Ordinary least squares

Given a training data set $S := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y \in \mathbb{R}$, ordinary least squares (OLS) is a regression algorithm for finding a linear model that minimizes the squared error on the training data. That is, given a data point $\mathbf{x} \in \mathbb{R}^d$, OLS considers hypotheses of the form

$$\begin{aligned} h_{\alpha,\boldsymbol{\beta}}(\mathbf{x}) &= \alpha + \sum_{i=1}^{n} \beta_i x_i \\ &= \alpha + \mathbf{x}^T \boldsymbol{\beta} \end{aligned}$$

Each hypothesis function is parameterized by the constant α and vector $\boldsymbol{\beta}$ where α is the translation of the dividing hyperplane and $\boldsymbol{\beta}$ are the coefficients. If we choose to append a 1 to each $\mathbf{x}$ vector and let the first element of $\boldsymbol{\beta}$ be α, then we can state the model more succinctly using only the $\boldsymbol{\beta}$ parameter:

$$h(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta} \tag{1}$$

The OLS algorithm specifically finds such a hypothesis that minimizes the squared error on the training set. That is, OLS solves

$$\begin{aligned} \hat{h} &:= \operatorname*{argmin}_h \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{squared}}(\mathbf{x}_i, y_i, h) \\ &= \operatorname*{argmin}_h \sum_{i=1}^{n} (y_i - h(\mathbf{x}_i))^2 \end{aligned}$$

Since each h is characterized by a $\boldsymbol{\beta}$, OLS finds

$$\hat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

As proven in Theorem 1, the solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

where $\mathbf{X}$ is the data matrix in which rows correspond to training samples and columns correspond to variables. An example of an OLS model fit to a dataset is illustrated in Figure 1.

Theorem 1

$$\underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} && \text{Note 1}
\end{aligned}$$

Now we will work to finding the $\boldsymbol{\beta}$ that minimizes this function. Note that since $\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$ is a quadratic equation in terms of $\boldsymbol{\beta}$, we can take the gradient with respect to $\boldsymbol{\beta}$ set it to the zero vector and solve for $\boldsymbol{\beta}$. This will find the $\boldsymbol{\beta}$ that minimizes the function.

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

Setting this to the zero vector and solving for $\boldsymbol{\beta}$ we get

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

Notes

1. In this step, we combined the second and third terms in the summation as follows: let's look at the second term $-\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$. If we take the transpose of the transpose of this object we get

$$\begin{aligned}
-\left(\left(\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \right)^T \right)^T &= -\left(\boldsymbol{\beta}^T \left(\mathbf{y}^T\mathbf{X} \right)^T \right)^T \\
&= \left(-\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} \right)^T
\end{aligned}$$

Note that the object inside the transpose is equal to the third term in the summation. Furthermore, we note that this term is actually a scalar when we examine the dimensions of this term:

$$\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} = \underset{1\times n}{\boldsymbol{\beta}^T} \underset{n\times m}{\mathbf{X}^T} \underset{m\times 1}{\mathbf{y}}$$

So we see that it is a 1×1 matrix, which is simply a scalar. Taking the transpose of a scalar results in a scalar, so it we can simply drop the transpose and combine the second and third terms.

2. In this step, we take the gradient of this function with respect to $\boldsymbol{\beta}$. The first term $\mathbf{y}^T\mathbf{y}$ clearly becomes the zero vector because there is no $\boldsymbol{\beta}$ present. Looking at the second term

$$\begin{aligned}
\nabla_{\beta} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} &= \begin{bmatrix} \frac{\partial(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y})}{\partial\beta_1} \\ \vdots \\ \frac{\partial(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y})}{\partial\beta_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial(-2\sum_{i=1}^n \beta_i(\mathbf{X}^T\mathbf{y})_i)}{\partial\beta_1} \\ \vdots \\ \frac{\partial(-2\sum_{i=1}^n \beta_i(\mathbf{X}^T\mathbf{y})_i)}{\partial\beta_n} \end{bmatrix} \\
&= \begin{bmatrix} -2(\mathbf{X}^T\mathbf{y})_1 \\ \vdots \\ -2(\mathbf{X}^T\mathbf{y})_n \end{bmatrix} \\
&= -2\mathbf{X}^T\mathbf{y}
\end{aligned}$$

And finally, we note that the third term is a quadratic form with $\mathbf{X}^T\mathbf{X}$ being the matrix of the quadratic form. Thus,

$$\nabla_{\beta}\, \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

□

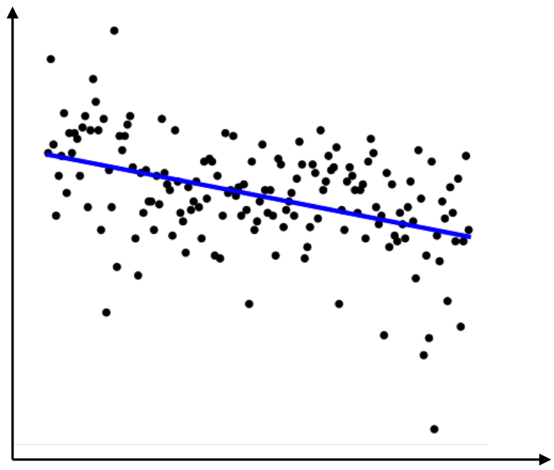

Figure 1: The blue line visualizes an OLS model fit to a set of data points in $\mathbb{R}^2$.