A Linear Algebra Interpretation of Linear Regression

Ajay Raj and Patrick Chao

October 23rd, 2018

Contents

1 Notation

Common notation:

- **A**: Bold capital letters represent matrices
- **x**: Bold lowercase letters represent vectors
- θ: Non-bold values represent scalars

2 The 1-Dimensional Case

The simplest case of linear regression is given points $(x_1, y_1), ..., (x_n, y_n)$, find a line of best fit through the origin to describe the points. For a motivating example, say that x is the temperature on any given day and y is the total ice cream sales you made for that day. Let's define some terms.

We are trying to find a model $f_\theta(x)$ such that $f_\theta(x)$ will predict, given the temperature on a day, the ice cream sales you will get. Specifically, we will utilize a simple linear model with one parameter:

$$f_\theta(x) = \theta x, \ \forall \theta \in \mathbb{R}.$$

Now, let's describe the loss of this model. Last week, you learned that we should use **RSS (residual sum of squares)** loss for this model, which is, for this model:

$$\sum_{i=1}^{n} (y_i - \theta x_i)^2. \tag{1}$$

2.1 Vector Reformulation

Now, let's express this using linear algebra. Instead of expressing that we have a set of points, we have two **vectors**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Since θ is a **scalar** (just a number, not a vector), we can still say that our model is $f_\theta(x)$, but now it looks like this:

$$f_\theta(x_1, x_2, ..., x_n) = \begin{bmatrix} \theta x_1 \\ \theta x_2 \\ ... \\ \theta x_n \end{bmatrix}$$

Keep in mind that multiplying a vector by a scalar is equivalent to multiplying each entry of the vector by the scalar.

Now, you can see that each element in the vector is the **predicted** value of x_i.

2.2 Vector Norms

Now, we need to express the loss of this model. To do this, we need to look at two fundamental concepts of linear algebra. In linear algebra, there exists a concept of a **norm** of a vector, which expresses various notions of **magnitude**. There are two **norms** that we will be discussing today.

$$\ell_1(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\ell_2(\mathbf{x}) = \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

The bars represent the norm, and the subscript represents the type of norm. The ℓ_2 norm is what you may be familiar with, it describes the length of the vector in the Euclidean sense. This is equivalent to the distance formula representation.

Can you think of an expression with the norm that expresses the loss of our model?

Next, let's discuss the notion of a **standard inner product**, which is a function that turns two vectors into a number. In general, an inner product can be thought as the similarity between two vectors. The definition of an inner product is

$$\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \cos(\theta).$$

where θ is the angle between the vectors. A more concise method of writing this with vectors in the Euclidean case is

$$\begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \mathbf{u}^T \mathbf{v}.$$

2.3 Transposes

The notation v^T represents the transpose of a vector or matrix. In general, this is defined as

$$\mathbf{A}_{ij}^T = \mathbf{A}_{ji}.$$

The transpose operation means the elements in the ith row and jth column become the elements in the jth row and ith column, or all elements are reflected over the main diagonal (top left to bottom right). In the case of a matrix $\mathbf{A}$

that is $n \times p$, n rows and p columns, $\mathbf{A}^T$ is of dimension $p \times n$. The transpose of a vector $\mathbf{x}$ of dimension $n \times 1$ becomes a row vector of dimension $1 \times n$. Since we may consider transposing matrices/vectors as reflecting them over the diagonal,

$$(\mathbf{A}^T)^T = \mathbf{A}.$$

A key identity of transposes is how they affect matrix-matrix products and matrix-vector products.

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \tag{2}$$

Notice that $\mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u}$. A very common technique is rewriting norms using transposes.

$$\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2 \tag{3}$$

We say that a vector $\mathbf{x}$ has **unit norm** if $\|\mathbf{x}\|_2 = 1$. This may be interpreted as the length of the vector being 1. We can turn any vector $\mathbf{x}$ into one that is unit norm by dividing by its norm:

$$\begin{aligned}\mathbf{v} &= \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \\ \mathbf{v}^T\mathbf{v} &= \left(\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right)^T \frac{\mathbf{x}}{\|\mathbf{x}\|_2} = \frac{\mathbf{x}^T\mathbf{x}}{\|\mathbf{x}\|_2^2} = 1\end{aligned}$$

2.4 Vectorized Loss

We are now equipped to solve for θ. We would like to rephrase the RSS in equation 1. Our loss is:

$$\mathbf{Loss}(\theta) = \sum_{i=1}^{n}(y_i - \theta x_i)^2 = \|\mathbf{y} - \theta\mathbf{x}\|_2^2.$$

Let's express the loss in terms of **inner products**. First we may start off with the definition of loss.

$$\mathbf{Loss}(\theta) = \|\mathbf{y} - \theta\mathbf{x}\|_2^2$$

Next, we may represent the ℓ_2 norm squared with equation 3.

$$\begin{aligned}\mathbf{Loss}(\theta) &= (\mathbf{y} - \theta\mathbf{x})^T(\mathbf{y} - \theta\mathbf{x}) \\ &= \mathbf{y}^T\mathbf{y} - \theta\mathbf{y}^T\mathbf{x} - \theta\mathbf{x}^T\mathbf{y} + \theta^2\mathbf{x}^T\mathbf{x} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{x}\theta + \mathbf{x}^T\mathbf{x}\theta^2. \end{aligned}\tag{4}$$

There is an important step between the second and third line. Since the loss is a scalar, we may consider it as a 1×1 matrix. Thus the transpose of a 1×1 matrix is itself, or more concretely $\mathbf{A} = \mathbf{A}^T$. In the case of $\theta\mathbf{y}^T\mathbf{x}$, we know this is equivalent to $(\theta\mathbf{y}^T\mathbf{x})^T$. Now applying the rule of switching order with transposition in equation 2,

$$(\theta\mathbf{y}^T\mathbf{x})^T = \theta(\mathbf{y}^T\mathbf{x})^T = \theta(\mathbf{x}^T(\mathbf{y}^T)^T) = \theta\mathbf{x}^T\mathbf{y}.$$

From the first to the second equation, we can pull out θ since it is a scalar.

2.5 Minimizing Loss in the 1-Dimensional Case

It is our goal, now, to minimize that loss. You may have remembered from calculus that in order to minimize a function, you take the derivative and set it equal to zero. Let's take the derivative with respect to θ. Notice that this is the same as deriving an equation that looks like $a\theta^2 + b\theta + c$, since we may consider the vector terms as constants.

We will differentiate the expanded form of the loss in equation 4.

$$\frac{\partial}{\partial \theta}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{x}\theta + \mathbf{x}^T\mathbf{x}\theta^2\right) = -2\mathbf{y}^T\mathbf{x} + 2\mathbf{x}^T\mathbf{x}\theta$$

Now, we set it equal to zero to solve for the extrema.

$$\begin{aligned} 0 &= -2\mathbf{y}^T\mathbf{x} + 2\mathbf{x}^T\mathbf{x}\theta \\ 2\mathbf{y}^T\mathbf{x} &= 2\mathbf{x}^T\mathbf{x}\theta \\ \mathbf{y}^T\mathbf{x} &= \mathbf{x}^T\mathbf{x}\theta \\ \theta &= \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} \end{aligned}$$

This is equivalent to the formula given last week!

$$\boxed{\theta = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}}$$

3 The p-Dimensional Case

Now say that instead of one feature, we have p features (note that this is usually the case—for example, ice cream sales don't just depend on temperature). We have a model $f_{\boldsymbol{\theta}}(x_1, x_2, ..., x_p)$. We would still like to model a single scalar output y_i. We continue our assumption that our model is linear, meaning we know that $f_{\boldsymbol{\theta}}$ looks like this:

$$f_{\boldsymbol{\theta}}(x_1, x_2, ..., x_p) = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

Because mathematicians are efficient (not lazy), we can write f_θ as an inner product, where θ is now a **vector** and x is now a p-dimensional vector:

$$f_{\boldsymbol{\theta}}(x) = \mathbf{x}^T \boldsymbol{\theta}$$

Notice that now we are attempting to calculate p values in θ instead of a single value. Now, let's express our loss. In general, our loss is

$$\textbf{Loss}(\theta) = \sum_{i=1}^{n} (y_i - f_\theta(\mathbf{x}_i))^2.$$

Previously, we were considering a single scalar input x to our model (temperature for the day). However, now we now have a vector $\mathbf{x}$ corresponding to the set of p features. We can consider each observation $\mathbf{x}_i$ as a vector of length p. With our current model, it looks like this:

$$\textbf{Loss}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}^T \boldsymbol{\theta})^2$$

3.1 Matrix-ized Loss

Let's try to express this without a sum. Earlier, we were able to construct a vector such that each element is the individual loss for that specific data point. We were able to stack each individual temperature into a vector corresponding to all the temperatures. In this case, we may stack each row into a matrix.

Let $\mathbf{X}$ be a matrix where each row is a single data point (with p features). That is, $\mathbf{X}$ is an $n \times p$ matrix. The vector specified above is $\mathbf{y} - \mathbf{X}\boldsymbol{\theta}$. Notice that $\mathbf{y}$ is the same since the dimension of our output (ice cream sales) does not change. Our loss is now:

$$\textbf{Loss}(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \tag{5}$$

As we did earlier, let's express this loss using inner products.

$$\begin{aligned}
\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 &= (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta} - (\mathbf{X}\boldsymbol{\theta})^T\mathbf{y} + (\mathbf{X}\boldsymbol{\theta})^T(\mathbf{X}\boldsymbol{\theta}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\theta}^T\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}.
\end{aligned}$$

We use the same trick of merging the middle terms. thus our final equation is

$$\mathbf{Loss}(\boldsymbol{\theta}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}. \tag{6}$$

3.2 Matrix Calculus Identities

How do we take the derivative of this? Since $\boldsymbol{\theta}$ is a vector now, it becomes a little more difficult. We won't go into why these properties are true today, but these tools help use derive this loss function. In vector calculus, we use $\nabla_{\mathbf{X}}$ instead of $\frac{d}{dx}$. The idea of a gradient is that you take the partial derivative with respect to each coordinate individually.

$$\nabla_{\mathbf{X}} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix}$$

There are some essential identities that you can perform with gradient operations.

$$\nabla_{\mathbf{X}}(\mathbf{a}^T\mathbf{X}) = \mathbf{a} \tag{7}$$

$$\nabla_{\mathbf{X}}(\mathbf{A}\mathbf{X}) = \mathbf{A} \tag{8}$$

$$\nabla_{\mathbf{X}}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = (\mathbf{A} + \mathbf{A}^T) \tag{9}$$

Notice that if $\mathbf{A}$ is symmetric, meaning $\mathbf{A} = \mathbf{A}^T$, then equation 9 becomes $\nabla_{\mathbf{X}}\mathbf{X}^T\mathbf{A}\mathbf{X} = 2\text{A}$.

We recommend that you go attempt to prove the first two identities yourself.

3.3 Minimizing Loss in the p-Dimensional Case

Now equipped with these identities, we may solve for the optimal value of $\boldsymbol{\theta}$.

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\right) &= \nabla_{\boldsymbol{\theta}}\left(\mathbf{y}^T\mathbf{y}\right) - \nabla_{\boldsymbol{\theta}}\left(2\mathbf{y}^T\mathbf{X}\boldsymbol{\theta}\right) + \nabla_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\right) \\ &= 0 - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \end{aligned}$$

We may now set this equal to 0 and solve for $\boldsymbol{\theta}$.

$$\begin{aligned} 0 &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \\ 2\mathbf{X}^T\mathbf{y} &= 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \\ \mathbf{X}^T\mathbf{y} &= \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \\ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \end{aligned}$$

$$\boxed{\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

The final equation is incredibly important, it is known as the **normal equation**. This boils down the entire linear regression problem into a simple and one line equation.

4 Closing Remarks

We have covered the 1-dimensional case intuition and built up to the general p-dimensional case. There are a few important key points to consider.

- Assuming the data has mean 0, what statistical interpretation do you have for the formula for the 1-dimensional case?
- Is $(\mathbf{X}^T\mathbf{X})^{-1}$ always invertible?
- Does this construction of linear regression contain constant terms? For example, in the 1-dimensional case, do we predict models of the form $\theta x + b$ where b is a constant? If not, how can we adjust our construction? Can this be expanded to the p-dimensional case?
- How does this change if we adjust our loss function to also include an ℓ_2 penalty to the magnitude of $\boldsymbol{\theta}$? Consider the following situation for $\lambda \geq 0$.

$$\mathbf{Loss}(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2.$$