Linear regression

Louis Wehenkel & Pierre Geurts

Institut Montefiore, University of Liège, Belgium

ELEN062-1
Introduction to Machine Learning
October 12, 2021

Outline

- Objects (or observations): $LS = \{o_1, \ldots, o_N\}$
- Attribute vector: $\boldsymbol{a}^i = (a_1(o_i), \ldots, a_n(o_i))^T$, $\forall i = 1, \ldots, N.$
- Attribute values: $\boldsymbol{a}_j = (a_j(o_1), \ldots, a_j(o_N))^T$ $\forall j = 1, \ldots, n.$
- Outputs: $y^i = y(o_i)$ or $c^i = c(o_i)$, $\forall i = 1, \ldots, N.$
- LS Table

o	$a_1(o)$	$a_2(o)$	$\ldots$	$a_n(o)$	$y(o)$
1	a_1^1	a_2^1	$\ldots$	a_n^1	y^1
2	a_1^2	a_2^2	$\ldots$	a_n^2	y^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	a_1^N	a_2^N	$\ldots$	a_n^N	y^N

- LS attribute matrix: $A = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^N)$ (n lines, N columns)
- LS ouput column: $\boldsymbol{y} = (y^1, \ldots, y^N)^T$

Linear regression models

- Output is numerical scalar
- All inputs are numerical scalars
- Linear regression tries to approximate output by

$$\hat{y}(o) = w_0 + \sum_{i=1}^{n} w_i a_i(o)$$

- Supervised learning problem:

  Choose the parameters $w_0, w_1, \ldots, w_n$ *so as to fit well LS and have good generalization to unseen objects*

Linear regression models

Linear in the parameters, not necessarily in the original inputs.

$$\hat{y}(o) = w_0 + \sum_{i=1}^{k} w_i \phi_i(\boldsymbol{a}(o))$$

Inputs can come from different sources:

- quantitative measurements
- transformations of quantitative measurements (log, square-root, etc.)
- basis expansions, such as $a_2(o) = a_1^2(o), a_2(o) = a_1^3(o)$, etc.
- numeric or "dummy" coding of qualitative inputs

Least mean square error solution

Posing, $a_0(o) = 1, \forall o$ and denoting by

1. $\boldsymbol{a}'(o_i) = (a_0(o_i), a_1(o_i), \ldots, a_n(o_i))^T$, and
2. $\boldsymbol{w}' = (w_0, w_1, \ldots, w_n)^T$, square error (SE) at o_i is defined by

$$SE(o_i, \boldsymbol{w}') = (y(o_i) - \hat{y}(o_i))^2 = \left(y(o_i) - \boldsymbol{w}'^T \boldsymbol{a}'(o_i)\right)^2$$

and the total squared error (TSE) by

$$TSE(LS, \boldsymbol{w}') = \sum_{i=1}^{N} \left(y(o_i) - \boldsymbol{w}'^T \boldsymbol{a}'(o_i)\right)^2$$

or in vector notation (denoting by $A' = (\boldsymbol{a}'^1, \ldots, \boldsymbol{a}'^N)$)

$$TSE(LS, \boldsymbol{w}') = \left(\boldsymbol{y} - A'^T \boldsymbol{w}'\right)^T \left(\boldsymbol{y} - A'^T \boldsymbol{w}'\right)$$

Least mean square error solution

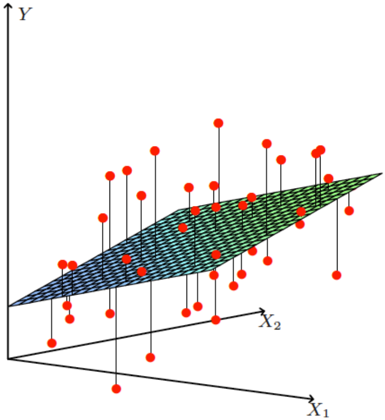


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Least mean square error solution: one dimension

Assuming only one input, the solution is computed as:

$$(w_0^*, w_1^*) = \arg\min_{w_0, w_1} \sum_{i=1}^{N} (y(o_i) - w_0 - w_1 a_1(o_i))^2$$

Canceling the derivative with respect to w_0 and w_1, one gets:

$$\begin{aligned} w_1^* &= \frac{\sum_{i=1}^{N}(a_1(o_i) - \overline{a}_1)(y(o_i) - \overline{y})}{\sum_{i=1}^{N}(a_1(o_i) - \overline{a}_1)^2} = \frac{\mathrm{cov}(a_1, y)}{\sigma_{a_1}^2} \\ w_0^* &= \overline{y} - w_1^* \overline{a}_1 \end{aligned}$$

where $\overline{a}_1 = N^{-1} \sum_{k=1}^{N} a_1(o_k)$ and $\overline{y} = N^{-1} \sum_{k=1}^{N} y(o_k)$.

Substituting the above into $\hat{y}(o) = w_0^* + w_1^* a_1(o)$:

$$\frac{\hat{y}(o) - \overline{y}}{\sigma_y} = \rho_{a_1, y} \frac{a_1(o) - \overline{a}_1}{\sigma_{a_1}},$$

with $\rho_{a_1,y}$ the correlation between a_1 and y, and σ_y, σ_{a_1} the standard deviations of y and a_1, all computed on the LS.

Least mean square error solution: multidimensional case

Choose $\boldsymbol{w}'$ to minimize

$$TSE(LS, \boldsymbol{w}') = \left(\boldsymbol{y} - A'^T \boldsymbol{w}'\right)^T \left(\boldsymbol{y} - A'^T \boldsymbol{w}'\right).$$

Differentiating w.r.t. $\boldsymbol{w}'$ (gradient)

$$\nabla_{w'} TSE(LS, \boldsymbol{w}') = -2A'(\boldsymbol{y} - A'^T \boldsymbol{w}')$$

and solving for $\nabla_{w'} TSE(LS, \boldsymbol{w'}^*) = 0$ we obtain

$$\boldsymbol{w'}^* = \left(A' A'^T\right)^{-1} A' \boldsymbol{y}$$

Note that $\nabla^2_{w'} TSE(LS, \boldsymbol{w}') = 2A' A'^T$ is symmetric positive (semi-) definite.

Shift invariance: suppose we define new attribute vector by $\boldsymbol{a}_c(o) = \boldsymbol{a}(o) + \boldsymbol{c}$ where $\boldsymbol{c}$ is a constant vector (i.e. independant of object).

Let $(w_0, \boldsymbol{w})$ be the optimal solution in the original attribute space. Then it is easy to see that $(w_0 - \boldsymbol{w}^T\boldsymbol{c}, \boldsymbol{w})$ is optimal in the new space.

Indeed, we have

$$\hat{y}_c(o) = w_0 - \boldsymbol{w}^T\boldsymbol{c} + \boldsymbol{w}^T\boldsymbol{a}_c(o) = w_0 + \boldsymbol{w}^T\boldsymbol{a}(o) = \hat{y}(o).$$

Hence, if $(w_0 - \boldsymbol{w}^T\boldsymbol{c}, \boldsymbol{w})$ is not optimal in the new space, $(w_0, \boldsymbol{w})$ couldn't be optimal in the original space.

Let us discuss the meaning of the table $(A'A'^T)$: element i, j is obtained by the scalar product of line i and line j of matrix A'. Thus we have

$$A'A'^T = N \left(\begin{array}{c|ccc} 1 & \overline{a}_1 & \dots & \overline{a}_n \\ \hline \overline{a}_1 & g_{1,1} & \dots & g_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{a}_n & g_{n,1} & \dots & g_{n,n} \end{array} \right)$$

where $\overline{a}_i = N^{-1} \sum_{k=1}^{N} a_i(o_k)$ and $g_{i,j} = N^{-1} \sum_{k=1}^{N} a_i(o_k) a_j(o_k)$

Assuming that the attributes have all a zero mean ($\overline{a}_i = 0$) we have $g_{i,j} = cov(a_i, a_j)$

Least mean square error solution

In the sequel we will use the notation Σ to denote the covariance matrix.

Thus if all the attributes are centered, we have

$$\boldsymbol{w'}^* = \begin{pmatrix} N^{-1} & \mathbf{0} \\ \mathbf{0} & N^{-1}\Sigma^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ A \end{pmatrix} \boldsymbol{y}.$$

In particular, $w_0^* = N^{-1} \sum_{k=1}^{N} y^k = N^{-1} \sum_{k=1}^{N} y(o_k) = \overline{y}$.

In other words, if both a_i and y are centered, $w_0^* = 0$.

Least mean square error solution

Assuming that the attributes have zero mean and unit variance ($g_{i,i} = 1$), we have

$$A'A'^T = N \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & \rho_{1,1} & \dots & \rho_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \rho_{n,1} & \dots & \rho_{n,n} \end{array} \right)$$

Note that $\rho_{i,i} = 1; \forall i = 1, \dots, n.$

- In this case the correlation and covariance matrices are identical.
- Pre-whiten the attributes before solving the linear system.
- Below, we assume attributes are pre-whitened and drop suffix $'$.

Least mean square error solution

Let us take a non-singular $n \times n$ matrix B and define the transformed attribute vector by $\boldsymbol{a}_B(o) = B\boldsymbol{a}(o)$.

For the transformed attributes, matrix A becomes matrix BA, and solution becomes:
$\boldsymbol{w}_B = ((BA)(BA)^T)^{-1}BA\boldsymbol{y} = (B^T)^{-1}(AA^T)^{-1}B^{-1}BA\boldsymbol{y} = B^{T^{-1}}\boldsymbol{w}$

In other words,
$\hat{y}_B = \boldsymbol{w}_B^T\boldsymbol{a}_b = (B^{T^{-1}}\boldsymbol{w})^T B\boldsymbol{a} = \boldsymbol{w}^T B^{-1}B\boldsymbol{a} = \boldsymbol{w}^T\boldsymbol{a}.$

$\Rightarrow$ Invariance with respect to (non-singular) linear transformation

Least mean square error solution

Discussion of matrix $N\Sigma = AA^T$: computation, singularity, inversion.

1. It is easy to see that $N\Sigma = \sum_{i=1}^{N} \boldsymbol{a}(o_i)\boldsymbol{a}^T(o_i)$.
2. Therefore, rank of Σ is at most N.
3. Thus, if $n > N$, Σ is rank deficient (and hence singular).
4. If Σ is singular, unicity of optimal solution is lost, but existence is preserved.
5. Need to impose other criteria to find unique solution, i.e. to build algorithm.
6. Several such solutions are discussed in the reference book, in particular regularization.

Regularization of least mean square error solution

Instead of choosing $\boldsymbol{w}$ to minimize

$$TSE(LS,\boldsymbol{w}) = \left(\boldsymbol{y} - A^T\boldsymbol{w}\right)^T \left(\boldsymbol{y} - A^T\boldsymbol{w}\right).$$

Let us minimize w.r.t. $\boldsymbol{w}$ and for given $\lambda > 0$

$$TSE_R(LS,\lambda,\boldsymbol{w}) = \left(\boldsymbol{y} - A^T\boldsymbol{w}\right)^T \left(\boldsymbol{y} - A^T\boldsymbol{w}\right) + \lambda \boldsymbol{w}^T\boldsymbol{w}$$

Differentiating w.r.t. $\boldsymbol{w}$ yields (I denotes the $n \times n$ identity matrix)

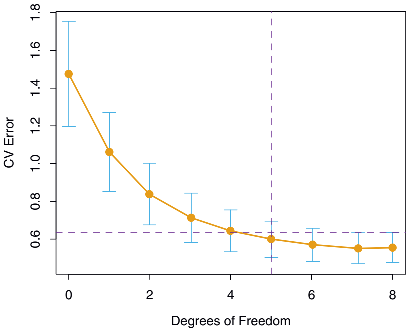
$$\nabla_w TSE_R(LS,\boldsymbol{w},\lambda) = -2A\left(y - A^T\boldsymbol{w}\right) + 2\lambda I\boldsymbol{w}$$

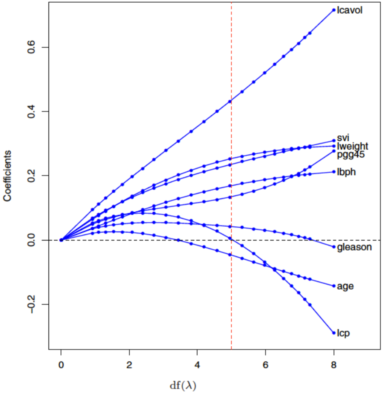
in other words

$$\boldsymbol{w}^*(\lambda) = \left(AA^T + \lambda I\right)^{-1} A\boldsymbol{y}$$

which has a unique solution, $\forall\lambda > 0$!

Illustration: effect of λ on CV error and optimal weights





(See Figures 3.7 and 3.8 in reference book)

$df(\lambda) = n$ when $\lambda = 0$ and $df(\lambda) \to 0$ when $\lambda \to \infty$

Algorithmics

Computational complexity:

- Building the covariance matrix: in the order of Nn^2 operations
- Solving the system for $\boldsymbol{w}^*$: in the order of n^3 operations

Various alternative techniques exist to solve system.

Some will be discussed in the sequel.

Other regularizations

- The above regularization method is called *Ridge Regression*. It belongs to the family of *shrinkage methods*.
- Other regularization for linear regression models:
  - LASSO: a shrinkage method replacing $\sum_i w_i^2 < t$ by $\sum_i |w_i| < t$ (discussed later in the course).
  - Subset selection: select an optimal subset of input attributes on which to regress. Various heuristics exist to determine the subset.

Residual fitting (a.k.a. Forward-Stagewise Regression)

Residual fitting: alternative algorithm, of general interest

- Start by computing w_0 for the no-variable case: $w_0 = \overline{y}$
- Introduce attributes (assumed of zero mean, unit variance) progressively, one at the time
  - Define residual at step k by
    $\Delta_k y(o) = y(o) - w_0 - \sum_{i=1}^{k-1} w_i a_i(o)$
  - Find best fit of residual with only attribute a_k:

    $w_k = \rho_{a_k, \Delta_k y} \sigma_{\Delta_k y}$.

    (since residuals have zero mean, and attributes are pre-whitened)

Note that this algorithm is in general suboptimal w.r.t. to the direct solution given previously, but it is linear in the number of attributes.

References

Chapter 3 from the reference book (Hastie *et al.*, 2009):

- Section 3.2: Linear regression models and least squares
- Section 3.4.1: Ridge regression
- Section 3.3.3: Forward-stagewise regression

Frequently asked questions

- How to choose a value for λ ?
- Asymptotic ($N \to \infty$) properties of LR and Ridge-regression
- Discuss LASSO vs Ridge regression
- Discuss computational complexity and interpretability