Linear regression

Patrick Huang

April 2025

1 Introduction

Linear regression is a simple method of fitting a linear model to data. It is simple enough that there exist closed form solutions on a given dataset.

In this paper, we present two methods of deriving the closed form expression: One using principles from machine learning, and the other from abstract linear algebra.

2 Problem statement

We are given a set of N data points $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$, $x_i, y_i \in \mathbb{R}$.

We have a model $\hat{y}_i = mx_i + b$, where $m, b \in \mathbb{R}$ are parameters.

The loss of the model is $L = \Sigma(\hat{y}_i - y_i)^2$, which is mean squared error or L2 loss.

The optimal model (i.e. choice of m, b) is that which minimizes L.

3 Applied method

In this method, we use principles from machine learning to solve for the ideal model as a closed form expression.

3.1 Rewrite as matrix equation

First, we rewrite the expressions slightly to turn the loss into a single matrix expression.

Let $u = (1, 1, ..., 1)$ be a vector of ones with length N. Then, our model is equivalent to $\hat{y} = mx + bu$.

Let the matrix $A = \begin{bmatrix} x & u \end{bmatrix} = \begin{bmatrix} x_1 & u_1 \\ x_2 & u_2 \\ ... & ... \end{bmatrix}$, i.e. the concatenation of the column vectors x and u. This is the feature matrix (where the "features" corresponding to the bias are one).

Let $v = \begin{bmatrix} m \\ b \end{bmatrix}$ be the column vector of the model's parameters.

We can rewrite our model as $\hat{y} = Av$. Thus, our loss becomes $L = (\hat{y} - y)^2 = \|Av - y\|^2$. Notice how we omit the summation, because all the indices are combined into one vector now.

In summary, we have 2 given vectors x, y (the data points), and 1 constant vector u (the vector of ones). The model still has two scalar parameters m, b. m multiplies x, and b multiplies u. We write the loss over the whole dataset as a single matrix equation.

3.2 Finding the optimal parameters

The loss function is a positive quadratic equation, which is differentiable, has no points of inflection, and has a single minimum. Therefore, we can take the derivative of the loss with respect to the parameters to obtain the optimal model.

$\frac{dL}{dv} = (\frac{dL}{dv_1}, \frac{dL}{dv_2}, ..., \frac{dL}{dv_n})$.

Because we are differentiating a scalar with respect to a vector, the result is the derivative of the scalar with respect to each component of the vector. Remember that $v = (m, b)$ will have two components.

Here, we will not separate the expressions component-wise, and instead use some results from matrix calculus. See appendix for a more detailed derivation.

First, we expand the square in the loss. Then, we take the derivative of each term.

$L = \|Av - y\|^2 = (Av - y)^\intercal(Av - y) = (v^\intercal A^\intercal - y^\intercal)(Av - y) = v^\intercal A^\intercal Av - v^\intercal A^\intercal y - y^\intercal Av + y^\intercal y$

$\frac{dL}{dv} = 2A^\intercal Av - A^\intercal y - y^\intercal A = 2A^\intercal Av - 2A^\intercal y = 0$

We set this to zero to find the global minimum. Solving for v, we obtain

$v = (A^\intercal A)^{-1}(A^\intercal y)$

3.3 More parameters

This method works with more than one input feature. Suppose each data sample has two features and a ground truth: x_i, w_i, and y_i.

Our linear model is $\hat{y}_i = mx_i + nw_i + b$.

To solve for the minimum, we still add a vector of ones u. The feature matrix

$A = \begin{bmatrix} x & w & u \end{bmatrix} = \begin{bmatrix} x_1 & w_1 & u_1 \\ x_2 & w_2 & u_2 \\ ... & ... & ... \end{bmatrix}$.

The vector of parameters $v = \begin{bmatrix} m & n & b \end{bmatrix}^\intercal$.

The closed form expression is still the same, as above.

In general, if we have N data points with k input features, our model will have $k + 1$ parameters (one for each feature, and a bias). The feature matrix will be shape $(N, k + 1)$, and the parameter vector will have $k + 1$ elements.

4 Abstract method

In this method, we use vector spaces and theorems from abstract linear algebra to find the ideal model.

4.1 Rewrite as vector spaces

Let $V = \mathbb{R}^N$. Similar to section 3.1, we rewrite the model as a vector equation in V.

Again, we define a vector of ones $u = (1, 1, ..., 1)$ of length N. Thus, $\hat{y} = mx + bu$. The vectors $\hat{y}, y, x, u \in V$.

Notice how $\hat{y}$ is a linear combination of the vectors x and u. Let $\beta = \{x, u\}$, and the space spanned $W = \text{span}(\beta)$, a two (possibly one) dimensional subspace. Therefore, $\hat{y} \in W$.

This means that a given model's predictions $\hat{y}$ is a single vector in W, and W is the space of *predictions of all possible models* given x.

4.2 Orthogonal decomposition

Consider y, the ground truth. If $y \in W$, then it is possible for the model $\hat{y}$ to fit the data exactly. Otherwise, we want to find the model that minimizes the loss defined in section 2.

Notice how the loss, $\Sigma(\hat{y}_i - y_i)^2$, is equivalent to $\|\hat{y} - y\|^2$, like in section 3.1 (for those pedantic, this is the standard L2 norm on $\mathbb{R}^N$). Therefore, we are trying to find the prediction vector $\hat{y} \in W$ "closest" to the ground truth y under the L2 norm.

By a theorem, a vector $v \in V$ can be uniquely expressed as $w + z$, where $w \in W$ and $z \in W^\perp$. $W^\perp$ is the orthogonal complement of W, defined as $\{p : p \perp w \quad \forall w \in W\}$. $a \perp b$ iff $a \cdot b = 0$.

Additionally, the resulting w is the vector in W "closest" to v; that is, $\|w - v\|^2 \leq \|a - v\|^2 \quad \forall a \in W$. See appendix for a proof.

Using this theorem, we decompose y as a sum $\hat{y} + z$. According to the theorem, the expression $\|\hat{y} - y\|^2$, which is also the loss, is minimized. This implies that $\hat{y}$ under this decomposition is the best model. z is the difference between the truth and the model's predictions, which we don't need.

4.3 Finding the optimal parameters

First, we need an orthonormal basis for W. Normalize u: $u_u = u/\sqrt{N}$. Orthogonalize x with respect to u_u: $x_\perp = x - (x \cdot u_u)u_u = x - \frac{u}{N}\sum x_i$.

5 Closed form

6 Appendix

6.1 Matrix calculus

The derivative of a scalar with respect to a vector, as was used with the loss function, is defined as:

$\frac{dL}{dv} = (\frac{dL}{dv_1}, \frac{dL}{dv_2}, ...)$

Similar to multivariable calculus, the components of v are independent variables; thus, the derivative of any one of them w.r.t. another one is zero, and w.r.t itself is one.

We can derive some simple derivatives by expanding the indices. Many formulas are quite similar to their scalar calculus counterparts.

$\frac{d}{dv}(\sum v_i) = \frac{d}{dv}(v_1 + v_2 + ...) = (1, 1, ...)$

$\frac{d}{dv}(v \cdot v) = \frac{d}{dv}(v^\intercal v) = \frac{d}{dv}(v_1^2 + v_2^2 + ...) = (2v_1, 2v_2, ...) = 2v$

The derivative of a vector w.r.t. scalar, vector w.r.t. vector, matrix w.r.t. scalar, scalar w.r.t. matrix also exist. In general, the derivative of an n dimensional quantity w.r.t. an m dimensional quantity is $n + m$ dimensional, to account for all the combinations of components.

For example, vector w.r.t vector is an N by M dimensional matrix:

$\frac{du}{dv} = \begin{bmatrix} \frac{du_1}{dv_1} & \frac{du_1}{dv_2} & ... \\ \frac{du_2}{dv_1} & \frac{du_2}{dv_2} & ... \\ ... & ... & ... \end{bmatrix}$ (or it's transpose, depending on convention).

$\frac{d}{dv}(Av) = \frac{d}{dv}(\sum A_{1i}v_i, \sum A_{2i}v_i, ...)^\intercal = \begin{bmatrix} A_{11} & A_{12} & ... \\ A_{21} & A_{22} & ... \\ ... & ... & ... \end{bmatrix} = A$

6.2 Orthogonal decomposition

Let V be a vector space with an inner product and induced norm. Let W be a subspace, and β be an orthonormal basis for W.

Any vector $v \in V$ can be uniquely expressed as $w + u$, where $w \in W$ and $u \in W^{\perp}$. Additionally, $\|w - v\| \leq \|x - v\| \quad \forall x \in W$.

$w = \sum \text{proj}_{\beta_i}(v) = \sum (v \cdot \beta_i)\beta_i$ (because β_i is normal). $u = v - w$.