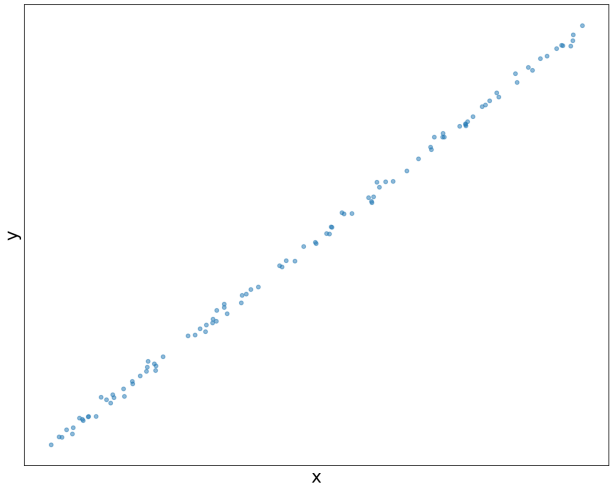
07-280 Linear Regression

Nihar B. Shah

Say you have data: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$.



You may want to understand the data or predict y from x. Let's fit a line!

Why do we care about fitting lines? (At least) two reasons:

- It is a simple, perhaps the simplest, thing to do when $\mathcal{Y}$ is real valued. It is a classical technique that works surprisingly well in many settings.
- Even in complex settings like neural networks or large language models, this can be useful. You have trained a neural network for one task, and now want to use it for another task. You may fix most of the trained neural network and modify only the "last layer" to tune it to your new task. Under real-valued labels, this amounts to fitting a linear model.

Now let us discuss how to actually fit a linear model, using the ML framework we had discussed in the first lecture.

Recall ERM:

$$\arg\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - h(x^{(i)}) \right)^2 \quad \leftarrow \text{squared loss since we have real valued labels}$$

Here, $\mathcal{H}$ = all functions $\mathbb{R} \to \mathbb{R}$ that can be represented as a line.

How do these functions look like? Two parameters $\omega \in \mathbb{R}$, $b \in \mathbb{R}$

$$h_{\omega,b}(x) = x \cdot \omega + b$$

How do we find the best fitting line, i.e., the arg min for ERM?

$$\arg\min_{\omega,b\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - (x^{(i)}\cdot\omega + b)\right)^2$$

For ease of understanding, suppose we consider $b = 0$. In other words, we want $h(0) = 0$. Then:

$$\arg\min_{\omega\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - x^{(i)}\cdot\omega\right)^2$$

$$= \arg\min_{\omega\in\mathbb{R}} \frac{1}{N}\left(\sum_{i=1}^{N} y^{(i)^2} + \omega^2\sum_{i=1}^{N} x^{(i)^2} - 2\omega\sum_{i=1}^{N} y^{(i)}x^{(i)}\right)$$

(by completing the squares)

$$= \arg\min_{\omega\in\mathbb{R}} \left(\omega\sqrt{\sum_{i=1}^{N} x^{(i)^2}} - \frac{\sum_{i=1}^{N} y^{(i)}x^{(i)}}{\sqrt{\sum_{i=1}^{N} x^{(i)^2}}}\right)^2$$

This is minimized when this term is zero, i.e., $\omega\cdot\sum_{i=1}^{N} x^{(i)^2} - \sum_{i=1}^{N} y^{(i)}x^{(i)} = 0$

$\therefore$ Minimizer is $\omega = \frac{\sum_{i=1}^{N} y^{(i)}x^{(i)}}{\sum_{i=1}^{N} x^{(i)^2}}$. The resulting h is a line $y = \omega x$ with slope ω.

(As a corner case, when will this solution behave weirdly? When the denominator is 0, that is, when $\sum_{i=1}^{N} x^{(i)^2} = 0$. Think about this situation – how would the plot of training points look like if this is the case?)

Let us now generalize this idea. Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$. Given training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$, we will consider functions $h : \mathbb{R}^d \to \mathbb{R}$ of the form $h(x) = x^T\omega + b$, where $\omega \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Without loss of generality, assume first coordinate of x is fixed at 1. Let $\theta = \begin{bmatrix} b \\ \omega \end{bmatrix}$. Consider $h_\theta(x) = x^T\theta$.

Then ERM: $\arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - x^{(i)^T}\theta\right)^2$.

Let $X = \begin{bmatrix} \leftarrow & x^{(1)^T} & \rightarrow \\ \leftarrow & x^{(2)^T} & \rightarrow \\ & \vdots & \\ \leftarrow & x^{(N)^T} & \rightarrow \end{bmatrix}$, $y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)}. \end{bmatrix}$.

Then ERM is $\arg\min_{\theta} \frac{1}{N}\|y - X\theta\|_2^2$. (By the way, can you simply set $\theta = X^{-1}y$? No.)

Expanding the squared term, we get ERM:

$$\arg\min_{\theta} (y - X\theta)^T(y - X\theta)$$

$$= \arg\min_{\theta} y^Ty - \theta^TX^Ty - y^TX\theta + \theta^TX^TX\theta$$

$$= \arg\min_{\theta} \; -2\theta^T X^T y + \theta^T X^T X \theta.$$

Let $A = X^T X$, $b = X^T y$ (relating to our previous simpler scalar case, $X^T X$ is like $\sum x_i^2$, $X^T y$ is like $\sum x_i y_i$). Assume A is invertible.

Then ERM: $\arg\min_{\theta} \; -2\theta^T b + \theta^T A \theta + b^T A^{-1} b - b^T A^{-1} b$

$$= \arg\min_{\theta} \; (\theta - A^{-1}b)^T A (\theta - A^{-1}b) - b^T A^{-1} b$$

$$= \arg\min_{\theta} \; \|X(\theta - (X^T X)^{-1} X^T y)\|_2^2$$

Minimized when $\theta = (X^T X)^{-1} X^T y$.

(As a corner case, one may be concerned about $X^T X$ not being invertible, that is, determinant of $X^T X$ being 0. In this case, instead of taking the inverse of $X^T X$, we take the pseudo-inverse. You don't have to worry about it for the purposes of this class, but are welcome to look it up online.)