Derivation of normal equation and some additional results for linear regression

Pawel Wocjan

February 6, 2019

Abstract

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

1 Notation

Let $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)}) \in \mathbb{R}^n \times \mathbb{R}$ denote the collection of training examples, where

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

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

Let

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \in \mathbb{R}^n$$

be the weight vector and $b \in \mathbb{R}$ the bias of the linear regression model. It predicts the value

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

when given the ith feature vector $\boldsymbol{x}^{(i)}$. Note that $\hat{y}^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + b$, that is, the inner (dot) product of the weight vector and feature vector plus the bias.

The mean squared error (MSE) on the training set is equal to

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

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

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

Observe that the MSE is equal to

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

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

Define the so-called design matrix by

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

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

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

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

2 Normal equation

Recall that the bias of the linear regression model is assumed to 0, that is, only the weight vector $\boldsymbol{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$ needs to be determined.

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

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

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 $[n] = \{1, \ldots, n\}$. Let ∂w_r denote the partial derivative operator

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

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

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

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

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

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

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

We have

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

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 $\boldsymbol{A}$ is symmetric, that is, $a_{rs} = a_{sr}$.

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

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

Proof. This is easy. Prove it yourself.

4 Proof of normal equation

To minimize the MSE, we compute its gradient and determine where it is equal to $\mathbf{0}$:

$$\begin{align}
\nabla_{\boldsymbol{w}}\mathrm{MSE}(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}}\frac{1}{m}\|\hat{\boldsymbol{y}}-\boldsymbol{y}\|_2^2 \tag{1}\\
&= \frac{1}{m}\nabla_{\boldsymbol{w}}(\hat{\boldsymbol{y}}-\boldsymbol{y})^T(\hat{\boldsymbol{y}}-\boldsymbol{y}) \tag{2}\\
&= \frac{1}{m}\nabla_{\boldsymbol{w}}(\boldsymbol{X}\boldsymbol{w}-\hat{\boldsymbol{y}})^T(\boldsymbol{X}\boldsymbol{w}-\hat{\boldsymbol{y}}) \tag{3}\\
&= \frac{1}{m}\nabla_{\boldsymbol{w}}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{y}^T\boldsymbol{y}) \tag{4}\\
&= \frac{2}{m}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}-\boldsymbol{X}^T\boldsymbol{y}) \tag{5}\\
&\implies \boldsymbol{w}=(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{6}
\end{align}$$

In eq. (4), we use Lemma 1 with $\boldsymbol{A}=\boldsymbol{X}^T\boldsymbol{X}$ and Lemma 2 with $\boldsymbol{v}=\boldsymbol{X}^T\boldsymbol{y}$ to compute the gradient. We also use that the term $\boldsymbol{y}^T\boldsymbol{y}$ does not depend on $\boldsymbol{w}$. The solution given by eq. 6 is known as the normal equation.

5 Proof of convexity of MSE

Theorem 2. *The MSE*

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

is convex.

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

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

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

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

□

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

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

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

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

Proof. We can write

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

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

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 $\boldsymbol{w}_b = (b, w_1, \ldots, w_n)^T \in \mathbb{R}^{n+1}$ and the augmented feature vectors $\boldsymbol{x}_b^{(i)} = (1, x_1^{(i)}, \ldots, x_n^{(i)})^T \in \mathbb{R}^{n+1}$. We have

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

References

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