PRML Note
C03 Linear Models for Regression

Yang Zhao

Department of Automation, Tsinghua University

- Given a training data set comprising N observations $\{\boldsymbol{x}_n\}$, together with corresponding target values $\{t_n\}$, the goal is to predict the value of t for a new value of $\boldsymbol{x}$.
- We can not only use the linear functions of the input variables, but also use the linear combinations of a fixed set of nonlinear functions of the input variables, known as *basis functions*. Such models are linear functions of the parameters and can be nonlinear with respect to the input variables.

1 Linear Basis Function Models

- The simplest linear model for regression is

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

We can extend it by considering linear combinations of fixed nonlinear functions of the input variables, of the form

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}) \tag{1}$$

where $\phi_j(\boldsymbol{x})$ are known as basis functions. We can also define an additional dummy basis function $\phi_0(\boldsymbol{x}) = 1$ so that

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{w} = (w_0, \cdots, w_{M-1})^T$ and $\boldsymbol{\phi} = (\phi_0, \cdots, \phi_{M-1})^T$.

- *polynomial basis function*

$$\phi_j(x) = x^j$$

One limitations of this basis functions is that they are global functions of the input variables, so that changes in one region of input space affect all other regions. This can be resolved by dividing the input space up into regions and fitting a different polynomial in each region, leading to *spline functions*.

- *Gaussian basis function*

$$\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2s^2}} \tag{3}$$

where the μ_j govern the locations of the basis functions in input space and the s governs their spatial scale.

- *sigmoidal basis function*

$$\phi_j(x) = \sigma\Big(\frac{x-\mu_j}{s}\Big) \tag{4}$$

where σ is the logistic sigmoid function defined by

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{5}$$

- We assume that the target variable t is given by a deterministic function $y(\boldsymbol{x}, \boldsymbol{w})$ with additive Gaussian noise so that

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$

where ϵ is a zero mean Gaussian random variable with precision β. Thus

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \tag{6}$$

- Consider a data set of inputs $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$ with corresponding target values $\boldsymbol{t} = (t_1, t_2, \cdots, t_N)^T$, which are drawn independently from the distribution 6, so we have

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1})$$

The gradient of the logliklihood function takes the form

$$\nabla lnp(\boldsymbol{t}|\boldsymbol{w}, \beta) = \beta \sum_{n=1}^{N} (t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))\boldsymbol{\phi}(\boldsymbol{x}_n) \tag{7}$$

Setting this gradient to zero gives

$$0 = \sum_{n=1}^{N} t_n\boldsymbol{\phi}(\boldsymbol{x}_n) - \Big(\sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n)\boldsymbol{\phi}(\boldsymbol{x}_n)^T\Big)\boldsymbol{w}$$

Solving for $\boldsymbol{w}$, we obtain

$$\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} \tag{8}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix}$$

- Considering the w_0, rewritten the error function, we have

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\boldsymbol{x}_n))$$

  Setting the derivative with respect to w_0 equal to zero, and solving for w_0, we obtain

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j\bar{\phi_j}$$

  where we define

$$\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n$$
$$\bar{\phi_j} = \frac{1}{N}\sum_{n=1}^{N} \phi_j(\boldsymbol{x}_n)$$

  Thus the bias w_0 compensates for the difference between the averages(over the training set) of the target values and the weighted sum of the averages of the basis function values.

- We can also maximize the log likelihood function with respect to the noise precision parameters β, giving

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \boldsymbol{w}_{ML}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2$$

- When $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is close to singular, solving this problem directly will have numerical difficulties. We can use *singular value decomposition*, or SVD to address this difficulties.

- *LMS algorithm.* This is also known as *least-mean-squares.* By applying the technique of *stochastic gradient descent*, we can update the $\boldsymbol{w}$

$$\begin{aligned}\boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} - \eta\nabla E_n \\ &= \boldsymbol{w}^{(t)} - \eta(\boldsymbol{w}^{(t)T}\boldsymbol{\phi}_n - t_n)\boldsymbol{\phi}_n\end{aligned}$$

  we can see how it works in figure 1

- By adding a regularization term to error function in order to control over-fitting, we get

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

  This choice of regularizer is known in the ML literature as *weight decay.* We obtain

$$\boldsymbol{w} = (\lambda\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} \tag{9}$$

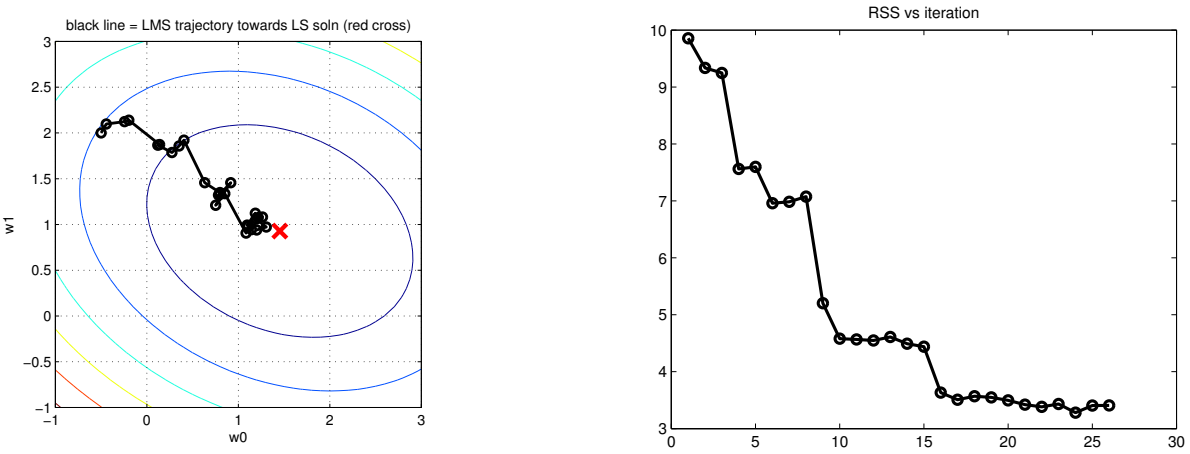

Figure 1: Illustration of LMS algorithm. Note that it doesn't decrease monotonically

- The problem of determining the optimal model complexity is then shifted from finding the appropriate number of basis functions to determining a suitable value of the regularization coefficient λ.
- For multiple outputs, we could solve this problem by introducing a different set of basis functions for each component of $\boldsymbol{t}$. But a more interesting and more common approach is to use the same set of basis functions to model all of the components of target vector, so that

$$y(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{10}$$

- Consider a linear basis function regression model for a multivariate target variable $\boldsymbol{t}$ having a Gaussian distribution of the form

$$p(\boldsymbol{t}|\boldsymbol{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{t}|y(\boldsymbol{x}, \boldsymbol{W}), \boldsymbol{\Sigma}) \tag{11}$$

where $y(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}^T \boldsymbol{\phi}(\boldsymbol{x})$, so the log-likelihood function is

$$lnp(\boldsymbol{T}|\boldsymbol{W}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} (\boldsymbol{t}_i - \boldsymbol{W}^T \phi(\boldsymbol{x}_i))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{t}_i - \boldsymbol{W}^T \phi(\boldsymbol{x}_i))$$

where

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \\ \vdots \\ \boldsymbol{t}_N \end{pmatrix} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1K} \\ t_{21} & t_{22} & \cdots & t_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{NK} \end{pmatrix}$$

So, let

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{W}} lnp(\boldsymbol{T}|\boldsymbol{W}, \boldsymbol{\Sigma}) &= \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) \phi(\boldsymbol{x}_i)^T \boldsymbol{W} \boldsymbol{\Sigma}^{-1} - \phi(\boldsymbol{x}_i) \boldsymbol{t}_i \boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{W} \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{0} \end{aligned} \tag{12}$$

we can get

$$\boldsymbol{W} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{T}$$

2 The Bias-Variance Decomposition

- Although the introduction of regularization terms can control over-fitting for models with many parameters, this raises the question of how to determine a suitable value for the regularization coefficient λ.
- Using an unbiased estimator is not always the best according to the bias-variance trade-off.
- For the squared loss function, the optimal prediction is given by

$$h(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{x}] = \int tp(t|\boldsymbol{x})dt$$

Accordingt to Chapter 1, we known that

$$\mathbb{E}[L] = \int (y(\boldsymbol{x}) - h(\boldsymbol{x}))^2 p(\boldsymbol{x})d\boldsymbol{x} + \iint (h(\boldsymbol{x}) - t)^2 p(t, \boldsymbol{x})dtd\boldsymbol{x} \tag{13}$$

The second term is independent of $y(\boldsymbol{x})$, representing the minimum achievable value of the expected loss. For any given data set $\mathcal{D}$, we can get a prediction function $y(\boldsymbol{x}; \mathcal{D})$, and the average over the different data sets denotes by $\mathbb{E}_{\mathcal{D}}(y(\boldsymbol{x}; \mathcal{D}))$. So the first term in the equation 13 can be rewritten by

$$\begin{aligned} \textit{the first term} &= \int (y(\boldsymbol{x}; \mathcal{D}) - h(\boldsymbol{x}))^2 p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int (y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}(y(\boldsymbol{x}; \mathcal{D})) + \mathbb{E}_{\mathcal{D}}(y(\boldsymbol{x}; \mathcal{D}) - h(\boldsymbol{x}))^2 p(\boldsymbol{x})d\boldsymbol{x} \\ &= \{\mathbb{E}_{\mathcal{D}}(y(\boldsymbol{x}; \mathcal{D})) - h(\boldsymbol{x})\}^2 + \mathbb{E}_{\mathcal{D}}(y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}\{y(\boldsymbol{x}; \mathcal{D}))^2\} \end{aligned} \tag{14}$$

- the first term in equation 14 is called *the squared bias*, which represents the extent to which the average prediction over all data sets differs from the desired regression.
- the second term, called the *variance*, measures the extent to which the solutions for individual data sets vary around their average.

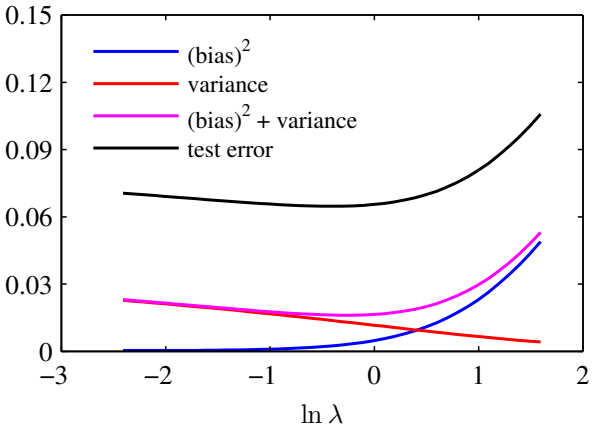

Figure 2: The bias-variance decomposition

- The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.
- The bias-variance decomposition is based on averages with respect to ensembles of data sets, whereas in practice we have only the single observed data set. And the larger data set can support the larger capability of the model. So spliting the data set into many sets is not good for most situations.

3 Bayesian Linear Regression

- Given the conjugate prior
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$
and the likelihood function
$$p(\boldsymbol{t}|\boldsymbol{w}) = p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1})$$
So we can write the posterior distribution directly in the form
$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{m}_N, \boldsymbol{S}_N)$$
where
$$\begin{aligned}\boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^T\boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\end{aligned}$$
- If data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point. By completing the square in the exponential, we can obtain
$$\begin{aligned}\boldsymbol{S}_{N+1}^{-1} &= \boldsymbol{S}_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T \\ \boldsymbol{m}_{N+1} &= \boldsymbol{S}_{N+1}(\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \beta\phi_{N+1}t_{N+1})\end{aligned}$$
- Sometimes, we are not interested in the value of $\boldsymbol{x}$ itself but rather in making predictions of t for new values of $\boldsymbol{x}$.
- The *predictive distribution*
$$p(t|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \int p(t|\boldsymbol{x}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)d\boldsymbol{w}$$
where
$$\begin{aligned}p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) &= \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \\ p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)\end{aligned}$$
So the predictive distribution takes the form
$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|\boldsymbol{m}_N^T(\phi(\boldsymbol{x})), \sigma_N^2(\boldsymbol{x}))$$
where
$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \phi(\boldsymbol{x})^T\boldsymbol{S}_N\phi(\boldsymbol{x})$$
The first term represents the noise on the data whereas the second term reflects the uncertainty associated with the parameters $\boldsymbol{w}$.

- Assume we have already observed N data points. By considering an additional $(\boldsymbol{x}_{N+1}, t_{N+1})$

$$\begin{aligned}\boldsymbol{S}_{N+1}^{-1} &= \boldsymbol{S}_N^{-1} + \beta\phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T \\ \boldsymbol{m}_{N+1} &= \boldsymbol{S}_{N+1}(\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \phi(\boldsymbol{x}_{N+1})t_{N+1})\end{aligned}$$

  So by completing the square in the exponential

$$\begin{aligned}\sigma_{N+1}^2(\boldsymbol{x}) - \sigma_N^2(\boldsymbol{x}) &= \frac{1}{\beta} + \phi(\boldsymbol{x})^T\boldsymbol{S}_N\phi(\boldsymbol{x}) - \Big(\frac{1}{\beta} + \phi(\boldsymbol{x})^T\boldsymbol{S}_{N+1}\phi(\boldsymbol{x})\Big) \\ &= \phi(\boldsymbol{x})^T(\boldsymbol{S}_{N+1} - \boldsymbol{S}_N)\phi(\boldsymbol{x}) \\ &= -\phi(\boldsymbol{x})^T\Big(\frac{\beta\boldsymbol{S}_N\phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N}{1+\beta\phi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N\phi(\boldsymbol{x}_{N+1})}\Big)\phi(\boldsymbol{x}) \\ &\leq 0\end{aligned}$$

  because $\boldsymbol{S}_{N+1} - \boldsymbol{S}_N$ is negative definite matrix. So the second term in the predictive variance will go to zero.

- Considering the predictive mean, we obtain

$$y(\boldsymbol{x}, \boldsymbol{m}_M) = \boldsymbol{m}_N^T\phi(\boldsymbol{x}) = \sum_{n=1}^{N}\beta\phi(\boldsymbol{x})^T\boldsymbol{S}_N\phi(\boldsymbol{x}_n)t_n \tag{15}$$

  Thus the mean of the predictive at a point $\boldsymbol{x}$ is given by a linear combination of the training set target variables t_n, so that we can write

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^{N}k(\boldsymbol{x}, \boldsymbol{x}_n)t_n \tag{16}$$

  where

$$k(\boldsymbol{x}, \boldsymbol{x}_n) = \beta\phi(\boldsymbol{x})^T\boldsymbol{S}_N\phi(\boldsymbol{x}_n)$$

  is known as the *smoother matrix* or the *equivalent kernel*, which can also be written as

$$k(\boldsymbol{x}, \boldsymbol{z}) = \psi(\boldsymbol{x})^T\psi(\boldsymbol{z})$$

  where $\psi(\boldsymbol{x}) = \beta^{1/2}\boldsymbol{S}_N^{1/2}\phi(\boldsymbol{x})$.

- We can see that $k(\boldsymbol{x}, \boldsymbol{x}_n)$ can be weight of the linear combination and the data points close to $\boldsymbol{x}$ are given higher weight than the data points further removed from $\boldsymbol{x}$.

- Note that the kernel function can be negative as well as positive, so although it satisfies a summation constraint, the corresponding predictions are not necessarily convex combinations of the training set target variables.

4 Bayesian Model Comparison

- Suppose we wish to compare a set of L models $\{\mathcal{M}_i\}$ where $i = 1, 2, \cdots, L$. Given a training set $\mathcal{D}$, we wish to caculate the posterior distribution $p(\mathcal{M}_i)$

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$$

The prior allows us to express a preference for different models. Let us simply assume that all models are given equal prior probability. The interesting term is the *model evidence* $p(\mathcal{D}|\mathcal{M}_i)$ which expresses the preference shown by the data for different models. The model evidence is sometimes also called the *marginal likelihood*. The ratio of model evidence $p(\mathcal{D}|\mathcal{M}_i)/p(\mathcal{D}|\mathcal{M}_j)$ for two models is known as a *Bayes factor*.

5 The Evidence Approximation

- In a fully Bayesian treatment of the linear basis function model, we would introduce prior distributions over the hyperparameters α and β and make predictions by marginalizing with respect to these hyperparameters as well as with respect to the parameters w, which is analytically intractable. We can make an approximation in which we set the hyperparameters to specific values determined by maximizing the *marginal likelihood function* obtained by first integrating over the parameters w. This framework is known as *empirical Bayes*, or *type 2 maximum likelihood*, or *generalized maximum likelihood*, or *evidence approximation*.

- The predictive distribution is obtained by marginalizing over $\boldsymbol{w}, \alpha, \beta$, so that

$$p(t|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{X}) = \iiint p(t|\boldsymbol{x}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)p(\alpha, \beta|\boldsymbol{t}, \boldsymbol{X})d\boldsymbol{w}d\alpha d\beta \tag{17}$$

where

$$\begin{aligned} p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) &= \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \\ p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N) \\ \boldsymbol{m}_N &= \beta \boldsymbol{S}_N \boldsymbol{\Phi}^T \boldsymbol{t} \\ \boldsymbol{S}_N^{-1} &= \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \end{aligned}$$

For posterior distribution $p(\alpha, \beta|\boldsymbol{t}, \boldsymbol{X})$, we can simply assume the it is sharply peaked around values $\hat{\alpha}$ and $\hat{\beta}$, so

$$p(t|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{X}) \simeq p(t|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{X}, \hat{\alpha}, \hat{\beta}) = \int p(t|\boldsymbol{x}, \boldsymbol{w}, \hat{\beta})p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}, \hat{\alpha}, \hat{\beta})d\boldsymbol{w}$$

and the posterior distribution for α and β is given by

$$p(\alpha, \beta|\boldsymbol{t}, \boldsymbol{X}) \propto p(\boldsymbol{t}|\alpha, \beta, \boldsymbol{X})p(\alpha, \beta)$$

If the prior is relatively flat, the value of $\hat{\alpha}$ and $\hat{\beta}$ are obtained by maximizing the marginal likelihood function $p(\boldsymbol{t}|\alpha, \beta, \boldsymbol{X})$. We can both use the analytially method by setting the derivative equal to zero and the expectation maximization algorithm.

- The marginal likelihood function $p(\boldsymbol{t}|\alpha, \beta, \boldsymbol{X})$ is obtained by integrating over the weight parameters $\boldsymbol{w}$, so that

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

where

$$\begin{aligned} lnp(\boldsymbol{t}|\boldsymbol{w}, \beta, \boldsymbol{X}) &= \frac{N}{2} ln\beta - \frac{N}{2} ln(2\pi) - \beta E_D(\boldsymbol{w}) \\ E_D(\boldsymbol{w}) &= \frac{1}{2} \|\boldsymbol{t} - \boldsymbol{\Phi w}\|^2 \\ p(\boldsymbol{w}|\alpha) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \end{aligned}$$

- We can write the evidence function in the form

$$p(\boldsymbol{t}|\alpha, \beta, \boldsymbol{X}) = \Big(\frac{\beta}{2\pi}\Big)^{N/2} \Big(\frac{\alpha}{2\pi}\Big)^{M/2} \int e^{-E(\boldsymbol{w})} d\boldsymbol{w}$$

where

$$E(\boldsymbol{w}) = \beta E_D(\boldsymbol{w}) + \alpha E_W(\boldsymbol{w}) = \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{\Phi w}\|^2 + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w}$$

we can complete the square over $\boldsymbol{w}$ giving

$$E(\boldsymbol{w}) = E(\boldsymbol{m}_N) + \frac{1}{2} (\boldsymbol{w} - \boldsymbol{m}_N)^T \boldsymbol{A} (\boldsymbol{w} - \boldsymbol{m}_N)$$

where we have introduced

$$\begin{aligned} \boldsymbol{A} &= \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \\ E(\boldsymbol{m}_N) &= \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{\Phi m}_N\|^2 + \frac{\alpha}{2} \boldsymbol{m}_N^T \boldsymbol{m}_N \\ \boldsymbol{m}_N &= \beta \boldsymbol{A}^{-1} \boldsymbol{\Phi}^T \boldsymbol{t} \end{aligned}$$

So the integral over $\boldsymbol{w}$ can now be evaluated as

$$\begin{aligned} \int e^{-E(\boldsymbol{w})} d\boldsymbol{w} &= e^{-E(\boldsymbol{m}_N)} \int e^{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^T \boldsymbol{A} (\boldsymbol{w} - \boldsymbol{m}_N)} d\boldsymbol{w} \\ &= e^{-E(\boldsymbol{m}_N)} (2\pi)^{M/2} |\boldsymbol{A}|^{-1/2} \end{aligned}$$

Thus we can obtain

$$lnp(\boldsymbol{t}|\alpha, \beta, \boldsymbol{X}) = \frac{M}{2} ln\alpha + \frac{N}{2} ln\beta - E(\boldsymbol{m}_N) - \frac{1}{2} ln|\boldsymbol{A}| - \frac{N}{2} ln(2\pi) \qquad (18)$$

- Another way to get the equation 18. The likelihood function can be rewritten by

$$p(\boldsymbol{t}|\boldsymbol{w}, \beta, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{\Phi w}, \beta^{-1} \boldsymbol{I}_N)$$

So, the result gives that

$$p(\boldsymbol{t}|\alpha, \beta, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \beta^{-1} \boldsymbol{I}_N + \alpha^{-1} \boldsymbol{\Phi} \boldsymbol{\Phi}^T)$$

Taking the log we obtain

$$lnp(\boldsymbol{t}|\alpha,\beta,\boldsymbol{X}) = -\frac{N}{2}ln(2\pi) - \frac{1}{2}ln|\beta^{-1}\boldsymbol{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T| - \frac{1}{2}\boldsymbol{t}^T(\beta^{-1}\boldsymbol{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{t}$$

Using the equation26, we have

$$\begin{aligned}
|\beta^{-1}\boldsymbol{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T| &= \beta^{-N}|\boldsymbol{I}_N + \beta\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T| \\
&= \beta^{-N}|\boldsymbol{I}_M + \alpha^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}| \\
&= \beta^{-N}\alpha^{-M}|\alpha\boldsymbol{I}_M + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}| \\
&= \beta^{-N}\alpha^{-M}|\boldsymbol{A}|
\end{aligned}$$

and using the *Woodbury matrix identity* introduced in Chapter2 appendix, we obtain

$$\begin{aligned}
-\frac{1}{2}\boldsymbol{t}^T(\beta^{-1}\boldsymbol{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{t} &= -\frac{1}{2}\boldsymbol{t}^T\Big(\beta\boldsymbol{I}_N - \beta\boldsymbol{\Phi}(\alpha\boldsymbol{I}_M + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\beta\Big)\boldsymbol{t} \\
&= -\frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t} + \frac{\beta^2}{2}\boldsymbol{t}^T\boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} \\
&= -\frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t} + \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{A}\boldsymbol{m}_N \\
&= -\frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 - \frac{\alpha}{2}\boldsymbol{m}_N^T \\
&= -E(\boldsymbol{m}_N)
\end{aligned}$$

and we have already got the same solutions as the equation 18.

- Maximizing the evidence function 18 with respect to α. First define the following eigenvector equation

$$\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i \tag{19}$$

Thus, we can know that the $\boldsymbol{A}$ has eigenvalues $\alpha + \lambda_i$. Because

$$\begin{aligned}
\frac{d}{d\alpha}ln|\boldsymbol{A}| &= \frac{d}{d\alpha}ln\prod_i(\alpha + \lambda_i) \\
&= \frac{d}{d\alpha}\sum_i ln(\alpha + \lambda_i) \\
&= \sum_i \frac{1}{\alpha + \lambda_i}
\end{aligned}$$

Thus the stationary points of equation 18 with respect to α satisfy

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha} \tag{20}$$

So

$$\alpha\boldsymbol{m}_N^T\boldsymbol{m}_N = M - \alpha\sum_i \frac{1}{\lambda_i + \alpha} = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \gamma \tag{21}$$

which equals to

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T\boldsymbol{m}_N} \tag{22}$$

Note that this is an implicit solution for α not only because γ depends on α but also because the model $\boldsymbol{m}_N$ of the posterior distribution itself depends on the choice of α, which means we have to use the iterative procedure to get the optimal α.

- Maximizing the evidence function 18 with respect to β. We note that the eigenvalues λ_i is proportional to β, and hence $d\lambda_i/d\beta = \lambda_i/\beta$ giving

$$\frac{d}{d\beta} ln|\boldsymbol{A}| = \frac{d}{d\beta}\sum_i(\lambda_i + \alpha) = \frac{1}{\beta}\sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta} \tag{23}$$

So

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}(t_n - \boldsymbol{m}_N^T\phi(\boldsymbol{x}_n))^2 \tag{24}$$

which is also an implicit solution for β.

- The quantity γ measures the effective tital number of well determined parameters.

6 Limitations of Fixed Basis Functions

- The difficulty from the assumption that the basis functions $\phi_j(\boldsymbol{x})$ are fixed before the training data set is observed. And the number of the basis functions needs to grow rapidly, often exponentially, with the dimensionality of the input space.
- The data vector typically lie close to a nonlinear manifold whose intrinsic dimensionality is small.
- Target variables may have significant dependence on only a small number of possible directions within the data manifold.

7 Appendix

- *The matrix identity for the sequential data*

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} \tag{25}$$

- If $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices of size $N \times M$, then

$$|\boldsymbol{I}_N + \boldsymbol{A}\boldsymbol{B}^T| = |\boldsymbol{I}_M + \boldsymbol{A}^T\boldsymbol{B}| \tag{26}$$

a useful special case is

$$|\boldsymbol{I}_N + \boldsymbol{a}\boldsymbol{b}^T| = 1 + \boldsymbol{a}^T\boldsymbol{b}$$

To prove the equation 26, we notice that

$$\begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{0} \\ \boldsymbol{A}^T & \boldsymbol{I}_M \end{bmatrix}\begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{B} \\ -\boldsymbol{A}^T & \boldsymbol{I}_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_M + \boldsymbol{A}^T\boldsymbol{B} \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{B} \\ -\boldsymbol{A}^T & \boldsymbol{I}_M \end{bmatrix}\begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{0} \\ \boldsymbol{A}^T & \boldsymbol{I}_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_N + \boldsymbol{B}\boldsymbol{A}^T & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_M \end{bmatrix}$$

So we have

$$det\begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_M + \boldsymbol{A}^T\boldsymbol{B} \end{bmatrix} = det\begin{bmatrix} \boldsymbol{I}_N + \boldsymbol{B}\boldsymbol{A}^T & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_M \end{bmatrix}$$

which equals to

$$|\boldsymbol{I}_M + \boldsymbol{A}^T\boldsymbol{B}| = |\boldsymbol{I}_N + \boldsymbol{B}\boldsymbol{A}^T| = |\boldsymbol{I}_N + \boldsymbol{A}\boldsymbol{B}^T|$$

where we use the result $det\boldsymbol{A} = det\boldsymbol{A}^T$.