Tuesday, February 3

- Reading available for every lecture
- Typos / mistakes now have to be submitted after class
- Research talk Friday 11-12:15 in Davidson Lecture Hall

Roadmap

ML used to solve problems like:

- predict temperature from present weather,
- classify objects in an image,

} supervised learning (= "right" answer)

- generate the next word in a sentence
- make an image from text description

} unsupervised learning (no right answer)

Supervised Learning

Input: Training data

$x^{(1)} \in \mathbb{R}^d$ $\quad y^{(1)} \in \mathbb{R}$

$x^{(2)}$ $\quad y^{(2)}$

$\vdots$ $\quad \vdots$

$x^{(n)}$ $\quad y^{(n)}$

e.g., weather, image classification, stock prices

Output: $f: \mathbb{R}^d \to \mathbb{R}$ s.t.

$$f(x^{(i)}) \approx y^{(i)}$$

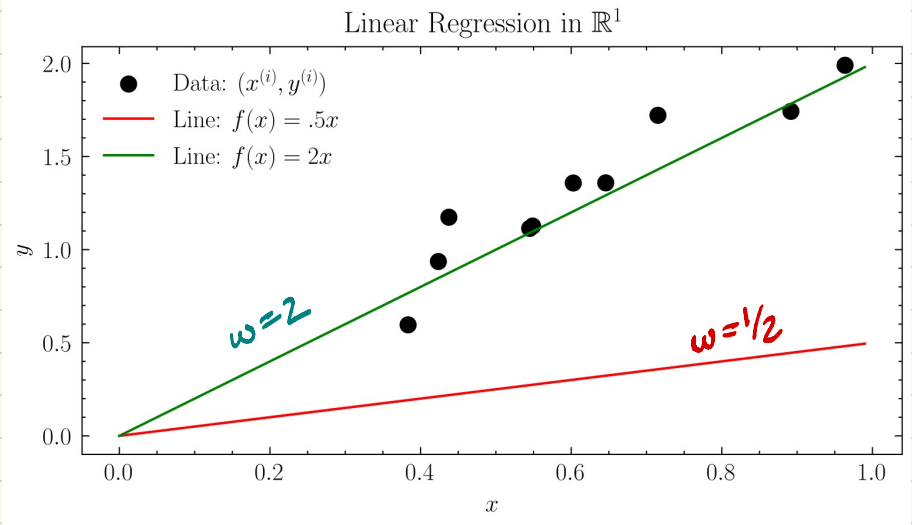
Our Strategy

1. Function class / model
2. Loss
3. Optimizer

Univariate Linear Regression

$x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}$ i.e., single variables

① Model: $f(x) = wx$ for some $w \in \mathbb{R}$



Q: Which line is better?

Let's formalize...

Attempt #1: $y^{(i)} - f(x^{(i)})$

Attempt #2: $|y^{(i)} - f(x^{(i)})|$

↑ absolute loss

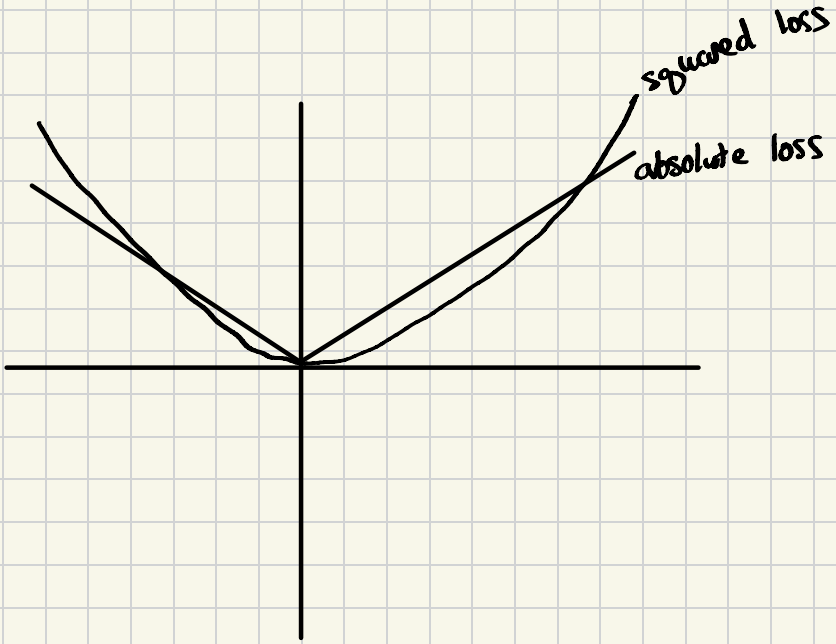
Attempt #3: $(y^{(i)} - f(x^{(i)}))^2$

↑ squared loss

② Loss: Mean Squared Error (MSE) Loss

$$\mathcal{L}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - f(x^{(i)})\right)^2$$

- differentiable
- penalizes bad predictions <u>more</u>



③ Optimization: Derive optimal w

Q: How do we optimize convex function?

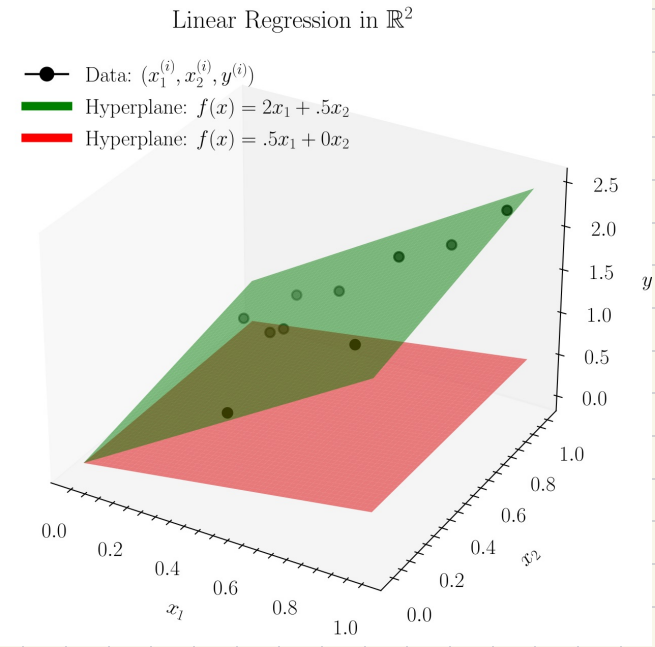
A: Set $\frac{\partial \mathcal{L}(w^*)}{\partial w} = 0$ and solve for w^*

Should get $w^* = \frac{\sum_{i=1}^{n} y^{(i)} x^{(i)}}{\sum_{i=1}^{n} (x^{(i)})^2}$

Multivariate Linear Regression

$x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$

① Model: $f(x) = w^T x = \sum_{j=1}^{d} w_j x_j$



② MSE Loss!

$$\mathcal{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \left(f(x^{(i)}) - y^{(i)} \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)T} w - y^{(i)} \right)^2$$

$$= \frac{1}{n} \| Xw - y \|_2^2$$

$$X = \underset{n \times d}{\begin{bmatrix} — x^{(1)T} — \\ \vdots \\ — x^{(n)T} — \end{bmatrix}} \qquad y = \underset{n \times 1}{\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}}$$

Thursday, February 5

- Shapley values for causal inference
  - ↳ tomorrow 11am @ Davidson Lecture Hall
- GEMS Mentors!
  - ↳ Saturday 9:30am @ Shanahan
    (Please email me)

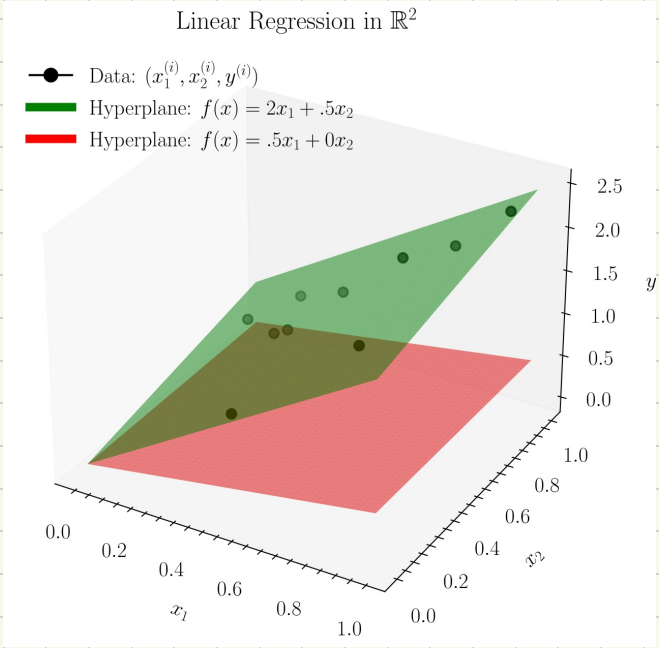
Plan

- Multivariate Linear Regression
- Empirical Risk Minimization

Multivariate Linear Regression

$x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$

① Model: $f(x) = \omega^T x = \sum_{j=1}^{d} \omega_j x_j$



② MSE Loss!

$$\mathcal{L}(\omega) = \frac{1}{n} \sum_{i=1}^{n} \left(f(x^{(i)}) - y^{(i)} \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)T} \omega - y^{(i)} \right)^2$$

$$= \frac{1}{n} \| X\omega - y \|_2^2$$

$$\underset{n \times d}{X} = \begin{bmatrix} — x^{(1)T} — \\ \vdots \\ — x^{(n)T} — \end{bmatrix} \qquad \underset{n \times 1}{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

(3) Exact Optimization

Set $\nabla_w \mathcal{L}(w^*) = 0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and solve for w^*

$$\nabla_w \mathcal{L}(w) = \begin{bmatrix} \frac{\partial}{\partial w_1} \mathcal{L}(w) \\ \vdots \\ \frac{\partial}{\partial w_d} \mathcal{L}(w) \end{bmatrix}$$

$$\frac{\partial \mathcal{L}(w)}{\partial w_i} = \lim_{\Delta \to 0} \frac{\mathcal{L}(w + \Delta e_i) - \mathcal{L}(w)}{\Delta}$$

$$= \lim_{\Delta \to 0} \frac{1}{n} \frac{\| X(w + \Delta e_i) - y \|_2^2 - \| Xw - y \|_2^2}{\Delta}$$

$$= \lim_{\Delta \to 0} \frac{1}{n} \frac{\| Xw - y \|_2^2 - 2 (Xe_i)^T (Xw - y) + \Delta^2 \| Xe_i \|_2^2 - \| Xw - y \|_2^2}{\Delta}$$

$$= \lim_{\Delta \to 0} \frac{1}{n} -2 (Xe_i)^T (Xw - y) + \Delta \| Xe_i \|_2^2$$

$$= -\frac{2}{n} (Xe_i)^T (Xw - y) = -\frac{2}{n} X_{:i}^T (Xw - y)$$

$$\nabla_w \mathcal{L}(w) = -\frac{2}{n} X^T (Xw - y)$$

$$\| a + b \|_2^2 = (a + b)^T (a + b)$$

$$= a^T a + a^T b + b^T a + b^T b$$

$$= \| a \|_2^2 + 2 a^T b + \| b \|_2^2$$

Solve for w^* when

$$\nabla_w \mathcal{L}(w^*) = -\frac{2}{n} X^T(Xw^* - y) = 0$$

$$X^T X w^* - X^T y = 0$$

$$X^T X w^* = X^T y$$

$$w^* = (X^T X)^{-1} X^T y$$

↑ assume invertible (not necessary)

Q: Time to compute $(X^T X)^{-1} X^T y$?

1. Time for $X^T y$
2. Time for $X^T X$
3. Time for $(X^T X)^{-1}$ is $O(d^3)$

Linear Algebraic View

Singular Value Decomposition

(aka eigendecomposition for rectangular matrices)

$$\underset{n \times d}{X} = \sum_{i=1}^{r} \sigma_i \underset{n \times 1}{u_i} \underset{1 \times d}{v_i^T}$$

$v_1, \ldots, v_r$ orthonormal

$u_1, \ldots, u_r$ orthonormal

$u_i^T v_j = \{ ?$

$$X^T = \sum_{i=1}^{r} \sigma_i v_i u_i^T$$

X^+ ← pseudoinverse

$$X^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T$$

Show: $(X^T X)^+ X^T = X^+$

Empirical Risk Minimization

Answer to why squared-error

Suppose we believe

$$y^{(i)} = \langle w^*, x^{(i)} \rangle + \eta^{(i)}$$

same as $w^{*T} x^{(i)}$ but pretty :)

for noise $\eta \sim \mathcal{N}(0, \sigma^2)$

↑ normal dist

Then $y^{(i)} \sim \mathcal{N}(\langle w^*, x^{(i)} \rangle, \sigma^2)$

$$Pr(y^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y^{(i)} - \langle w, x^{(i)} \rangle)^2}{2\sigma^2} \right)$$

Idea: Find w that maximizes likelihood of observing data

$$\begin{aligned}
&\operatorname*{argmax}_w Pr(y^{(1)}, \dots, y^{(n)}) \\
&= \operatorname*{argmax}_w \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(y^{(i)} - \langle w, x^{(i)} \rangle)^2}{2\sigma^2} \right) \\
&= \operatorname*{argmax}_w \exp\left(\sum_{i=1}^{n} - \frac{(y^{(i)} - \langle w, x^{(i)} \rangle)^2}{2\sigma^2} \right) \\
&= \operatorname*{argmin}_w - \log\left(\exp\left(\sum_{i=1}^{n} - \frac{(y^{(i)} - \langle w, x^{(i)} \rangle)^2}{2\sigma^2} \right) \right) \\
&= \operatorname*{argmin}_w - \sum_{i=1}^{n} - \frac{(y^{(i)} - \langle w, x^{(i)} \rangle)^2}{2\sigma^2} \\
&= \operatorname*{argmin}_w \sum_{i=1}^{n} (y^{(i)} - \langle w, x^{(i)} \rangle)^2
\end{aligned}$$