

# Linear Regression Normal Equation – Additional Results

Pawel Wocjan

January 14th, 2019

## Abstract

We derive the normal equation for linear regression and show that the mean-squared-error is a convex function.

## 1 Notation

Let  $(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(m)}, y^{(m)}) \in \mathbb{R}^n \times \mathbb{R}$  denote the collection of training examples. The  $i$ th training example consists of the feature vector  $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})^T \in \mathbb{R}^n$  and the label  $y^{(i)} \in \mathbb{R}$ .

Set

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & \dots & x_n^{(2)} \\ \vdots & \ddots & \vdots \\ x_1^{(m)} & \dots & x_n^{(m)} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

and

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

$\mathbf{X}$  is called the design matrix. Its rows correspond to the feature vectors of the training examples.

## 2 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 weight vector  $\mathbf{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$  needs to be determined.

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

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

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

### 3 Additional results

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

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

**Lemma 1** (Gradient of quadratic form). *Let  $\mathbf{A} = (a_{rs}) \in \mathbb{R}^{n \times n}$  be an arbitrary symmetric matrix and  $\mathbf{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$  an arbitrary column vector. Define the quadratic form  $f(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w}$ . Its gradient is given by*

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{A} \mathbf{w}.$$

*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 of the above equation is the column vector whose entries are the partial derivatives

$$\partial w_r f(\mathbf{w})$$

for  $r \in [n]$ . This follows from the definition of the nabla operator

$$\nabla_{\mathbf{w}} = \begin{pmatrix} \partial w_1 \\ \vdots \\ \partial w_n \end{pmatrix}.$$

We have

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

We use that either  $t = r$  and  $s = r$  or  $t = r$  and  $s \neq r$ . Otherwise the partial derivative  $\partial w_r(w_t a_{ts} w_s)$  is equal to 0. We also use that  $\mathbf{A}$  is symmetric, that is,  $a_{rs} = a_{sr}$ .

**Lemma 2.** Let  $\mathbf{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$  and  $\mathbf{v} = (v_1, \dots, v_n)^T \in \mathbb{R}^n$  be an arbitrary column vectors. Define the function  $g(\mathbf{w}) = \mathbf{w}^T \mathbf{v}$ . Its gradient is given by

$$\nabla_{\mathbf{w}} g(\mathbf{w}) = \mathbf{v}.$$

*Proof.* This is easy. Prove it yourself.

## 4 Proof of normal equation

Let  $\hat{y}^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)}$  denote the prediction of the linear regression model with weight vector  $\mathbf{w}$  when given the  $i$ th feature vector  $\mathbf{x}^{(i)}$ .<sup>1</sup> Define

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(n)} \end{pmatrix}$$

The mean squared error (MSE) is given by

$$\begin{aligned} \text{MSE} &= \frac{1}{m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2 \\ &= \frac{1}{m} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2, \end{aligned}$$

so the error increases whenever the Euclidean distance between the predictions and the targets (labels) increases.

To minimize MSE, we can simply solve for where its gradient is  $\mathbf{0}$ :

$$\nabla_{\mathbf{w}} \text{MSE} = \mathbf{0} \tag{1}$$

$$\Rightarrow \nabla_{\mathbf{w}} \frac{1}{m} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 = \mathbf{0} \tag{2}$$

$$\Rightarrow \frac{1}{m} \nabla_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \hat{\mathbf{y}}\|_2^2 = \mathbf{0} \tag{3}$$

$$\Rightarrow \frac{1}{m} \nabla_{\mathbf{w}} (\mathbf{X}\mathbf{w} - \hat{\mathbf{y}})^T (\mathbf{X}\mathbf{w} - \hat{\mathbf{y}}) = \mathbf{0} \tag{4}$$

$$\Rightarrow \frac{1}{m} \nabla_{\mathbf{w}} (\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}) = \mathbf{0} \tag{5}$$

$$\Rightarrow 2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y} = \mathbf{0} \tag{6}$$

$$\Rightarrow \mathbf{w} = (\mathbf{X}^T \mathbf{X} \mathbf{w})^{-1} \mathbf{X}^T \mathbf{y} \tag{7}$$

---

<sup>1</sup>Note that  $\hat{y}^{(i)}$  is the inner product between the weight vector  $\mathbf{w}$  and the feature vector  $\mathbf{x}^{(i)}$ . We also have  $\hat{y}^{(i)} = (\mathbf{x}^{(i)})^T \mathbf{w}$  because the inner product is symmetric.

In eq. (5), we can use Lemma 1 with  $\mathbf{A} = \mathbf{X}^T \mathbf{X}$  and Lemma 2 with  $\mathbf{v} = \mathbf{X}^T \mathbf{y}$  to compute the gradient. The system of equations whose solutions is given by eq. 7 is known as normal equations.

The general case of linear regression with non-zero bias  $b$  can also be solved with the help of the normal equation. Define the augmented weight vector  $\mathbf{w}_b = (b, w_1, \dots, w_n)^T \in \mathbb{R}^{n+1}$  and the augmented feature vectors  $\mathbf{x}_b^{(i)} = (1, x_1^{(i)}, \dots, x_n^{(i)})^T \in \mathbb{R}^{n+1}$ . We have

$$\hat{y}^{(i)} = \mathbf{w}_b^T \mathbf{x}_b^{(i)} = \mathbf{w}^T \mathbf{y}^{(i)} + b.$$

## 5 Proof of convexity of mean-squared-error

We now show that MSE is convex in  $\mathbf{w}$ . We can write

$$\text{MSE}(\mathbf{w}) = \phi(\|\mathbf{X}\mathbf{w} - \mathbf{y}\|),$$

where  $\phi : [0, \infty) \rightarrow \mathbb{R}$ ,  $\phi(x) = \frac{1}{m}x^2$  is non-decreasing and convex. Therefore, it suffices to show that  $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|$  is convex. This is seen by invoking the following simple lemma.

**Lemma 3.** *Suppose  $\phi : [0, \infty) \rightarrow \mathbb{R}$  is non-decreasing and convex and  $f : \mathbb{R}^n \rightarrow [0, \infty)$  is convex. For  $p \in [0, 1]$  and  $\mathbf{r}, \mathbf{s} \in \mathbb{R}^n$ , we have*

$$\begin{aligned} \phi(f(p\mathbf{r} + (1-p)\mathbf{s})) &\leq \phi(f(p\mathbf{r}) + (1-p)f(\mathbf{s})) \\ &\leq p\phi(f(\mathbf{r})) + (1-p)\phi(f(\mathbf{s})). \end{aligned}$$

Hence,  $\phi \circ f$  is convex.

Let  $\mathbf{w}$  and  $\tilde{\mathbf{w}}$  be two weight vectors. We have

$$\begin{aligned} f(p\mathbf{w} + (1-p)\tilde{\mathbf{w}}) &= \|\mathbf{X}(p\mathbf{w} + (1-p)\tilde{\mathbf{w}}) - \mathbf{y}\| \\ &\leq \|p(\mathbf{X}\mathbf{w} - \mathbf{y}) + (1-p)(\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y})\| \\ &\leq p\|\mathbf{X}\mathbf{w} - \mathbf{y}\| + (1-p)\|\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y}\| \\ &= pf(\mathbf{w}) + (1-p)f(\tilde{\mathbf{w}}). \end{aligned}$$

In the above derivation, we have used the triangle inequality  $\|\mathbf{r} + \mathbf{s}\| \leq \|\mathbf{r}\| + \|\mathbf{s}\|$  and  $\|\lambda\mathbf{r}\| = |\lambda|\|\mathbf{r}\|$ , which hold for arbitrary  $\lambda \in \mathbb{R}$  and  $\mathbf{r}, \mathbf{s} \in \mathbb{R}^n$ .

We see that the mean-squared-error is a convex function by combining all results.

## References

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