

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, where

$$\mathbf{x}^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{pmatrix} \in \mathbb{R}^n$$

is the i th the feature vector of the i th training example and $y^{(i)} \in \mathbb{R}$ is label.

Let $\mathbf{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$ be the weight vector and the $b \in \mathbb{R}$ bias of the linear regression model. Let

$$\hat{y}^{(i)} = \sum_{j=1}^n w_j x_j^{(i)} + b.$$

denote its prediction when given the i th feature vector $\mathbf{x}^{(i)}$. Note that $\hat{y}^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + b$. The mean squared error (MSE) on the training set is equal to

$$\text{MSE}(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2.$$

To simplify the presentation, we assume that the bias b is set to 0, that is, only the weight vector \mathbf{w} has to be determined. We will see later that this is not a restriction.

We now introduce additional notation to express the MSE in a linear-algebraic way. Define the vectors

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

Observe that the MSE is equal to

$$\text{MSE}(\mathbf{w}) = \frac{1}{m}(\hat{\mathbf{y}} - \mathbf{y})^T(\hat{\mathbf{y}} - \mathbf{y}) = \frac{1}{m}\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2,$$

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

Define the so-called design matrix by

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

Observe that

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

due to the symmetry of the inner (dot) product $\hat{y}^{(i)} = \mathbf{x}^{(i)T}\mathbf{w} = \mathbf{w}^T\mathbf{x}^{(i)}$.

2 Normal equation

Recall that the bias of the linear regression model is assumed 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 on gradients

We introduce some abbreviations. Let $[n] = \{1, \dots, n\}$. Let ∂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

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

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 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

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

$$\nabla_{\mathbf{w}} \text{MSE}(\mathbf{w}) = \mathbf{0} \quad (1)$$

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

$$\Rightarrow \frac{1}{m} \nabla_{\mathbf{w}} (\hat{\mathbf{y}} - \mathbf{y})^T (\hat{\mathbf{y}} - \mathbf{y}) = \mathbf{0} \quad (3)$$

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

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

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

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

In eq. (5), we 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. We also use that the term $\mathbf{y}^T \mathbf{y}$ does not depend on \mathbf{w} . The solution given by eq. 7 is known as the normal equation.

5 Proof of convexity of MSE

Theorem 2. *The MSE*

$$\text{MSE}(\mathbf{w}) = \frac{1}{m} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

is convex.

To prove this theorem, we will rely on the following two simple lemmata.

Lemma 3. *Suppose $\phi : [0, \infty) \rightarrow \mathbb{R}$ is non-decreasing and convex and $f : \mathbb{R}^n \rightarrow [0, \infty)$ is convex. Then, $\phi \circ f$ is convex.*

Proof. For $p \in [0, 1]$ and $\mathbf{r}, \mathbf{s} \in \mathbb{R}^n$, we have

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

□

Lemma 4. *The function $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2$ is convex.*

Proof. Let $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^n$ be two arbitrary weight vectors. We have

$$\begin{aligned} f(p\mathbf{w} + (1-p)\tilde{\mathbf{w}}) &= \|\mathbf{X}(p\mathbf{w} + (1-p)\tilde{\mathbf{w}}) - \mathbf{y}\|_2 \\ &\leq \|p(\mathbf{X}\mathbf{w} - \mathbf{y}) + (1-p)(\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y})\|_2 \\ &\leq p\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2 + (1-p)\|\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y}\|_2 \\ &= pf(\mathbf{w}) + (1-p)f(\tilde{\mathbf{w}}). \end{aligned}$$

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 $\mathbf{r}, \mathbf{s} \in \mathbb{R}^n$. \square

Proof. We can write

$$\text{MSE}(\mathbf{w}) = \phi(\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2),$$

where $\phi : [0, \infty) \rightarrow \mathbb{R}, \phi(x) = \frac{1}{m}x^2$. It is important that ϕ is non-decreasing and convex. The theorem follows now by applying the above two lemmata. \square

6 General case of linear regression with non-zero bias

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

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

References

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