Linear least-square regression

Thibaud Taillefumier

1 Problem set-up

Suppose we have a m-dimensional vector $\boldsymbol{y} = \{y_1, \dots, y_m\}$ whose components represent scalar output measurements to be related n m-dimensional input vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. We want to find weights $\boldsymbol{w}$ so that one can predict the measurement outcomes $\boldsymbol{y}$ as a linear combination of input vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$:

$$\boldsymbol{y} \approx \sum_{i=1}^{n} w_i \boldsymbol{x}_i .$$

In principle, we can repeat measurement at will so that m can be very large, whereas n is set by the complexity of the model and should be assumed comparatively small $n < m$. Because the vector y lies into a much larger m-dimensional space than the at most n-dimensional space spanned by $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, it is in general impossible to perfectly reconstruct $\boldsymbol{y}$. For this reason, the of linear least-square regression is to minimize the prediction error rather than achieving perfect reconstruction. Specifically, in linear least-square regression, we look for weights $\boldsymbol{w}^\star$ that minimize the squared prediction error $E(\boldsymbol{w})$, which can be formally stated as follows:

$$\boldsymbol{w}^\star = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w}) \quad \text{with} \quad E(\boldsymbol{w}) = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right)^2 .$$

The squared prediction error $E(\boldsymbol{w})$ can be interpreted geometrically as the squared Euclidean length of the residual vector defined by $\boldsymbol{y} - \sum_{i=1}^{m} w_i \boldsymbol{x}_i$. Thus $E(\boldsymbol{w})$ can also be written as

$$E(\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|^2 ,$$

where $\|\boldsymbol{u}\|$ denotes the length of vector $\boldsymbol{u}$ and where X is the matrix whose columns are $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. As expected, the squared prediction error is a non-negative number that is zero should the output $\boldsymbol{y}$ lies in the span of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. However, we know that perfect reconstruction is in general impossible due to the dimensionality mismatch $m > n$ and we resort to look for weights $\boldsymbol{w}^\star$ that minimize the Euclidean square length of the residual vector. We are going to fulfill this program in the next section via a combination of calculus and linear algebra. Before we go ahead, remember that the nexus of the problem stems from the dimensionality mismatch $m < n$, which can be stated concretely by saying that the matrix X has many more rows than columns.

2 Solution via calculus and linear algebra

The first step to linear least-square regression is to compute the derivative of the squared prediction error E with respect to the weight w_k, while holding all the other weights fixed:

$$\begin{aligned}\frac{\partial E(\boldsymbol{w})}{\partial w_k} &= \sum_{i=1}^{m} \frac{\partial}{\partial w_k} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right)^2 , \\ &= 2 \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right) \left[\frac{\partial}{\partial w_k} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right) \right] .\end{aligned}$$

The term in between square bracket is actually much simpler than it looks as it is the derivative of a linear function of w_k with linear coefficient x_{ik}.

$$\frac{\partial}{\partial w_k} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right) = x_{ik} \, .$$

The weights $\boldsymbol{w}^\star$ that minimized the squared prediction error E are those weights for which the derivatives of E with respect to any w_k is zero. Based on our computation of the derivative of the squared prediction error, this means that the weights $\boldsymbol{w}^\star$ satisfy the following set n linear equations:

$$\frac{\partial E(\boldsymbol{w})}{\partial w_k} = 2 \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right) x_{ik} = 0 \, , \quad \text{with} \quad 1 \leq k \leq n \, . \tag{1}$$

The above set of equations can be conveniently expressed in matrix form by using the transpose operation. Indeed, using the fact that $x_{ik} = (X^T)_{ki}$, we can write the system of equation in matrix form as

$$X^T (\boldsymbol{y} - X\boldsymbol{w}) = 0 \, ,$$

showing that the weights $\boldsymbol{w}^\star$ are solution of the matrix equation

$$\left(X^T X\right) \boldsymbol{w} = X^T \boldsymbol{y}\,.$$

The matrix $\left(X^T X\right)$ is a n-by-n square matrix that is invertible if $n \leq m$ (which is true) and if the matrix X has rank n, i.e. if the columns of X are linearly independent (which need to be checked). Under this assumption of invertibility, the weight $\boldsymbol{w}^\star$ are obtained via matrix inversion:

$$\boldsymbol{w}^\star = \left(X^T X\right)^{-1} X^T \boldsymbol{y}\,.$$

Moreover the best approximation to the original vector $\boldsymbol{y}$, denoted by $\boldsymbol{y}^\star$, is given by:

$$\boldsymbol{y}^\star = X\boldsymbol{w}^\star = X \left(X^T X\right)^{-1} X^T \boldsymbol{y}\,.$$