

## 1.2.6 Bayesian curve fitting

Note: the Gaussian curve problem demonstrated, can be/is solved analytically

In a fully Bayesian approach, we should consistently apply the sum and product rules of probability, which requires, as we shall see shortly, that we integrate over all values of  $w$ . Such marginalizations lie at the heart of Bayesian methods for pattern recognition.

given the training data  $\mathbf{x}$  and  $\mathbf{t}$ , along with a new test point  $x$ , and our goal is to predict the value of  $t$ .

we assume that the parameters  $\alpha$  and  $\beta$  are fixed and known in advance (in later chapters we shall discuss how such parameters can be inferred from data in a Bayesian setting).

predictive  
distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}. \quad (1.68)$$

$x$  = test data  
 $\mathbf{x}$  = training data

is the posterior distribution  
can be found by normalizing the right-hand side of (1.66).

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha). \quad (1.66)$$

posterior  $\propto$  likelihood  $\times$  prior

omitted the dependence on  $\alpha$  and  $\beta$  to simplify the notation

target values  $\mathbf{t} = (t_1, \dots, t_N)^T$ . (sometimes  $t$  is labels)

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j \quad (1.1)$$

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \quad (1.60)$$

$\beta$  corresponding to the inverse variance of the distribution.

## 1.2.6 Bayesian curve fitting (cont.)

solving 1.60:

■ solve for  $\mathbf{w}_{\text{ML}}$  by minimizing the negative log-likelihood:

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi). \quad (1.62)$$

$$\frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 = 0$$

$$\text{where } y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

One can use regularization to avoid overfitting by a polynomial order that is too high:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

where  $\|\mathbf{w}\|^2 \equiv \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$ , and the coefficient  $\lambda$  governs the relative importance of the regularization term compared with the sum-of-squares error term. Note that often the coefficient  $w_0$  is omitted from the regularizer because its

■ solve for  $\beta_{\text{ML}}$

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2. \quad (1.63)$$

## 1.2.6 Bayesian curve fitting (cont.)

then 1.60 becomes 1.64:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \quad (1.60)$$

$$p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1}). \quad (1.64)$$

<https://github.com/zjost/bayesian-linear-regression/blob/master/src/bayes-regression.ipynb>

Can be expressed using the parameters  $\mathbf{w}$  and basis functions  $\phi$ :  $\text{Norm}(t_n | \vec{w}^T \vec{\phi}(\vec{x}_n), \beta^{-1})$

for the prior of 1.66 consider a Gaussian over the polynomial coefficients of  $\mathbf{w}$ :

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2} \mathbf{w}^T \mathbf{w}\right\} \quad (1.65)$$
$$= \mathcal{N}(\mathbf{w} | \vec{m}_0, \vec{S}_0)$$

where  $\alpha$  is the precision of the distribution, and  $M+1$  is the total number of elements in the vector  $\mathbf{w}$  for an  $M^{\text{th}}$  order polynomial. Variables such as  $\alpha$ , which control

■ solve for  $\mathbf{w}$  by minimizing the negative log-likelihood of 1.66:

$$\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}. \quad (1.67)$$

Thus we see that maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function encountered earlier in the form (1.4), with a regularization parameter given by  $\lambda = \alpha/\beta$ .

Integration of 1.68 can be solved analytically:

predictive  
distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}. \quad (1.68)$$

$$= \mathcal{N}(t_n | \tilde{\mathbf{w}}^T \tilde{\phi}(\tilde{\mathbf{x}}_n), \beta^{-1}) \underbrace{\mathcal{N}(\mathbf{w} | \tilde{\mathbf{m}}_0, \tilde{\mathbf{S}}_0)}_{\substack{\text{prior} \\ \text{mean} \quad \text{cov}}}$$

<https://github.com/zjost/bayesian-linear-regression/blob/master/src/bayes-regression.ipynb>

$$= \text{Norm}(\tilde{\mathbf{w}} | \tilde{\mathbf{m}}_N, \tilde{\mathbf{S}}_N)$$

corrections to Bishop's 1.70-1.72 are present in Bishop's chap 3 and in ctgk PRML GitHub

Note that if data comes in sequentially, the posterior of the previous step becomes the prior of the current step and we only need to calculate the updates to  $\mathbf{m}_N$  and  $\mathbf{S}_N^{-1}$

$$\mathbf{w}_k = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}_k = \Phi^\dagger \mathbf{t}_k \quad (3.35)$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0) \quad (3.48)$$

eral result (2.116), which allows us to write down the posterior distribution directly in the form

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N) \quad (3.49)$$

where

$$\mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) \quad (3.50)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \Phi^T \Phi. \quad (3.51)$$

Note that because the posterior distribution is Gaussian, its mode coincides with its mean. Thus the maximum posterior weight vector is simply given by  $\mathbf{w}_{\text{MAP}} = \mathbf{m}_N$ . If we consider an infinitely broad prior  $\mathbf{S}_0 = \alpha^{-1} \mathbf{I}$  with  $\alpha \rightarrow 0$ , the mean  $\mathbf{m}_N$  of the posterior distribution reduces to the maximum likelihood value  $\mathbf{w}_{\text{ML}}$  given by (3.15). Similarly, if  $N = 0$ , then the posterior distribution reverts to the prior.

Furthermore, if data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point, such that the new posterior distribution is again given by (3.49).

For the remainder of this chapter, we shall consider a particular form of Gaussian prior in order to simplify the treatment. Specifically, we consider a zero-mean isotropic Gaussian governed by a single precision parameter  $\alpha$  so that

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}) \quad (3.52)$$

and the corresponding posterior distribution over  $\mathbf{w}$  is then given by (3.49) with

$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t} \quad (3.53)$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^T \Phi. \quad (3.54)$$

