

## Linear regression

When we have multidimensional inputs, we consider linear functions of the form:

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

where  $\phi$ s are basis function and  $w_0$  is a bias.

**Big Warning:** the book uses wacky notation. Everyone I know just encodes the basis functions into  $\mathbf{x}$ , and calls  $w_0$  “ $b$ ” for bias. Everyone else calls the function  $f$  or  $h$  instead of  $y$ . Thus, the more standard way to write this is:

$$f(\mathbf{x}, \mathbf{w}) = b + \sum_d w_d x_d \quad (2)$$

$$= \mathbf{w}^\top \mathbf{x} + b \quad (3)$$

The key idea in this section is the relationship between least squares regression and maximum likelihood estimation under Gaussian error.

Let’s consider regressing on house price as a function of square footage:

Square footage	Price
1200	\$120
1340	\$125
1390	\$105
1400	\$130
1420	\$135
1500	\$145
1550	\$160
1700	\$155
1900	\$140
2150	\$130
2300	\$135

Linear prediction model:

$$\text{price} \approx w_0 + w_1 \times \text{square footage} \quad (4)$$

We can write this as:

$$[w_1 \quad w_0] \begin{bmatrix} 1200 & 1 \\ 1340 & 1 \\ 1390 & 1 \\ 1400 & 1 \\ 1420 & 1 \\ 1500 & 1 \\ 1550 & 1 \\ 1700 & 1 \\ 1900 & 1 \\ 2150 & 1 \\ 2300 & 1 \end{bmatrix}^\top = \begin{bmatrix} 120 \\ 125 \\ 105 \\ 130 \\ 135 \\ 145 \\ 160 \\ 155 \\ 140 \\ 130 \\ 135 \end{bmatrix} \quad (5)$$

It is easy to check that no  $\{w_0, w_1\}$  exists that satisfies this.

Define a cost function, *least-squares*:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N [\mathbf{w}^\top x_n - y_n]^2$$

This cost function penalizes outliers.

Now, we've changed the learning problem to an optimization problem: find  $\mathbf{w}$  to minimize  $E(\mathbf{w})$ .

It turns out that we can actually obtain a solution in closed form. Let  $\mathbf{X}$  be the data matrix, let  $\mathbf{y}$  be a (column) vector containing the targets. Then  $\mathbf{X}\mathbf{w} - \mathbf{y}$  is a column vector whose  $n$ th element is  $\mathbf{w}^\top x_n - y_n$ . So:

$$E(\mathbf{w}) = \frac{1}{2} [\mathbf{X}\mathbf{w} - \mathbf{y}]^\top [\mathbf{X}\mathbf{w} - \mathbf{y}]$$

Then, we can compute the gradient:

$$\begin{aligned} \nabla_{\mathbf{w}} E(\mathbf{w}) &= \nabla_{\mathbf{w}} \frac{1}{2} [\mathbf{X}\mathbf{w} - \mathbf{y}]^\top [\mathbf{X}\mathbf{w} - \mathbf{y}] \\ &= \frac{1}{2} \nabla_{\mathbf{w}} [\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{y}^\top \mathbf{y}] \\ &= \frac{1}{2} \nabla_{\mathbf{w}} \text{tr} [\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{y}^\top \mathbf{y}] \\ &= \frac{1}{2} \nabla_{\mathbf{w}} [\text{tr} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2 \text{tr} \mathbf{y}^\top \mathbf{X}\mathbf{w}] \\ &= \frac{1}{2} [\mathbf{X}^\top \mathbf{X}\mathbf{w} + \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}] \\ &= \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} \end{aligned}$$

Thus, setting the gradient equal to zero, we obtain:

$$\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$$

So:

$$\mathbf{w} = [\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{y}$$

### Maximum Likelihood

An alternative formulation:  $y = \mathbf{w}^\top x + \epsilon$ , where  $\epsilon \sim \mathcal{Nor}(0, \sigma^2)$ . Then  $y \sim \mathcal{Nor}(\mathbf{w}^\top x, \sigma^2)$ . Now, find  $\mathbf{w}$  to maximize likelihood of the training set.

$$\begin{aligned} L(D; \mathbf{w}) &= \prod_n \mathcal{Nor}(y_n \mid \mathbf{w}^\top x_n, \sigma^2) \\ &= \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y_n - \mathbf{w}^\top x_n)^2\right] \end{aligned}$$

Take log to yield *log-likelihood*:

$$\begin{aligned} \ell(D; \mathbf{w}) &= \log L(D; \mathbf{w}) \\ &= \sum_n \left\{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \mathbf{w}^\top x_n)^2 \right\} \\ &= -m \frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_n (y_n - \mathbf{w}^\top x_n)^2 \end{aligned}$$

The  $\mathbf{w}$  that maximizes this is clearly the same as in the least-squares formulation!