Hands on Machine Learning

Linear Regression

Benedikt Zönnchen

February 11, 2022

Given m data points, e.g.,

$$(a_{1,1},\ldots,a_{1,n-1},b_1),\ldots,(a_{m,1},\ldots a_{m,n-1},b_m)$$

Let $\mathbf{a}_i = (a_{i,1},\ldots,a_{i,n})^T$. We search for a linear function f with $f:\mathbb{R}^n \to \mathbb{R}$, such that,

$$\forall i \in \{1,\ldots,m\} : f(\mathbf{a}_i) = x_0 + x_1 \cdot a_{i,1} + \ldots x_n \cdot a_{i,n-1} = \langle \mathbf{a}_i^T, \mathbf{x} \rangle = b_i$$

In other words, we search for $\mathbf{x}$, such that

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

with

$$\mathbf{A} = \begin{bmatrix} 1 & a_{1,1} & \ldots & a_{1,n-1} \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \ldots & \ldots & a_{m,n-1} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_0 \\ \vdots \\ x_{n-1} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

We want to solve the linear equation

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m > n$ (more rows than columns).

It is very likely that $\mathbf{A}^{-1}$ does not exists and therefore, there is no solution, i.e., no $\mathbf{x}$ satisfies the equation.

In general we search for

$$\arg\min_{\hat{\mathbf{x}} \in \mathbb{R}^n} \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\| = \|\mathbf{e}\| \tag{2}$$

$\|\mathbf{e}\|$ is the minimal Euclidean distance from $\mathbf{b}$ to $C(\mathbf{A})$ (column space of $\mathbf{A}$). Therefore, $\mathbf{A}\hat{\mathbf{x}} = \mathbf{p}$ is the projection of $\mathbf{b}$ onto $C(\mathbf{A})$.

Since $\mathbf{b} - \mathbf{A}\hat{\mathbf{x}} = \mathbf{e} \perp C(\mathbf{A})$, it follows that

$$\mathbf{A}^T\mathbf{p} = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0} \tag{3}$$

Therefore, we solve

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b} \tag{4}$$

for which a solution exists!

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{5}$$

We can also compute the projection matrix

$$\mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \tag{6}$$