Linear least-square regression

Thibaud Taillefumier

1 Problem set-up

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

\[
\mathbf{y} \approx \sum_{i=1}^{n} w_i \mathbf{x}_i.
\]

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 \( \mathbf{y} \) lies into a much larger \( m \)-dimensional space than the at most \( n \)-dimensional space spanned by \( \mathbf{x}_1, \ldots, \mathbf{x}_n \), it is in general impossible to perfectly reconstruct \( \mathbf{y} \). For this reason, the objective 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 \( \mathbf{w}^{\star} \) that minimize the squared prediction error \( E(\mathbf{w}) \), which can be formally stated as follows:

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

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

\[
E(\mathbf{w}) = \| \mathbf{y} - X\mathbf{w} \|^2,
\]

where \( \| \mathbf{u} \| \) denotes the length of vector \( \mathbf{u} \) and where \( X \) is the matrix whose columns are \( \mathbf{x}_1, \ldots, \mathbf{x}_n \). As expected, the squared prediction error is a non-negative
number that is zero should the output \( y \) lies in the span of \( x_1, \ldots, x_n \). However, we know that perfect reconstruction is in general impossible due to the dimensionality mismatch \( m > n \) and we resort to look for weights \( w^* \) 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:

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

\[
= 2 \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right) \left[ \frac{\partial}{\partial w_k} \left(y_i - \sum_{j=1}^{n} x_{ij} w_j \right) \right].
\]

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 \( w^* \) 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 \( w^* \) satisfy the following set \( n \) linear equations:

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

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

\[
X^T (y - Xw) = 0,
\]
showing that the weights $w^*$ are solution of the matrix equation

$$(X^T X) w = X^T y.$$ 

The matrix $(X^T X)$ 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 $w^*$ are obtained via matrix inversion:

$$w^* = (X^T X)^{-1} X^T y.$$ 

Moreover the best approximation to the original vector $y$, denoted by $y^*$, is given by:

$$y^* = X w^* = X (X^T X)^{-1} X^T y.$$