Linear Regression Normal Equation – Additional Results

Pawel Wocjan

January 9th, 2019

Abstract

We derive the normal equation for linear regression.

1 Normal equation

To simplify the discussion, consider first the case that the bias of the linear regression model is set to 0, that is, only the weights $w_1, \ldots, w_n$ are trained.

Let $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}) \in \mathbb{R}^n \times \mathbb{R}$ be the training examples. Set

$$X = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

and

$$y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} \in \mathbb{R}^{m}.$$

Theorem 1. *The optimal weight vector $w = (w_1, \ldots, w_n)^T \in \mathbb{R}^n$, that is, the one that minimizes the mean squared error is given by the formula*

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

This is proved in [1, 5.1.4 Example: Linear Regression]. I have derived some additional results so you can understand every step of the proof.

2 Additional results

We introduce some abbreviations. Let $[n] = \{1, \dots, n\}$. Let ∂w_r denote the partial derivative operator

$$\frac{\partial}{\partial w_r}.$$

Lemma 1. *Let $A = (a_{rs}) \in \mathbb{R}^{n\times n}$ be an arbitrary symmetric matrix. Let $w = (w_1, \dots, w_n)^T \in \mathbb{R}^n$ be an arbitrary column vector. Define the function $f(w) = w^T A w$. We have*

$$\nabla_w f(w) = 2Aw\,.$$

Proof. The right hand side is the column vector whose entries are given by

$$2\sum_{s=1}^{n} a_{rs} w_s.$$

for $r \in [n]$. This follows simply by carrying out the matrix-vector-multiplication.

The left hand side is the column vector whose entries are the partial derivatives

$$\partial w_r f(w)$$

for $r \in [n]$. We have

$$\begin{aligned}
\partial w_r f(w) &= \partial w_r \left(\sum_{t,s=1} w_t a_{ts} w_s \right) \\
&= \partial w_r \left(w_r^2 a_{rr} + 2 \sum_{s\neq r} w_r a_{rs} w_s \right) \\
&= 2 w_r a_{rr} + 2 \sum_{s\neq r} a_{rs} w_s \\
&= 2 \sum_{s=1}^{n} a_{rs} w_s.
\end{aligned}$$

We use that

- either t and s are both equal to r
- or t is equal to r and s is not equal to r.

Otherwise the partial derivative $\partial w_r(w_t a_{ts} w_s)$ is equal to 0.

Lemma 2. *Let* $w = (w_1, \ldots, w_n)^T \in \mathbb{R}^n$ *be an arbitrary column vector. Let* $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ *be an arbitrary column vector. Define the function* $g(w) = w^T v$. *We have*

$$\nabla_w g(w) = v\,.$$

Proof. This is easy. Prove it yourself.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, 2006, `http://www.deeplearningbook.org`