Outline

BAYESIAN REGRESSION: POSTERIOR DISTRIBUTION

- Let's assume the regression weights have a Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$ and that the bias is zero
- The log posterior is

$$\log p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = -\frac{\beta}{2}\sum_{j=1}^{N}[t_j - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_j})]^2 - \alpha\mathbf{w}^T\mathbf{w} + const$$

- Hence the posterior is Gaussian

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m_N}, \mathbf{S_N})$$

with mean and variance

$$\mathbf{m_N} = \beta\mathbf{S_N}\boldsymbol{\Phi}^T\mathbf{t}$$

$$\mathbf{S_N}^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

BAYESIAN REGRESSION: PREDICTIVE DISTRIBUTION

- Given the posterior, one can find the MAP estimate. However, in a fully Bayesian treatment, one makes predictions by integrating out the parameters via their posterior distribution.
- The predictive distribution

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{t}, \mathbf{w}, \alpha, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w}$$

is Gaussian

$$p(t|\mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m_N}^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

with mean $\mathbf{m_N}^T \boldsymbol{\phi}(\mathbf{x})$ and variance

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S_N} \boldsymbol{\phi}(\mathbf{x})$$

MARGINAL LIKELIHOOD

- We can find α and β by maximising the marginal likelihood: $p(\mathbf{t}|\alpha,\beta)$
- The log-marginal likelihood is:

$$\log p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(\mathbf{m_N}) - \frac{1}{2}\log|\mathbf{S_N}^{-1}| - \frac{N}{2}\log 2\pi$$

with

$$E(\mathbf{m_N}) = \boxed{\frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m_N}\|^2 + \frac{\alpha}{2}\mathbf{m_N}^T\mathbf{m_N}}$$

OPTIMISING THE MARGINAL LIKELIHOOD

- We will present a fix-point algorithm: we will write the gradient equations equal to zero as fix-point equations and iterate until convergence.
- In taking the derivative w.r.t α or β, the most challenging term is the log of the determinant of $\mathbf{S_N}^{-1} = \alpha\mathbf{I} + \boxed{\beta\mathbf{\Phi}^T\mathbf{\Phi}}$.
- To deal with it, let λ_i be the eigenvalues of $\boxed{\beta\mathbf{\Phi}^T\mathbf{\Phi}}$, so that $|\mathbf{S_N}^{-1}| = \prod_{i=0}^{M-1}(\alpha + \lambda_i)$.
- We then have that

$$\partial \log|\mathbf{S_N}^{-1}|/\partial\alpha = \sum_i \frac{1}{\alpha + \lambda_i}$$

$\Phi^T\Phi$ eigenvalues $\mu_1, \ldots, \mu_M$

$\lambda_i = \beta \cdot \mu_i$

- Moreover, λ_i are proportional to β, so that $\partial\lambda_i/\partial\beta = \lambda_i/\beta$

OPTIMISING THE MARGINAL LIKELIHOOD

- Now, define

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$

(which measures the number of well determined parameters)

- By deriving the log-marginal w.r.t. α and setting derivative to zero, we obtain:

$$\alpha = \frac{\gamma}{\mathbf{m_N}^T \mathbf{m_N}} = g_\alpha(\alpha, \beta)$$

- By deriving the log-marginal w.r.t. β and setting derivative to zero, we obtain:

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^{N} [t_n - \mathbf{m_N}^T \boldsymbol{\phi}(\mathbf{x_n})]^2 = \frac{1}{g_\beta(\alpha, \beta)}$$

- We start from an initial value α_0 and β_0 and iterate $\alpha_{n+1} = g_\alpha(\alpha_n, \beta_n)$, $\beta_{n+1} = g_\beta(\alpha_n, \beta_n)$ until convergence.

TASK 3

- Implement Bayesian regression, with type II likelihood optimisation of α and β.
- For the 1d non-linear dataset, use polynomial model of degree 12.
- Plot predictions and 95% confidence intervals, from the predictive distribution.
- For the 2d non-linear dataset, use the Gaussian functions models. How can we set the lengthscale γ?