1.1 - Linear Regression

Linear regression

We try to approximate f by a function $\hat{f} : \mathbb{R} \to \mathbb{R}$, $x \to wx + b$, where $w, b \in \mathbb{R}$
How to select w and b that approximate the data quite well?

Question: How to find $w, b \in \mathbb{R}$ such that $\hat{f}(x) \sim f(x)$?
--> Find w, b such that $\forall i, \hat{f}(x_i) \sim y_i$

Sum of squared deviations

$$L(w,b) = \frac{1}{2} \sum_{i=1}^{N} |wx_i + b - y_i|^2$$

$(wx_i + b - y_i)$ is the distance from data points to estimated function
This will give an exact fit for the data
We want to find w, b which minimizes $L(w, b)$
You may be wondering why there's a $\frac{1}{2}$ is in the loss function. Well scaling a loss function by a constant does not change the location of its minimum (optimal w and b remain the same whether $\frac{1}{2}$ is included or not). Also it simplifies derivatives because when taking the derivate the square cancels out the $\frac{1}{2}$ which leaves a cleaner expression.

1st observation: L is **differentiable**
2nd observation: L is convex
(w, b) is a minimum of $L \Leftrightarrow \partial_w L(w,b) = \partial_b L(w,b) = 0$
((is the same as)$\Leftrightarrow \nabla L(w,b) = 0$)

1.3.3

$$\partial_w L(w,b) = \frac{1}{2} \sum_{i=1}^{N} (wx_i + b - y_i)^2$$

Apply chain rule $f'(g(x))g'(x)$ where $f(x) = x^2$ and $g(x) = (wx_i + b - y_i)$

$$\text{chain rule} = 2(wx_i + b - y_i) \cdot \frac{\partial}{\partial w}(wx_i + b - y_i)$$

$$= \frac{1}{2} \sum_{i=1}^{N} \cdot 2(wx_i + b - y_i) \cdot x_i$$

$$= \sum_{i=1}^{N} x_i(wx_i + b - y_i)$$

Now expanding $(wx_i + b - y_i)x_i = w_x i^2 + bx_i - y_i x_i$

$$= w\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i y_i$$

$$\partial_b L(w,b) = \sum_{i=1}^{N}(wx_i + b - y_i)$$

Also applying chain rule here:

$$\frac{1}{2}\sum_{i=1}^{N} 2(wx_i + b - y_i) \cdot \frac{\partial}{\partial b}(wx_i + b - y_i)$$

$$= \sum_{i=1}^{N}(wx_i + b - y_i)$$

$$= w\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} 1 - \sum_{i=1}^{N} y_i$$

$b\sum_{i=1}^{N} = b \cdot N$, thus:

$$= w\sum_{i=1}^{M} x_i + Nb - \sum_{i=1}^{N} y_i$$

The necessary and sufficient optimality conditions for

$$\min_{w,b\in\mathbb{R}} \mathcal{L}(w,b)$$

are given by $\frac{\partial}{\partial w}\mathcal{L}(w,b) = 0$ and $\frac{\partial}{\partial b}\mathcal{L}(w,b)$ and are equivalent to the linear system

1.3.4

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}$$

Explanation

Let's reconsider matrix multiplications to understand the linear system given above:
The first element (11) is multiplicated with (1) of the vector and added up with the product of $\sum i = 1^N x_i * b$. Those products are added up so:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} wx_i^2 + \sum_{i=1}^{N} x_i b \\ \sum_{i=1}^{N} x_i w + Nb \end{pmatrix} = \begin{pmatrix} w\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i \\ w\sum_{i=1}^{N} x_i + Nb \end{pmatrix}$$

But how do we conclude to

$$\begin{pmatrix} \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i \end{pmatrix}$$

(here: $\frac{\partial}{\partial w}$) and after the AND its $\frac{\partial}{\partial b}$

So how do we conclude the vector $\begin{pmatrix} \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i \end{pmatrix}$? Look at the equ

derivaties above. Consider

$$w \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i - \sum_{i=1}^N x_i y_i = 0 \quad / + \left(\sum_{i=1}^N x_i y_i\right)$$

$$\Leftrightarrow \quad w \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i = \sum_{i=1}^N x_i y_i$$

AND

$$w \sum_{i=1}^N x_i + Nb - \sum_{i=1}^N y_i = 0 \quad / + \left(\sum_{i=1}^N y_i\right)$$

$$\Leftrightarrow \quad w \sum_{i=1}^N x_i + Nb = \sum_{i=1}^N y_i$$

We can substitute this in the linear equation system and thus get

$$\begin{pmatrix} \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i \end{pmatrix}$$

The explanation above should be enough to conclude that we get to the final equation in 1.3.4

This can be further simplified by defining the averages $\overline{x} := \frac{1}{N}\sum_{i=1}^N x_i$ and $\overline{y} = \frac{1}{N}\sum_{i=1}^N y_i$ Now by dividing every entry of the matrices by N we get to:

$$\begin{pmatrix} \frac{1}{N}\sum_{i=1}^N x_i^2 & \overline{x} \\ \overline{x} & 1 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N x_i y_i \\ \overline{y} \end{pmatrix}$$

The determinant of the system matrix is:

$$\frac{1}{N}\sum_{i=1}^N x_i^2 - \overline{x}^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \overline{x})^2$$

and is non-zero **exactly when** there are at least **two distinct** data points x_i. In this case we can explicitly calculate w and b, and the solution is given by (see the exercise sheet)

Problem 2

We can rewrite it as a system of linear equations:

$$\begin{cases} w \frac{1}{N} \sum_{i=1}^{N} x_i^2 + b \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i \\ w \bar{x} + b = \bar{y} \end{cases}$$

Where we can immediately derive b:

$$b = \bar{y} - w \bar{x}$$

Let's substitute b into the first equation:

$$w \cdot \frac{1}{N} \sum_{i=1}^{N} x_i^2 + (\bar{y} - w\bar{x}) \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i \equiv$$

$$\equiv w \cdot \frac{1}{N} \sum_{i=1}^{N} x_i^2 + \bar{x}\bar{y} - w \bar{x}^2 = \frac{1}{N} \sum_{i=1}^{N} x_i y_i \equiv$$

$$\equiv w \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 - \bar{x}^2 \right) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

where $\frac{1}{N} \sum_{i=1}^{N} x_i^2 = E[x_i^2]$, $\bar{x}^2 = E[x_i]^2$,

$\frac{1}{N} \sum_{i=1}^{N} x_i y_i = E[x \cdot y]$.

We also know, that $Var(X) = E[(x_i - E[x])^2] =$

$= E[x_i^2] - E[x_i]^2$

and $Cov(X, Y) = E[x \cdot y] - E[x] \cdot E[y] =$

$= E[(x_i - E[x])(y_i - E[y])]$.

We can rewrite it:

$$w = \frac{cov(X, Y)}{Var(X)} = \frac{\sum_{i=1}^{N} (x - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

1.3.5

$$w = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

$$b = \overline{y} - w\overline{x}$$

Unlike nearest-neighbor interpolation, the linear regression function $\hat{f}(x) := wx + b$ is continuous and easy to evaluate. Additionally, it is more robust to errors in the data (x_i, y_i)
Remark: Let

$$X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix} \in \mathbb{R}^{Nx2}$$

$$\beta = \begin{pmatrix} w \\ b \end{pmatrix} \in \mathbb{R}^2$$

$$y = \begin{pmatrix} y_i \\ . \\ . \\ . \\ y_N \end{pmatrix}$$

we can rewrite the sum of squared deviations 1.3.3

$$L(\beta) = \frac{1}{2}||X\beta - y||^2$$

where $||\cdot||$ denotes the euclidian norm on $\mathbb{R}^N$ (normal Euclidian norm: $\sqrt{x_1^2 + x_2^2 + \cdots + x_N^2}$)

Explanation

Why does this work?
When multiplying $X\beta$ we get

$$X\beta = \begin{pmatrix} x_1 w + b \\ x_2 w + b \\ x_n w + b \end{pmatrix}$$

Remember the euclidian norm above. This effectively replaces the sum $\sum_{i=1}^{N} wx_i + b$. Thus

$$L(\beta) = \frac{1}{2}||(X\beta - y)_i^2 = \frac{1}{2}||X\beta - y||^2$$

Exercise 1.3.1

Show that for a model of the form $\hat{f}(x) = wx$ with parameter $w \in \mathbb{R}$, the least squares solution is given by

$$w = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}$$

Solution: We have $\mathcal{L}(w) := \frac{1}{2}\sum_{i=1}^N (wx_i - y_i)^2$
$\frac{\partial}{\partial w}\mathcal{L} = \sum_{i=1}^N (wx_i - y_i)x_i$. Since we want to minimize the function we set $\frac{\partial}{\partial w}\mathcal{L} = 0$:

$$\begin{aligned}
\sum_{i=1}^N (wx_i - y_i)x_i &= 0 \\
\sum_{i=1}^N wx_i^2 - y_i x_i &= 0 \quad | + \sum_{i=1}^N x_i y_i \\
w\sum_{i=1}^N x_i^2 &= \sum_{i=1}^N x_i y_i \quad | \div \sum_{i=1}^N x_i^2 \\
w &= \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}
\end{aligned}$$

$\square$

Exercise 1.3.2

Show that the gradient of the function L in 1.3.3 is given by:

$$\nabla L(\beta) = X^T(X\beta - y)$$

Remember that for any vector,

$$|| \cdot ||^2 = v^T v$$

Since $|| \cdot ||^2$ is just adding up the squared matrix elements its equivalent (multiplying each element with itself and then add all elements up)
So:

$$||X\beta - y||^2 = (X\beta - y)^T(X\beta - y)$$

And the loss function would be

$$\mathcal{L}(\beta) = \frac{1}{2}(X\beta - y)^T(X\beta - y)$$

$$(X\beta - y)^T(X\beta - y) = (X\beta)^T X\beta - 2(X\beta)^T y + y^T y$$

Since the scalar transpose ($(X\beta)^T$ would be a $1 \times m$ row vector and we have a property for scalar values: $a^T = a$) does not change the value:
Remember that $(X\beta)^T y = y^T(X\beta)$. Since $(X\beta)^T = (X\beta)$ if its a scalar we can rewrite

$$= \beta^T X^T X\beta - 2y^T X\beta + y^T y$$

We add $\frac{1}{2}$ again:

$$\mathcal{L}(\beta) = \frac{1}{2}[\beta^T X^T X\beta - 2y^T X\beta + y^T y]$$

Thus:

$$\mathcal{L}(\beta) = \frac{1}{2}[\beta^T X^T X\beta - 2y^T X\beta + y^T y]$$

Now for any quadratic from like $\beta^T A\beta$, where A is a symmetric matrix, the gradient with respect to β is:

$$\nabla_\beta(\beta^T A\beta) = 2A\beta$$

Since in $(\beta^T X^T X\beta)$ $X^T X$ is symmetric, applying this formula gives:

$$\nabla_\beta(\beta^T X^T X\beta) = 2X^T X\beta$$

However, we still have $\frac{1}{2}$ in front of it so we get:

$$\nabla_\beta(\frac{1}{2}\beta^T X^T X\beta) = X^T X\beta$$

Differentiating second term $-2y^T X\beta$

For a linear term like $c^T\beta$ the gradient is simply the coefficient:

$$\nabla_\beta(c^T\beta) = c$$

Here, treating $X^T y$ as a constant vector we get

$$\nabla_\beta(\frac{1}{2} - 2y^T X\beta) = -X^T y$$

Now differentiating $y^T y$ which does not contain β this is 0
Combining the results we get

$$\nabla_\beta\mathcal{L}(\beta) = X^T X\beta - X^T y$$
$$\nabla\beta\mathcal{L}(\beta) = X^T(X\beta - y)$$