Linear Regression

Recap

Thomas Wedenig

April 10, 2024

Institute of Theoretical Computer Science
Graz University of Technology, Austria

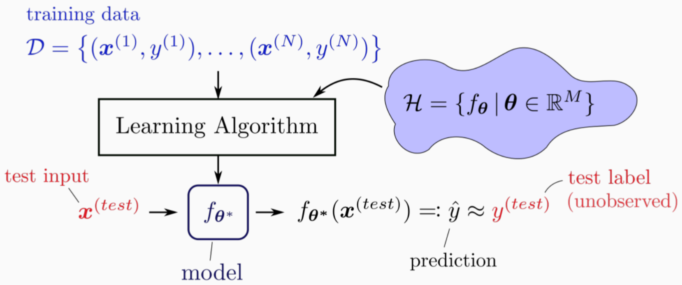
- Inputs: features $x_1, x_2, \ldots, x_D$, collected in $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T \in \mathbb{R}^D$
- Given $\mathbf{x}$, we want to predict a **real-valued target** $y \in \mathbb{R}$

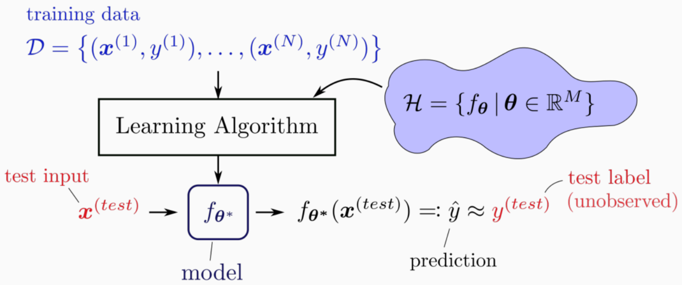
Example

Given knee height in cm (x_1) and arm span in cm (x_2),
predict the body height in cm (y).

- We collect a **training dataset** with many such input-output pairs

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)} \right) \right\}$$


training data
$\mathcal{D} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$
$\mathcal{H} = \{f_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \mathbb{R}^M\}$
Learning Algorithm
test input
$\boldsymbol{x}^{(test)}$
$f_{\boldsymbol{\theta}^*}$
model
$f_{\boldsymbol{\theta}^*}(\boldsymbol{x}^{(test)}) =: \hat{y} \approx y^{(test)}$
prediction
test label
(unobserved)



- **Linear Regression:** We assume an (approx.) **affine relationship** between $\mathbf{x}$ and y:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = b + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_D \cdot x_D$$

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = b + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_D \cdot x_D$$

- Collect $\boldsymbol{\theta} = (b, w_1, w_2, \ldots, w_D)^T$
- Let's add a **constant feature** (dummy feature) to each $\mathbf{x}$:

$$\mathbf{x} = (1, x_1, \ldots, x_D)^T$$

- We can now compactly write

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta}$$

$$x = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{pmatrix}, \quad \theta = \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$$

Loss function !

- We need a way to **quantify how "good" a given $\boldsymbol{\theta}$** is
- We call this the **loss** $\mathcal{L}(\boldsymbol{\theta})$
  - Small values are "good"

- In regression, we often use the **Mean Squared Error** as the loss:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \left(f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)} \right)^2$$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 \qquad f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\theta}$$

- Collect data points in the **rows** of a matrix (**data matrix** X) and targets in a **vector**

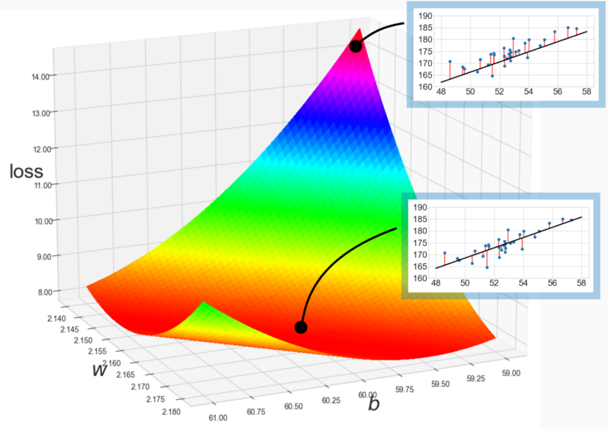
$$X = \begin{pmatrix} {\mathbf{x}^{(1)}}^T \\ {\mathbf{x}^{(2)}}^T \\ \vdots \\ {\mathbf{x}^{(N)}}^T \end{pmatrix} \qquad \mathbf{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(N)}\right)^T \in \mathbb{R}^n$$

$$X\theta = \begin{bmatrix} — x^{(1)T} — \\ \vdots \\ — x^{(N)T} — \end{bmatrix} \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_D \end{pmatrix} = \begin{pmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(N)} \end{pmatrix}$$

- Then, we can re-write the loss as

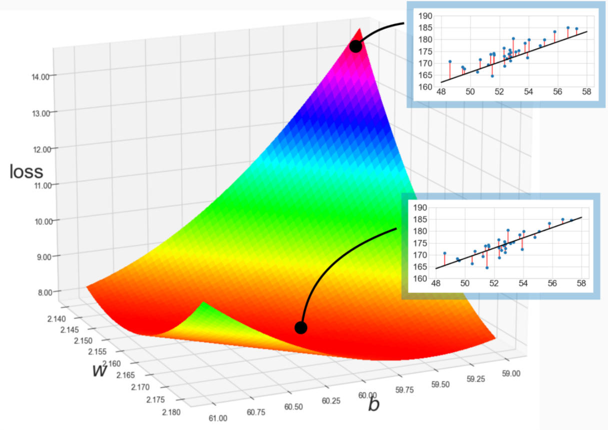
$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\|X\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

This choice of $\mathcal{L}$ and f_θ yields a loss function that is **convex in θ** 🥰



Minimizing the loss function

This choice of $\mathcal{L}$ and $f_{\boldsymbol{\theta}}$ yields a loss function that is **convex in $\boldsymbol{\theta}$** 🥰



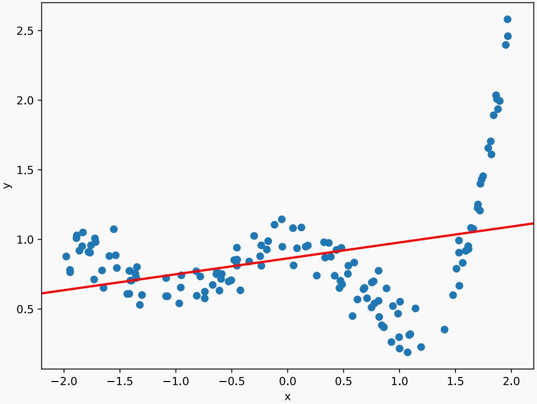
- **Goal:** find

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\theta})$$

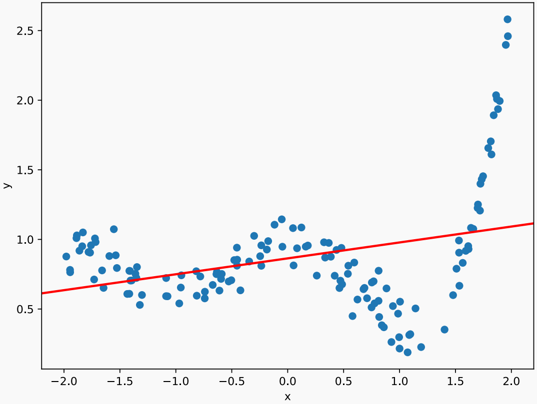
- $\boldsymbol{\theta}^*$ is **unique** iff X has **full rank**
- **There exists a closed-form solution for the global minimum** 😍

$$\boldsymbol{\theta}^* = \underbrace{(X^T X)^{-1} X^T}_{\text{Pseudoinverse}} \, \mathbf{y}$$

A word on Linearity

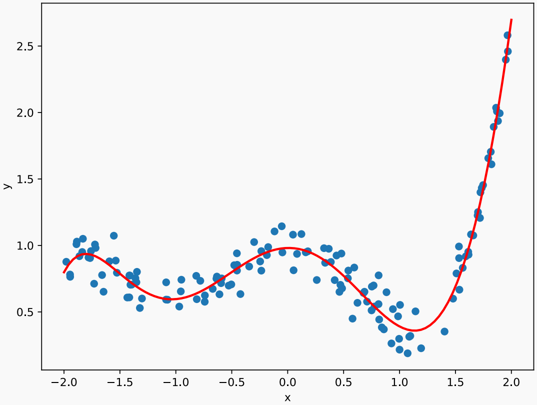


Is this the result of **linear** regression? 🤔

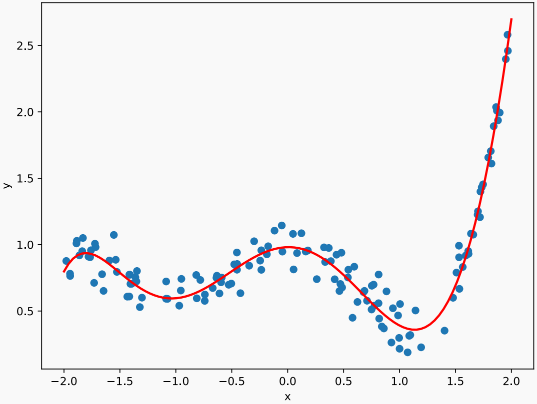


Yes! The function in red is of the form

$$f_{\boldsymbol{\theta}}(x) = b + w \cdot x$$



Is this still **linear** regression? 🤔



Polynomial Regression is still **linear** in $\boldsymbol{\theta}$!

$$f_{\boldsymbol{\theta}}(x) = b + w_1 \cdot x + w_2 \cdot x^2 + \cdots + w_6 \cdot x^6$$

$$f_{\boldsymbol{\theta}}(x) = b + w_1 \cdot x + w_2 \cdot x^2 + \cdots + w_6 \cdot x^6$$

- It's just **non-linear** in x
- Let $\phi(x) = (1 \;\; x \;\; x^2 \;\; x^3 \ldots \;\; x^6)^T$

$$f_{\boldsymbol{\theta}}(x) = \phi(x)^T \boldsymbol{\theta}$$

Dimensionality

- **Note:** x is 1D while $\phi(x)$ and $\boldsymbol{\theta}$ are 7D !
- We lift x into a higher dimensional space
- Hypothesis: There is an affine relationship between the **transformed features** $\phi(x)$ and the target

Linear Regression Demo