Lecture 3: Straight Line in Bayesian Framework

Linear Regression

The equation of a straight line is:

$$y = mx + c \tag{1}$$

where m is the slope and c the intercept on the y–axis at $x = 0$. y is a predicted value derived by best–fitting the equation to a set of n experimental values of (t_i, x_i) for $i = 1, \ldots n$.

The same equation can be rewritten as

$$y = \sum_{j=1}^{2} w_j \phi_j \qquad \text{where} \qquad \phi_1 = x \qquad \text{and} \qquad \phi_2 = 1$$

$$w_1 \equiv m \qquad \text{and} \qquad w_2 \equiv c$$

Using the best–fit values of the weights w_1 and w_2 can, however, be misleading when dealing with finite sets of data. If a different finite–dataset is assembled from the population of data, then it is possible that a different set of weights will be obtained. This uncertainty in the line that best represents the entire population of data can be expressed by determining a distribution of the weights, rather than a single set of best–fit weights (Fig. 1a). A particular set of weights in the weight–space can be identified as a vector $\mathbf{w}^T = (w_1, w_2)$.

In the absence of data, we may have some prior beliefs about the variety of straight lines as illustrated in Fig. 1b. The distribution of

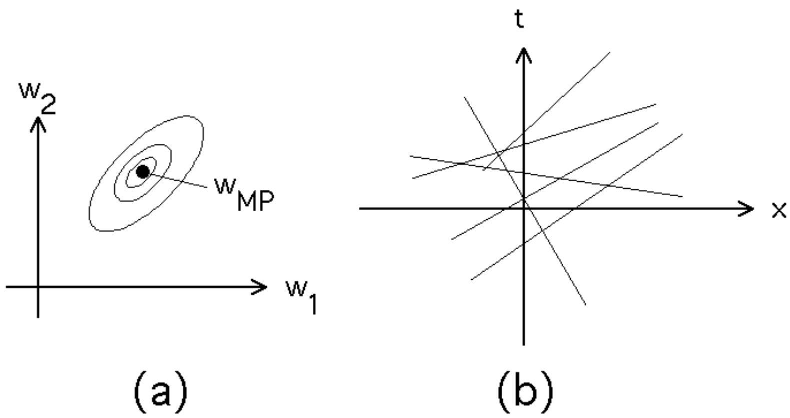

Fig. 1: (a) Weight space showing a distribution of weights about the most probable set. (b) Prior beliefs about straight line models.

lines is represented by a two–dimensional Gaussian with variables w_1 and w_2:

$$\mathrm{P}(\mathbf{w}) = \frac{1}{Z_w} \exp\left\{ -\frac{\alpha}{2} \sum_{j=1}^{2} w_j^2 \right\} \tag{2}$$

where Z_w is the usual normalising factor and $\alpha = 1/\sigma_w^2$, where σ_w^2 is the variance. (Notice that it appears that the prior favours smaller weights, but the value of α can be made sufficiently small to make the distribution approximately flat, so that the set of lines which fall within the distribution is roughly random.)

Now suppose that we have some experimental data (consisting of $(t_i, \mathbf{x}_i)$ with $i = 1 \ldots n$; t represents the measured values of y and is often referred to as the *target*) then it becomes possible to assign likelihoods to each of the lines using another two–dimensional Gaussian with variables w_i and w_2:

$$\mathrm{P}(\mathbf{T}|\mathbf{w}, \mathbf{X}) = \frac{1}{Z_D} \exp\left\{ -\frac{\beta}{2} \sum_{i=1}^{n} (t_i - y_i)^2 \right\} \tag{3}$$

where Z_D is the normalising factor and $\beta = 1/\sigma_\nu^2$, where σ_ν^2 is the variance.

The actual probability distribution of weights is then obtained by scaling the prior with the the likelihood,

$$\begin{aligned} P(\mathbf{w}|\mathbf{T},\mathbf{X}) &\propto P(\mathbf{T}|\mathbf{w},\mathbf{X}) \times P(\mathbf{w}) \\ &= \frac{\exp\{-M\{\mathbf{w}\}\}}{Z_M} \end{aligned} \tag{4}$$

where

$$M\{\mathbf{w}\} = \frac{\alpha}{2}\sum_i w_i^2 + \frac{\beta}{2}\sum_{m=1}^{N}\left(t_m - \sum_i w_i x_{m,i}\right)^2$$

Using a Taylor expansion about the most probable $\mathbf{w} = \mathbf{w}_{MP}$ gives

$$M\{\mathbf{w}\} \simeq M\{\mathbf{w}_{MP}\} + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{MP})^T \underbrace{\left[\alpha\mathbf{I} + \beta\sum_n \mathbf{x}\mathbf{x}^T\right]}_{\mathbf{V}^{-1}}(\mathbf{w} - \mathbf{w}_{MP})$$

where $\mathbf{I}$ is a 2×2 identity matrix and $\mathbf{V}$ is the variance–covariance matrix. This can be used to find the uncertainty in the prediction of $y\{\mathbf{x}\}$ at a particular location in the input space†:

$$\sigma_y^2 = \mathbf{x}^T\mathbf{V}\mathbf{x} \tag{5}$$

The variation in σ_y as a function of $\mathbf{x}$ is illustrated schematically in Fig. 2.

---

† equation 4 describes the Guassian distribution of weights whereas what we want is the variance in y. Using a Taylor expansion,

$$y = y\{\mathbf{w}_{MP}\} + \frac{\partial y}{\partial \mathbf{w}}\Delta\mathbf{w}$$

$$\sigma_y^2 = \frac{\partial y}{\partial \mathbf{w}}\sigma_w^2\frac{\partial y}{\partial \mathbf{w}} = \mathbf{x}^T\mathbf{V}\mathbf{x}$$

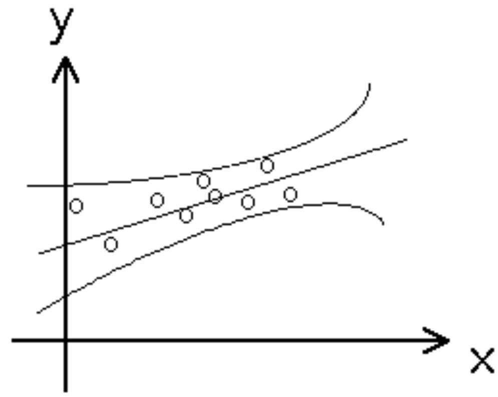

Fig. 2: Error bounds calculated using equation 5