CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Gaussian Processes II

Lecturer: Michael I. Jordan *Scribes: Bhaskara Marthi*

First, we conclude our discussion of the Gaussian Process view of Bayesian regression. At the end of the last lecture, we had shown that, in the Gaussian Process view of linear prediction, a new value Z_* is Gaussian with mean and covariance given by

$$\begin{aligned} E(Z_*|Y_1 = y_1, \ldots, Y_n = y_n) &= \phi_*^T \Sigma \Phi^T P^{-1} y \\ Var(Z_*|Y_1 = y_1, \ldots, Y_n = y_n) &= \phi_*^T \Sigma \phi_* - \phi_*^T \Sigma \Phi^T P^{-1} \Phi \Sigma \phi_* \end{aligned}$$

where $P = \Phi \Sigma \Phi^T + \sigma^2 I$.

We now confirm that is the same result as would be predicted by Bayesian linear regression.

To show that the means are the same involves showing that $\beta A^{-1} \Phi^T = \Sigma \Phi^T P^{-1}$. This is shown by

$$\begin{aligned} A \Sigma \Phi^T &= (\beta \Phi^T \Phi + \Sigma^{-1}) \Sigma \Phi^T \\ &= \beta \Phi^T \Phi \Sigma \Phi^T + \Phi^T \\ &= \beta \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I) \\ &= \beta \Phi^T P \end{aligned}$$

As for the variances,

$$\begin{aligned} A^{-1} &= (\beta \Phi^T \Phi + \Sigma^{-1})^{-1} \\ &= \Sigma + \Sigma \Phi^T (\sigma^2 I + \Phi \Sigma \Phi^T)^{-1} \Phi \Sigma \end{aligned}$$

by the matrix inversion lemma. Now multiply on the left by ϕ_*^T, and on the right by ϕ_* to obtain the desired result.

It should be noted that we can use this to get a MAP estimate

$$\begin{aligned} Z(x_*) &= \phi^T(x_*) \hat{\theta}_{MAP} \\ &= Ly \end{aligned}$$

i.e, it can be viewed as a linear function of the data.

Suppose there are n data points and m parameters (or basis functions). Bayesian regression involves inverting an m by m matrix, while the Gaussian Process view requires inverting an n by n matrix. So if the number of basis functions is less than the number of data points, then Bayesian regression is more efficient. However, there are situations where the number of basis functions is large or even infinite; in these cases Gaussian processes are the preferred way to treat the problem.

An example of this is infinite neural networks. Neural networks are models of the form

$$Z(x) = \theta_0 + \sum_{i=1}^{m} \theta_i \phi_i(\eta_i^T x)$$

The ϕ_i are assumed to be bounded. A common choice is for the ϕ_i to be logistic functions. It was shown by Neal that, under suitable independence assumptions, $Z(x)$ converges to a Gaussian process as $m \to \infty$.

Another example is to consider radial basis functions

$$\phi_i(x) = \exp\left(\frac{-(x-i)^2}{2\lambda^2}\right)$$

and let the prior covariance be $\Sigma = \text{diag}\frac{S}{\Delta}$, where S is a constant, and Δ is the number of basis functions contained in the interval being considered. The idea is to hold the interval fixed, but let the number of functions tend to infinity, in which case the sum in the covariance formula turns into an integral :

$$\begin{aligned}
\text{cov}(Z(x_n), Z(x'_n)) &= \sigma_i \phi_i(x_n)\frac{S}{\Delta}\phi_i(x'_n) \\
&\approx S\int dh\, \phi_h(x_n)\phi_h(x'_n) \\
&= S\int dh\, e^{\frac{-(x_n-h)^2}{2\lambda^2}} e^{\frac{-(x'_n-h)^2}{2\lambda^2}} \\
&= \sqrt{\pi\lambda^2} S e^{-\frac{(x_n-x'_n)^2}{4\lambda^2}}
\end{aligned}$$

which is a Gaussian kernel. So now, we can make predictions just using this kernel. Even though the number of basis functions is infinite, we only deal with the N by N matrix

$$Z(x_*) = k^T K^{-1} y$$