Week 2

9/2/2025

↳ Quiz!

↳ How many read the notes? Important!

↳ Encourage practice, problem sets due every week

- First 2 problems still 9/22
- In general, due Mondays 11:59 pm

↳ Office hours change:
Monday + Wednesday 12:30 to 2

↳ Post questions, come to office hours :

↳ username on discord, so I can follow up!

↳ How many on discord?

Last week was warm up for machine learning!

Review

What about when M is not square?

$A \in \mathbb{R}^{d \times d}$

$$A = \sum_{i=1}^{r} v_i \lambda_i w_i^T$$

Use: Page Rank where we powered up matrix multiplication!

$\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ eigenvalues

$v_1, \ldots, v_r \in \mathbb{R}^d$ right eigenvectors $\quad A v_i = \lambda_i v_i$

$w_1, \ldots, w_r \in \mathbb{R}^d$ left eigenvectors $\quad w_i^T A = \lambda_i w_i^T$

$$v_i^T v_j = \mathbb{1}[i = j] = w_i^T w_j$$

$$w_j^T A v_i = w_j^T \lambda_i v_i \iff w_j^T v_i = \mathbb{1}[i = j]$$

$$\lambda_j w_j^T v_i = \lambda_i w_j^T v_i$$

$$(\lambda_j - \lambda_i) w_j^T v_i = 0$$

Supervised Learning

Problems like:

- predicting temperature
- identifying objects in an image
- generating next word

Labelled data

$\underline{x} \in \mathbb{R}^d$ $y \in \mathbb{R}$

n points...

$(\underline{x}^{(1)}, y^{(1)}), \ldots, (\underline{x}^{(n)}, y^{(n)})$

Goal: Find function $f: \mathbb{R}^d \to \mathbb{R}$

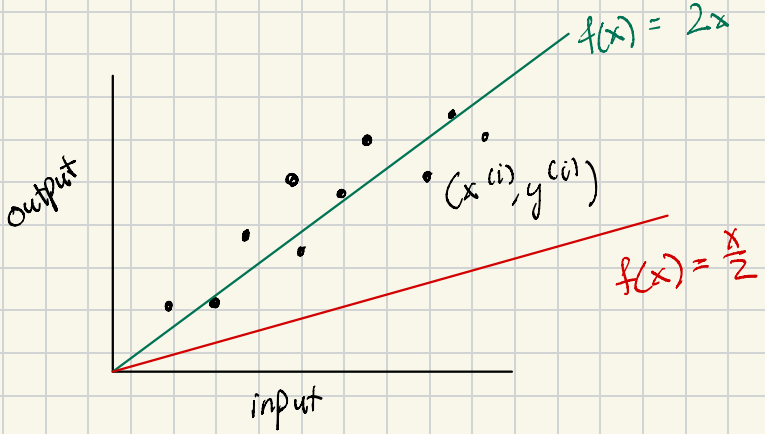
so that $f(x^{(i)}) \approx y^{(i)} \quad \forall\, i$

Empirical risk minimization:

1. Function class $\mathcal{F}$ from which to select f
2. Loss to measure how well f fits data
3. Optimizer, method to select f with low loss

Univariate Linear Regression

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



① Function Class:

$f(x) = wx$ for $w \in \mathbb{R}$

② Loss

Attempt #1: $f(x) - y$

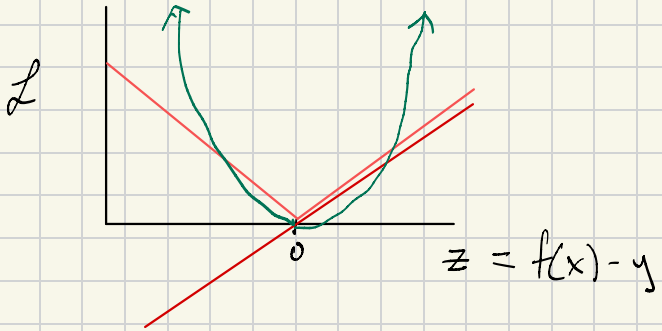
☹ $y > f(x) \Rightarrow$ very negative

Attempt #2: $|f(x) - y|$

☹ not differentiable at 0

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

☺ far gets penalized more



Mean Squared Error (MSE) Loss

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

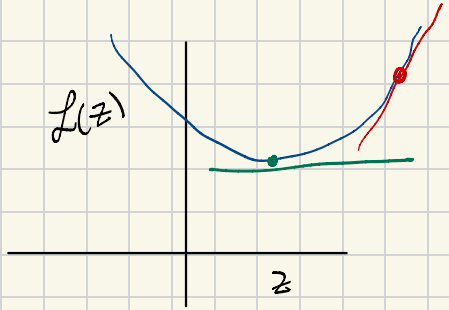
③ Exact Optimization

Key insight: Squared loss is convex

↳ differentiable

↳ single minima

Q: How do we find minima?



no improvement in any direction!

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

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

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

set $= 0$

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

$$w^* = \frac{\sum_{i=1}^{n} y^{(i)} x^{(i)}}{\sum_{i=1}^{n} \left[x^{(i)}\right]^2}$$

In general, $x^{(i)} \in \mathbb{R}^d$

Again, but with some linear algebra

9/4/2025

Reminders

- Notes?
- Pset 2 due Monday
  (self grade following Friday)
- Quiz
  - grades released
  - let's chat ☺

Review

$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

Goal: f so that $f(x^{(i)}) \approx y^{(i)}$

1. Function Class
2. Loss
3. Optimizer

	Tuesday	Thursday
dim?	$d=1$	$d \geq 1$
1	Linear	Linear
2	MSE	MSE
3	Exact	Exact

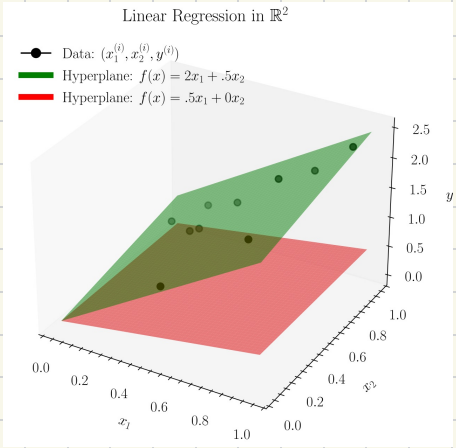
Multivariate Linear Regression

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

(1) Function Class

Let weights $\underline{w} \in \mathbb{R}^d$ (coefficient for each feature)

$$f(\underline{x}) = \langle \underline{x}, \underline{w} \rangle = \underline{x}^T \underline{w} = \sum_{k=1}^{d} w_k x_k$$



$f(x) = [2 \;\; \frac{1}{2}] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$f(x) = [1/2 \;\; 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(2) MSE

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

$\underline{X} \in \mathbb{R}^{n \times d}$ $\qquad \underline{y} \in \mathbb{R}$

$$\begin{bmatrix} - \; \underline{x}^{(1)T} \; - \\ - \; \underline{x}^{(2)T} \; - \\ \vdots \\ - \; \underline{x}^{(n)T} \; - \end{bmatrix} \qquad \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$\frac{1}{n} \| \underline{X} \underline{w} - \underline{y} \|_2^2 = \mathcal{L}(w)$$

$$\frac{1}{n} \left\| \begin{bmatrix} \langle \underline{x}^{(1)}, \underline{w} \rangle \\ \langle \underline{x}^{(2)}, \underline{w} \rangle \\ \vdots \\ \langle \underline{x}^{(n)}, \underline{w} \rangle \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} \right\|_2^2$$

③ Optimizer

Key insight:

$\underline{w}^*$ when no improvement in any direction

$$\nabla_w \mathcal{L}(w) = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_1} \\ \frac{\partial \mathcal{L}}{\partial w_2} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial w_d} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\underline{e}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i\text{th position}$$

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

$$= \lim_{\Delta \to 0} \frac{\| \underline{X}^T(\underline{w} + \Delta \underline{e}_i) - \underline{y} \|_2^2 - \| \underline{X}^T \underline{w} - \underline{y} \|_2^2}{\Delta} = (*)$$

$$\| \underline{a} + \underline{b} \|_2^2 = \sum_{i=1}^{d} (a_i + b_i)^2 = (\underline{a} + \underline{b})^T (\underline{a} + \underline{b})$$

$$= \underline{a}^T \underline{a} + \underline{a}^T \underline{b} + \underline{b}^T \underline{a} + \underline{b}^T \underline{b}$$

$$= \| \underline{a} \|_2^2 + 2 \langle \underline{a}, \underline{b} \rangle + \| \underline{b} \|_2^2$$

$$\underline{a} = \underline{X}^T \underline{w} - \underline{y} \qquad \underline{b} = \Delta \underline{X}^T \underline{e}_i$$

$$(*) = \lim_{\Delta \to 0} \frac{\| \underline{X}^T \underline{w} - \underline{y} \|_2^2 + 2 \langle \underline{X}^T \underline{w} - \underline{y}, \Delta \underline{X}^T \underline{e}_i \rangle + \| \Delta \underline{X}^T \underline{e}_i \|_2^2 - \| \underline{X}^T \underline{w} - \underline{y} \|_2^2}{\Delta}$$

$$= \lim_{\Delta \to 0} \frac{2 \Delta \langle \underline{X}^T \underline{w} - \underline{y}, \underline{X}^T \underline{e}_i \rangle + \Delta^2 \| \underline{X}^T \underline{e}_i \|_2^2}{\Delta}$$

$$= 2 \langle \underline{X}^T \underline{w} - \underline{y}, \underline{X}^T \underline{e}_i \rangle$$

$$\underline{X}\,\underline{e}_i = \underline{X}_i$$

$$\begin{bmatrix} — & \underline{x}^{(1)T} & — \\ & \vdots & \\ — & \underline{x}^{(n)T} & — \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$\nabla_{\underline{w}} \mathcal{L}(\underline{w}) = 2 \begin{bmatrix} \underline{X}_1^T (\underline{X}\underline{w} - \underline{y}) \\ \vdots \\ \underline{X}_i^T (\underline{X}\underline{w} - \underline{y}) \end{bmatrix} = 2\,\underline{X}^T (\underline{X}\,\underline{w} - \underline{y})$$

$$\nabla_{\underline{w}} \mathcal{L}(\underline{w}^*) = \underline{0} = 2\,\underline{X}^T (\underline{X}\,\underline{w}^* - \underline{y})$$

$$\underline{X}^T \underline{X}\,\underline{w}^* = \underline{X}^T \underline{y}$$

$$\underline{w}^* = (\underline{X}^T \underline{X})^+ \underline{X}^T \underline{y}$$

Singular Value Decomposition

assume features indep

$$\underline{X} = \sum_{i=1}^{d} \sigma_i \underbrace{\underline{u}_i}_{n \times 1} \underbrace{\underline{v}_i^T}_{1 \times d}$$

pseudo inverse

$$\underline{X}^+ = \sum_{i=1}^{d} \frac{1}{\sigma_i} \underline{v}_i \underline{u}_i^T$$

$d \times 1 \quad 1 \times n$

$$\underline{u}_i^T \underline{u}_j = \mathbb{1}[i = j] = \underline{v}_i^T \underline{v}_j$$

$$\underline{X}^T \underline{X} = \sum_{i=1}^{d} \underline{v}_i \sigma_i^2 \underline{v}_i^T$$

$$(\underline{X}^T \underline{X})^+ \underline{X}^T \underline{y} = \sum_{i=1}^{d} \underline{v}_i \frac{1}{\sigma_i^2} \underline{v}_i^T \sum_{j=1}^{d} \sigma_j \underline{v}_j \underline{u}_j^T \underline{y}$$

$$= \sum_{i=1}^{d} \underline{v}_i \frac{1}{\sigma_i} \underline{u}_j^T \underline{y}$$

$$= \underline{X}^+ \underline{y}$$

Q: Why squared loss?

Before:

↳ differentiable

↳ penalize far, more

Now:

Q2: Why use linear model?

$y = \langle x, w^* \rangle + \eta$ (w^*: unknown)

where $\eta \sim \mathcal{N}(0, \sigma^2)$ (0: mean; σ^2: unknown variance)

~ Central Limit Theorem ~

Equivalently,

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

ERM says find model that most likely generated data ...

$\Pr(y \text{ drawn from model } w)$

$= \Pr(y \sim \mathcal{N}(\langle x, w \rangle, \sigma^2))$

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

$\arg\max_w \Pr(y^{(1)} \sim \mathcal{N}(\langle x^{(1)}, w \rangle, \sigma^2), \ldots, y^{(n)} \sim \mathcal{N}(\langle x^{(n)}, w \rangle, \sigma^2))$

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

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

$$= \arg\max_w \log(\cdot)$$

$$= \arg\max_w \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) + \sum_{i=1}^{n} \log\left(\exp\left(-\frac{(y^{(i)} - \langle x^{(i)}, w\rangle)^2}{2\sigma^2}\right)\right)$$

$$= \arg\max_w \sum_{i=1}^{n} -(y^{(i)} - \langle x^{(i)}, w\rangle)^2$$

$$= \arg\min_w \sum_{i=1}^{n} (y^{(i)} - \langle x^{(i)}, w\rangle)^2$$

$$= \arg\min_w \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \langle x^{(i)}, w\rangle)^2$$