basis Function Model

Recall the linear basis function model, which we can write as

$$t \;\sim\; N(\Phi w,\, \sigma^2 I)$$

where here,

– t is the vector of observed targets

– w is the vector of regression coefficients

– σ^2 is the "noise" variance

– Φ is the matrix of basis function values in the training cases.

Suppose that our prior for w is $N(m_0, S_0)$. This is a conjugate prior, with the posterior for w also being normal.

Posterior for Linear Basis Function Model

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

$$-\tfrac{1}{2}\Big[(t-\Phi w)^T(\sigma^2 I)^{-1}(t-\Phi w)\Big] \;+\; \cdots \quad = \quad -\tfrac{1}{2}\tfrac{1}{\sigma^2}\Big[w^T\Phi^T\Phi w \;-\; 2w^T\Phi^T t\Big] \;+\; \cdots$$

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

This combines with the prior for w in the same way as seen earlier, with the result that the posterior distribution for w 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

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

Predictive Distribution for a Test Case

We can write the target, t, for some new case with inputs x as

$$t \;=\; \phi(x)^T w + n$$

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

Since the posterior distribution for w is $N(m_N,\, S_N)$, the posterior distribution for $\phi(x)^T w$ will be $N(\phi(x)^T m_N,\, \phi(x)^T S_N \phi(x))$.

Hence the predictive distribution for t 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 (since the model is linear in the parameters).

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

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

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:

$$\begin{aligned} m_N &= S_N(1/\sigma^2)\Phi^T t \\ &= (S_0^{-1} \,+\, (1/\sigma^2)\Phi^T\Phi)^{-1}(1/\sigma^2)\Phi^T t \\ &= (\sigma^2 S_0^{-1} \,+\, \Phi^T\Phi)^{-1}\Phi^T t \end{aligned}$$

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 w in which the w_j are independent, all with variance ω^2, except that w_0 (the intercept) 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 w_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 (t) given values for σ^2 and ω^2. [We need to set the prior variance of w_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.

Note: Notation in Bishop's book is $\beta = 1/\sigma^2$ and $\alpha = 1/\omega^2$. The marginal likelihood is sometimes called the "evidence".

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

The marginal likelihood for σ^2 and ω^2 given a vector of observed targets values, t, is found by integrating over w with respect to its prior:

$$P(t \,|\, \sigma^2,\, \omega^2) \;\;=\;\; \int P(t \,|\, w, \sigma^2)\, P(w \,|\, \omega^2)\, dw$$

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 w, 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 w, but we now need to pay attention to some factors we ignored before. I'll fix the prior mean for w to $m_0 = 0$.

The log of the probability density of the data is

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}(t-\Phi w)^T(t-\Phi w)/\sigma^2$$

The log prior density for w is

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

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

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

and these terms that do involve w:

$$-\frac{1}{2}w^T\Phi^T\Phi w/\sigma^2 \;+\; w^T\Phi^T t/\sigma^2 \;-\; \frac{1}{2}w^T S_0^{-1} w$$

More Details...

We can combine the quadratic terms that involve w, giving

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

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

$$S_N \ = \ \Big[S_0^{-1} \ + \ (1/\sigma^2)\Phi^T\Phi\Big]^{-1}, \qquad m_N \ = \ S_N\Phi^T t/\sigma^2$$

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

$$\begin{aligned} & -\frac{1}{2}\Big[w^T S_N^{-1} w \ - \ 2w^T S_N^{-1} m_N\Big] \\ & \quad = \ -\frac{1}{2}\Big[w^T S_N^{-1} w \ - \ 2w^T S_N^{-1} m_N \ + \ m_N^T S_N^{-1} m_N\Big] \ + \ \frac{1}{2} m_N^T S_N^{-1} m_N \\ & \quad = \ -\frac{1}{2}(w - m_N)^T S_N^{-1}(w - m_N) \ + \ \frac{1}{2} m_N^T S_N^{-1} m_N \end{aligned}$$

The second term above doesn't involve w, 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 w:

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

and this term that does involve w:

$$-\frac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)$$

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

$$\int \exp\Big(-\frac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)\Big)\,dw \;=\; (2\pi)^{M/2}\,|S_N|^{1/2}$$

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

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

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\left(\frac{|S_0|}{|S_N|}\right) \;-\; \frac{1}{2}t^T t/\sigma^2 \;+\; \frac{1}{2}m_N^T S_N^{-1} m_N$$

Compare equation (3.86) in Bishop's book…Have I made a mistake?

Correspondence with Bishop's Marginal Likelihood Formula

I didn't make a mistake. My formula from the previous slide:

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\left(\frac{|S_0|}{|S_N|}\right) \;-\; \frac{1}{2}t^Tt/\sigma^2 \;+\; \frac{1}{2}m_N^TS_N^{-1}m_N$$

Bishop's formula, after translating notation and minor rearrangement:

$$-\frac{N}{2}\log(2\pi) \;-\; \frac{N}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\left(\frac{|S_0|}{|S_N|}\right) \;-\; \frac{1}{2}||t-\Phi m_N)||^2/\sigma^2 \;-\; \frac{1}{2}m_N^TS_0^{-1}m_N$$

The difference is in the last two terms. Expanding these in Bishop's formula gives

$$\begin{aligned}
&-\frac{1}{2}||t-\Phi m_N)||^2/\sigma^2 \;-\; \frac{1}{2}m_N^TS_0^{-1}m_N \\
&\quad= \;-\frac{1}{2}t^Tt/\sigma^2 \;+\; m_N^T\Phi^Tt/\sigma^2 \;-\; \frac{1}{2}m_N^T\Phi^T\Phi m_N/\sigma^2 \;-\; \frac{1}{2}m_N^TS_0^{-1}m_N \\
&\quad= \;-\frac{1}{2}t^Tt/\sigma^2 \;+\; m_N^TS_N^{-1}m_N \;-\; \frac{1}{2}m_N^TS_N^{-1}m_N \\
&\quad= \;-\frac{1}{2}t^Tt/\sigma^2 \;+\; \frac{1}{2}m_N^TS_N^{-1}m_N
\end{aligned}$$

Bishop's formula is probably better numerically — the roundoff error when computing t^Tt could sometimes be large.