LINEAR REGRESSION AND LEAST SQUARES

AVNEESH SALUJA

Abstract. The linear regression model is one of the simplest and most widely used tools in statistics and machine learning. Put simply, the linear regression algorithm computes an output value y based on a weighted combination of the input features ($n+1$ features for $X \in \mathbb{R}^n$, since we include the offset/bias). Assuming a least-mean squares criterion, we can derive a closed-form solution for the parameter weights through a variety of techniques. This "tutorial" provides two derivations for the parameter weights and aims to develop an intuition of the linear regression problem and the multiple ways to view the solution.

1. Background

Let us first define the learning environment. Suppose we have a series of training examples $\{x_i, y_i\}_1^m$, where m is the size of our training set, $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$.

The task that we are faced with is to design a learner that, based on the training examples we have provided along with the correct values, can predict values of y given future values of x. We can express the model for this task in a number of different ways, but the focus of this document is the linear regression model.

In the linear regression model, we can express our hypothesis $h_\theta(x)$ as a function of the input x as well as a set of parameter values θ (we choose this notation to emphasize the dependence of our hypothesis on the parameter values). This, we can express our model in the following form:

$$h_\theta(x_i) = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_n x_{in}$$

which can be written concisely as:

$$h_\theta(x_i) = \sum_{j=0}^{n} \theta_j x_{ij} = \theta^T \cdot x_i$$

The question now is how do we estimate the parameter weights θ. One way to accomplish this goal is, in the training step, to minimize the squared difference between our hypothesis $h_\theta(x_i)$ and the true value of the training example, y_i. Thus, the optimal parameters θ can be expressed as:

Date: April 6, 2011.

$$\theta = \arg\min_{\theta} \sum_{i=0}^{m} (h_\theta(x_i) - y_i)^2 \tag{1}$$

This minimization criterion is the least squares criterion and it is this criterion that we will use in subsequent derivations. In Section 2, we first show the link between the least squares minimization criterion and maximum likelihood estimation (MLE). Sections 3 and 4 go over the two main ways one can minimize this objective function, and Section 5 provides a linear algebra-motivated intuition to the problem.

2. Least Mean Squares & MLE

In this section, we derive the maximum likelihood estimates of the parameters θ (we need to make some assumptions so that we can get a closed form solution for these estimates), and show that the form is equivalent to the least-squares minimization problem.

We must first make the assumption that the difference between the true value of the training example and our hypothesis $h_\theta(x_i)$, which we can call ϵ_i, or the error, is normally distributed: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Thus, we can write:

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\epsilon_i^2}{2\sigma^2}\right)$$

We must now express $p(y_i|x_i;\theta)$ in closed form. Note that the relationship is $y_i = \theta^T x_i + \epsilon_i$, and so we can express $p(y_i|x_i;\theta)$ as a Gaussian $\mathcal{N}(\theta^T x_i - y_i, \sigma^2)$.

In the following steps, we look to maximize the likelihood, which we first express as a log likelihood. Explicitly maximizing this log likelihood will get us the answer.

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{m} p(y_i \mid x_i;\theta) \rightarrow \\
LL(\theta) &= \sum_{i=1}^{m} \log p(y_i \mid x_i;\theta) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\theta^T x_i - y_i)^2}{2\sigma^2}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{m} (\theta^T x_i - y_i)^2 \qquad (2)
\end{aligned}$$

Equation 2 gives the log-likelihood as a function of θ. Taking the derivative of this expression with respect to θ and setting equal to 0, one finds that the first term has no dependence on θ, and maximizing the second term is equivalent to minimizing the negative of that term, which is simply Equation 1, along with some constants that do not change the optimization problem.

Thus, we have shown that if the training error per sample is normally distributed with mean 0 and variance σ^2, then the maximum likelihood estimate for θ is equivalent to the least squares minimization value for θ.

3. Derivation with Matrix Calculus I

Instead of just looking at one training example, let us formulate a matrix X, which is an $m \times n+1$ ($n+1$ instead of n because of the intercept term, which we set to 1) matrix, where m is the size of our training set, and each training example $x_i \in \mathbb{R}^{n+1}$. Note that each row of the matrix X is a transpose of each training example x_i. Thus, we can write $H_\theta = X\theta$ where H_θ is an $m \times 1$ matrix with the hypotheses for each training example.

We can formulate 1 in the following manner:

$$\sum_{i=0}^{m}(h_\theta(x_i) - y_i)^2 = \|X\theta - Y\|_2^2 \tag{3}$$

$$= (X\theta - Y)^T(X\theta - Y) \tag{4}$$

$$= \theta^T X^T X\theta - 2Y^T X\theta + Y^T Y \tag{5}$$

Lines 3 and 4 make use of the equality $\sum_i z_i^2 = \|z\|_2^2 = z^T z$ for vectors, and $(X\theta - Y)$ is an $m \times 1$ vector. Line 5 expands and multiplies the expression using transpose properties (note that $\theta^T X^T Y = Y^T X\theta$ due to the transpose properties, so we can collapse these terms together).

We need to take the gradient and set it to zero:

$$\arg\min_\theta \theta^T X^T X\theta - 2Y^T X\theta + Y^T Y = \nabla_\theta\ \theta^T X^T X\theta - 2Y^T X\theta + Y^T Y$$

$$= 2X^T X\theta - 2X^T Y \tag{6}$$

where Line 6 makes use of the following standard matrix derivatives:

$$\nabla_\theta\ \theta^T X\theta = 2X\theta \text{ and } \nabla_\theta\ Y^T\theta = Y$$

If we set this last expression to zero and solve for θ, we get:

$$\theta = (X^T X)^{-1} X^T Y \tag{7}$$

Equation 7 is known as the normal equations, and provides a closed-form solution to θ estimation.

4. Derivation with Matrix Calculus II

A second way of deriving the normal equations is similar to Section 3, except instead of using properties of matrix derivatives, we use properties of the trace derivative.

First, we alter the objective function slightly to add a coefficient of $\frac{1}{2}$, so we express Equation 1 with a factor of $\frac{1}{2}$. Then, we note that Equation 5 is actually a scalar, and so

we can write $\nabla_\theta\ \theta^T X^T X\theta - 2Y^T X\theta + Y^T Y = \nabla_\theta\ \mathrm{tr}\left(\theta^T X^T X\theta - 2Y^T X\theta + Y^T Y\right)$ without loss of generality. So, let us rederive the normal equations from Equation 5:

$$\begin{aligned}
\frac{1}{2}\nabla_\theta\ \theta^T X^T X\theta - 2Y^T X\theta + Y^T Y &= \frac{1}{2}\nabla_\theta\ \mathrm{tr}\left(\theta^T X^T X\theta - 2Y^T X\theta + Y^T Y\right) \\
&= \frac{1}{2}\nabla_\theta\ \left(\mathrm{tr}\theta^T X^T X\theta - 2\mathrm{tr}Y^T X\theta\right) \\
&= \frac{1}{2}\left(X^T X\theta + X^T X\theta - X^T Y\right) \qquad (8) \\
&= X^T X\theta - X^T Y
\end{aligned}$$

Setting equal to zero and solving, we get Equation 7. Line 8 is due to the properties of the trace derivatives, namely:

$$\nabla_A \mathrm{tr}AB = B^T \qquad (9)$$

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T \qquad (10)$$

$$\nabla_A \mathrm{tr}ABA^T C = CAB + C^T AB^T \qquad (11)$$

The first property is used in the first step to collapse the two terms together as in Equation 5. The second property is used in conjunction with the third, as follows:

Let $f(A) = \mathrm{tr}ABA^T C$. So, as per Equation 10,

$$\begin{aligned}
\nabla_{A^T} f(A) &= (\nabla_A f(A))^T \rightarrow \\
\nabla_{A^T} \mathrm{tr}ABA^T C &= (\nabla_A \mathrm{tr}ABA^T C)^T \\
&= (CAB + C^T AB^T)^T \\
&= B^T A^T C^T + BA^T C \qquad (12)
\end{aligned}$$

Let us show that these three properties are true.

4.1. **Deriving Equation 9.** This derivation is fairly straightforward:

$$\begin{aligned}
\nabla_A \mathrm{tr}AB &= \nabla_A \sum_{i=1}^{m} (AB)_{ii} \qquad (13) \\
&= \nabla_{a_{ij}} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} B_{ji} \right) \qquad (14) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \nabla_{a_{ij}} A_{ij} B_{ji} \qquad (15) \\
&= B_{ji}
\end{aligned}$$

where line 13 is from the definition of the trace, line 14 is from the definition of matrix multiplication, and line 15 is because of the linearity of the gradient operator. Therefore, $\nabla_A \mathrm{tr}AB = B^T$.

4.2. **Deriving Equation 10.** For this derivation, it is easier to visualize the matrix directly:

$$\begin{aligned}\nabla_{A^T} f(A) &= \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{21}} & \cdots & \frac{\partial f(A)}{\partial A_{n1}} \\ \frac{\partial f(A)}{\partial A_{12}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{n2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{1n}} & \frac{\partial f(A)}{\partial A_{2n}} & \cdots & \frac{\partial f(A)}{\partial A_{nn}} \end{bmatrix} \\ &= (\nabla_A f(A))^T\end{aligned}$$

4.3. **Deriving Equation 11.** This derivation is fairly tricky. First, let us say $f(A) = AB$. Then:

$$\begin{aligned}\nabla_A \mathrm{tr} ABA^TC &= \nabla_A \mathrm{tr} f(A)A^TC \\ &= \nabla_A \mathrm{tr} f(A)A^TC + \nabla_A \mathrm{tr} f(A)A^TC \quad (16) \\ &= (A^TC)^T f'(A) + (\nabla_{A^T} \mathrm{tr} f(A)A^TC)^T \quad (17) \\ &= C^TAB^T + ((Cf(A))^T)^T \quad (18) \\ &= C^TAB^T + CAB\end{aligned}$$

Line 16 breaks down the derivative into the product rule, and line 17 follows from the property in line 9 for the first term, and line 10 for the second term. Line 18 is the result of the evaluation of the simplified forms from line 17 (note that $f'(A) = \nabla_A AB = B^T$, similar operations are performed on the second term). Thus, using these trace properties, we can show the least squares derivation.

4.4. **Tying it all together.** Using Equation 12, we can proceed by substituting $A^T = \theta, B = X^TX = B^T, C = I$ into equation 12. From there, we get equation 8.

5. Linear Algebra Interpretation

Given the linear principles underlying the linear least squares problem, a natural place to look is to linear algebra and results from that field for intuition.

Recall that in Section 3 we defined series of data points in the form of a matrix H_θ. We can think about the least squares problem as finding a point in a linear subspace closest to the point y_i. The linear subspace can be expressed in terms of a matrix θ operating on an $m \times n + 1$ matrix X, where $x_i \in \mathbb{R}^{n+1}$, and thus the goal is to solve the problem $Y = X\theta$. The problem is that the points y_i may not necessarily lie in the column space of X, i.e. it's not in the linear subspace. So, the aim is to find a matrix θ such that given an input x_i, we can find the projection of y_i, i.e. the point in the linear subspace spanned by columns of the matrix X, i.e. the basis vectors for the subspace, closest to the point y_i in some sense.

If we define the linear projection from y_i to the column space of X as $|X\theta - Y|$, which is equivalent to the error term, then we can show that the normal equations we came up

with before satisfy some interesting properties, for example the fact that $X\theta - Y$ is in the null space of X:

$$\begin{aligned} X^T(X\theta - Y) &= X^T X\theta - X^T Y \\ &= X^T X(X^T X)^{-1} X^T Y - X^T Y \quad (19) \\ &= X^T Y - X^T Y \\ &= 0 \end{aligned}$$

where equation 19 is simply substituting from the normal equations. This shows that the error term lies in the null space of the subspace spanned by the data, and this also means that the error term $X\theta - Y$ is orthogonal to the column space of the data X. Therefore, the linear projection $X\theta - Y$ is orthogonal to the subspace spanned by X, meaning that the normal equations we derived in Sections 3 and 4 are optimal, since an orthogonal projection is the shortest distance between the point y_i and the subspace. This is known as the orthogonality principle.

Thus, one can say that linear algebra provides a good intuition as to why the normal equations are optimal.