1 Linear regression

In this problem, we are given a data set $D = \{x_m, t_m\}_{m \in \{1,2,\ldots,N\}}$, that is to say N points of $\Re^2$, which we want to represent by a function $y = ax + b$. We note further $y_m = ax_m + b$.

You surely have encountered the 'least-square fitting' method in which the error $E_D = \sum_{m=1}^{N} (t_m - y_m)^2$ is minimised to identify the most suitable parameters a and b. This is easily done by solving:

$$\begin{aligned} \frac{\partial E_D}{\partial a} &= 0 \\ \frac{\partial E_D}{\partial b} &= 0 \end{aligned}$$

which leads to:

$$\begin{aligned} a_{MP} &= \frac{NS_{XT} - S_X S_T}{NS_{XX} - {S_X}^2} \\ b_{MP} &= \frac{S_{XX} S_T - S_X S_{XT}}{NS_{XX} - {S_X}^2} \end{aligned}$$

where $S_X = \sum_{m=1}^{N} x_m$, $S_{XT} = \sum_{m=1}^{N} x_m t_m$, etc.

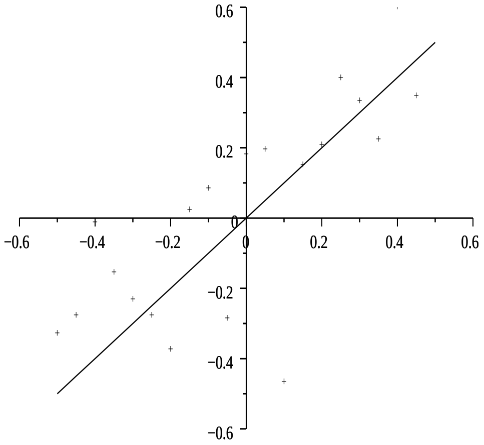

Figure 1: A dataset generated from $y = x$ by adding a random noise with gaussian distribution of variance 0.15.

2 Bayesian approach to the problem

As mentioned earlier, we can denote H our hypothesis 'the data can be fitted with a line, *ie* $y = ax + b$'. Our problem is to evaluate:

$$\mathrm{P}(a, b|D, H) = \frac{\mathrm{P}(D|a, b, H)\mathrm{P}(a, b|H)}{\mathrm{P}(D|H)}$$

Recall that $\mathrm{P}(a, b|H)$ is the *prior*, *ie* what we think of the values a and b might take before we have seen the data.

2.1 Choosing a prior

We will assume that the dataset has been normalised in $[-0.5 : 0.5]$ in both x and t. If the data fit perfectly a straight line, it should have $a = 1$ and $b = 0$, however if the noise level is high, these parameters may be significantly different. We might therefore want to chose a prior which is maximum for $a = 1$ and $b = 0$, or for simplicity a flat prior (that is, we assume that a and b take any value in given intervals with uniform probabilities).

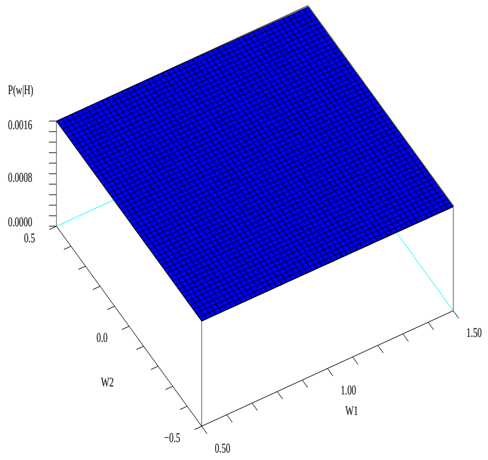

Figure 2: Prior on $w1$ (a) and $w2$ (b), flat over the region we would hope the parameters to fall in ($\sigma_w = 10$).

In this example, we will attribute to a and b the same prior as would be used in more complex regression methods (discussed later), that is:

$$\mathrm{P}(a, b|H) = \frac{1}{2\pi\sigma_w^2} \exp\left(-\frac{a^2 + b^2}{2\sigma_w^2}\right)$$

and let σ_w be large so that the prior will be almost flat in the region where we hope the parameters will be (a close to 1, b close to 0).

Note that this prior is chosen for convenience but will favour (if a small value of σ_w is chosen) values of a close to 0 rather than 1, which does not make sense.

2.2 The likelihood

We now need to evaluate $P(D|a,b,H)$, the probability of the data given some parameters a and b. What exactly can we mean by probability of data ? Certainly, the dataset is given ?

To attribute a probability to the data, we need to define a noise model: the targets t_m are in fact measurements of some 'real' values y_m with a noise ν_m.

$$t_m = y_m + \nu_m$$

where $y_m = ax_m + b$. If we can estimate the probability distribution of the noise that may occur during any measurement, we can in turn attribute a probability to the distance between the actual data t_m and the predicted value y_m which we think is the 'real' one.

This noise is typically attributed a gaussian distribution of zero mean and mean σ_ν:

$$\mathrm{P}(\nu_m) = \frac{1}{\sqrt{2\pi{\sigma_\nu}^2}} \exp\left(-\frac{{\nu_m}^2}{2{\sigma_\nu}^2}\right)$$

the probability of the data can therefore be thought of as:

$$\begin{aligned}\mathrm{P}(D|a,b,H) &= \mathrm{P}(\nu_1,\nu_2,\ldots,\nu_N|a,b,H) = \prod_{m=1}^{N} \mathrm{P}(\nu_m|a,b,H) \\ &= \frac{1}{(2\pi{\sigma_\nu}^2)^{N/2}} \exp\left(-\frac{\sum_{m=1}^{N} {\nu_m}^2}{2{\sigma_\nu}^2}\right) \\ &= \frac{1}{(2\pi{\sigma_\nu}^2)^{N/2}} \exp\left(-\frac{\sum_{m=1}^{N} (t_m - y_m)^2}{2{\sigma_\nu}^2}\right) \qquad (1)\end{aligned}$$

Note how the set (a_{MP}, b_{MP}) which minimises E_D as defined earlier, now maximises the likelihood.

2.3 The posterior distribution

Having defined the prior and likelihood, we have:

$$\mathrm{P}(a,b|D,H) = \frac{\mathrm{P}(D|a,b,H)\mathrm{P}(a,b|H)}{\mathrm{P}(D|H)} = \frac{\mathrm{P}(D|a,b,H)\mathrm{P}(a,b|H)}{\int_a\int_b \mathrm{P}(D|a,b,H)\mathrm{P}(a,b|H)\mathrm{d}a\mathrm{d}b}$$

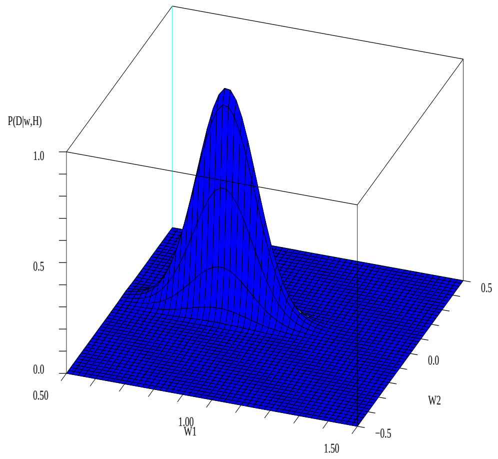

Figure 3: The likelihood $\mathrm{P}(D|\mathbf{w}, H)$ for the data in figure 1, having assumed $\sigma_\nu = 0.2$. Note that the curve has been normalised for easy representation. Given that the prior is flat over the region, the posterior distribution would look exactly the same as the likelihood.

If we are trying to evaluate the most probable parameters a_{MP} and b_{MP}, the normalising constant is irrelevant and it is sufficient to know that:

$$\mathrm{P}(a, b|D, H) \propto \exp\left(-\frac{E_D}{2{\sigma_\nu}^2}\right) \exp\left(-\frac{a^2+b^2}{2{\sigma_w}^2}\right) \tag{2}$$

To simplify further calculations, we approximate this distribution to its 2^{nd} order Taylor expansion around (a_{MP}, b_{MP}). Let us note

$$M(a, b) = (E_D/2{\sigma_\nu}^2) + ((a^2+b^2)/2{\sigma_w}^2)$$

giving the Taylor expansion around the minimum:

$$M(a, b) \approx M(a_{MP}, b_{MP}) + \frac{1}{2}\Delta\mathbf{w}^T A \Delta\mathbf{w} \tag{3}$$

where $A = \nabla\nabla(E_D + (a^2 + b^2))$ is evaluated in (a_{MP}, b_{MP}), with $\nabla = (\partial/\partial a, \partial/\partial b)$; $\Delta\mathbf{w}$ refers to the vector $(a - a_{MP}, b - b_{MP})$, and from now on, $\mathbf{w}$ refers to (a, b). With the help of 2 and 3, we have:

$$\mathrm{P}(\mathbf{w}|D, H) \approx \mathrm{P}(\mathbf{w}_{MP}|D, H) \exp\left(-\frac{1}{2}\Delta\mathbf{w}^T A \Delta\mathbf{w}\right) \tag{4}$$

which is a gaussian centred in $\mathbf{w}_{MP}$, of variance matrix A^{-1}.

2.4 Prediction and error bars

Calculation of y for a new datum can now be accompanied by an error bar: if we consider the Taylor expansion of y around $(\mathbf{w}_{MP})$:

$$y(x_{N+1}, \mathbf{w}) \approx y(x_{N+1}, \mathbf{w}_{MP}) + \Delta\mathbf{w}\, \nabla y|_{\mathbf{w}_{MP}} \tag{5}$$

With this linearisation, the distribution of $y(x_{N+1}, \mathbf{w})$ is conditioned by that of $\mathbf{w}$, and its variance is:

$$\nabla y|_{\mathbf{w}_{MP}}^{T} A^{-1} \nabla y|_{\mathbf{w}_{MP}}$$

This is analog to the case of a single random variable $Z' = \alpha Z + \beta$ for which:

$$\mathrm{var}(Z') = \alpha^2 \mathrm{var}(Z)$$