Week 2

9/2/2025

↳ Quiz!

↳ How many read the notes? Important!

↳ Encourage practice, problem sets due every week

- First 2 problems still 9/19
- In general, due Mondays 11:59 pm

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

↳ Post questions, come to office hours :)

Last week was warm up for machine learning!

Review

What about when M is not square?

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

$$M = \sum_{i=1}^{r} \underline{v}_i \lambda_i \underline{w}_i^T$$

Use: Page Rank where we powered up matrix multiplication!

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

$\underline{v}_1, \ldots, \underline{v}_r \in \mathbb{R}^d$ right eigenvectors $\quad A \underline{v}_i = \lambda_i \underline{v}_i$

$\underline{w}_1, \ldots, \underline{w}_r \in \mathbb{R}^d$ left eigenvectors $\quad \underline{w}_i^T A = \lambda_i \underline{w}_i^T$

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

$$\underline{w}_j^T A \underline{v}_i = \underline{w}_j^T \lambda_i \underline{v}_i$$

$$\lambda_j \underline{w}_j^T \underline{v}_i = \lambda_i \underline{w}_j^T \underline{v}_i$$

$$(\lambda_j - \lambda_i) \underline{w}_j^T \underline{v}_i = 0$$

$\Leftrightarrow$

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

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 $f: \mathbb{R}^d \to \mathbb{R}$

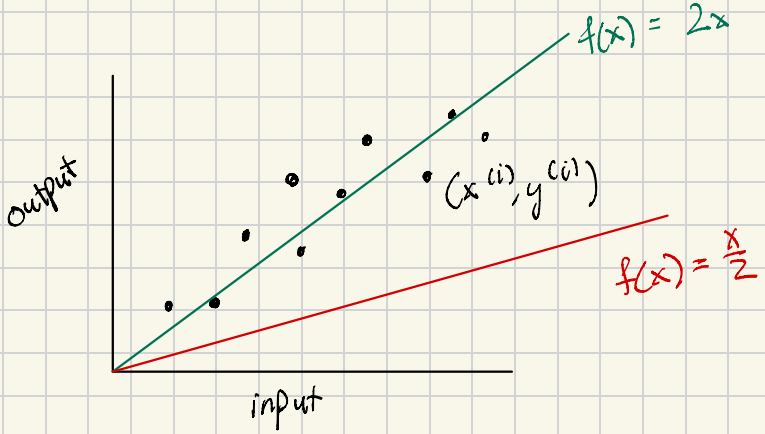
so that $f(x^{(i)}) \approx y^{(i)} \quad \forall\, i$

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: $f(x) - y$

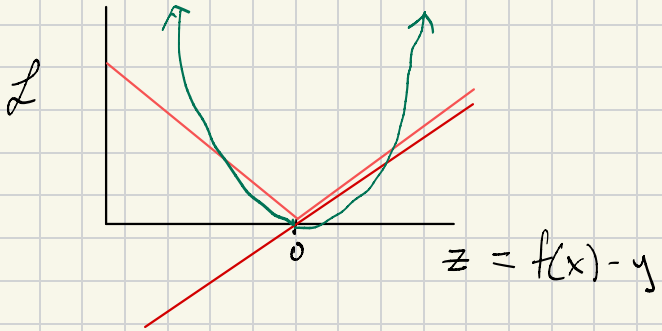
☹ $y > f(x) \Rightarrow$ very negative

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

☹ not differentiable at 0

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

☺ far gets penalized more



Mean Squared Error (MSE) Loss

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

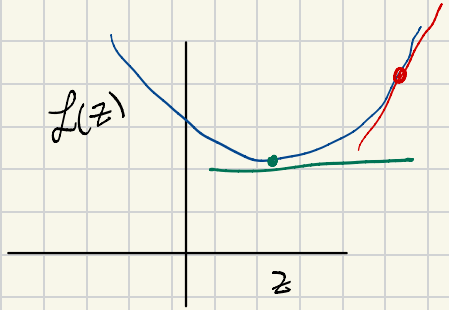
③ Exact Optimization

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

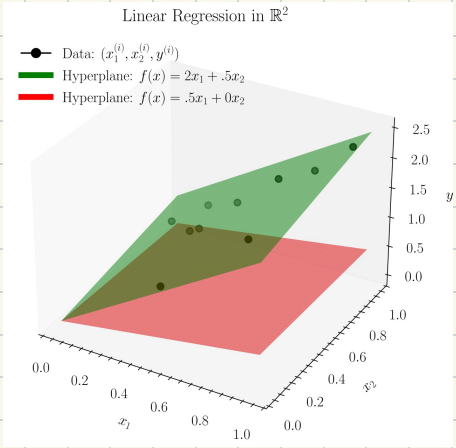
Multivariate Linear Regression

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

(1) Function Class

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

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



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

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

(2) MSE

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

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

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

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

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

③ Optimizer

Key insight:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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