Week 02 - Part 2

Review:

- Supervised Learning
  - discrete y_n: Classification
  - Continuous y_n: Regression

Today:

- We study a specific type of regression.

  Linear Regression

- Least squares Solution.

Linear Regression

Training Set: $\mathcal{D} = \{(\underline{x}_n, y_n)\}_{n=1}^{N}$ $\quad \underline{x}_n \in \mathbb{R}^d \quad y_n \in \mathbb{R}$

Decision Rule ("Hypothesis Set", "Learning Model"):

$$h(\underline{w}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

Define the augmented form. It makes life easier!

$$\underline{x} = (x_0 = 1, x_1, x_2, \ldots, x_d) \in \{1\} \times \mathbb{R}^d$$

$$h_{\underline{w}}(\underline{x}) = \underline{w}^T \underline{x}$$

Criterion:

$$E_{in}(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{(y_n - \hat{y}_n)}_{e_n(\underline{w}):\ \text{error for } n_{th} \text{ Sample}}{}^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \underline{w}^T \underline{x}_n)^2$$

Average squared error

Goal: Give $\mathcal{D}$, find $\underline{w}$ that minimizes $E_{in}(\underline{w})$

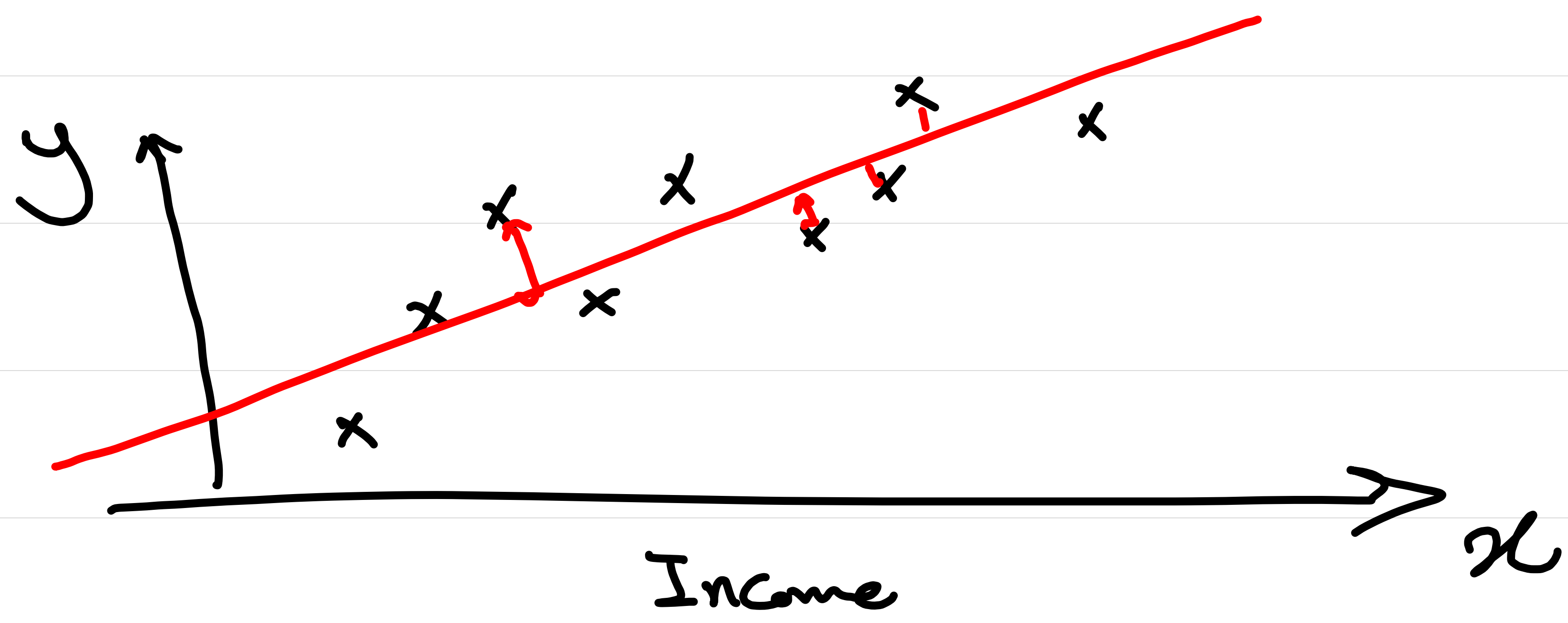
E.g.: The bank wants to set a proper credit limit for each customer.

x = customer's income

y = credit limit

Historical Data:

$$D = \{(x_n, y_n)\}_{n=1}^{N}$$



Fit a linear model

$$\hat{y} = w_0 + w_1 x$$

In reality: $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} \text{Income} \\ \text{age} \\ \text{years of experience} \\ \vdots \end{bmatrix}$

$\hat{y} = w_0 + w_1 x_1 + \dots + w_d x_d$ $\rightsquigarrow$ larger w_i, more important factor in assigning credit limit

Matrix-Vector Algebraic Representation

1) Data matrix:

$$X = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix} \in \mathbb{R}^{N \times (d+1)}$$

2) Target vector:

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$$

3) Weight vector: $\underline{w} = (w_0, \dots, w_d) \in \mathbb{R}^{d+1}$

4) Model:

$$\hat{\underline{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \underline{w}^T \underline{x}_1 \\ \underline{w}^T \underline{x}_2 \\ \vdots \\ \underline{w}^T \underline{x}_N \end{bmatrix} = \begin{bmatrix} \underline{x}_1^T \underline{w} \\ \underline{x}_2^T \underline{w} \\ \vdots \\ \underline{x}_N^T \underline{w} \end{bmatrix} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix} \underline{w} = X \underline{w}$$

5) Error : $E_{in}(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$

$$= \frac{1}{N} \left\| \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix} \right\|^2 = \frac{1}{N} \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} \right\|^2 = \frac{1}{N} \left\| \underline{y} - \underline{\hat{y}} \right\|^2$$

$$= \frac{1}{N} \left\| \underline{y} - \underline{X} \underline{w} \right\|^2 .$$

Remark: $\underline{P} = \begin{bmatrix} P_1 \\ \vdots \\ P_K \end{bmatrix}$, $\|\underline{P}\| = \sqrt{P_1^2 + \cdots + P_K^2}$. Hence, $P_1^2 + \cdots + P_K^2 = \|\underline{P}\|^2$.

Q) When is $E_{in}(\underline{w}) = 0$? for all datapoints, $y_n = \hat{y}_n = \underline{w}^T \underline{x}_n$

$$\begin{cases} y_1 = \underline{w}^T \underline{x}_1 \\ y_2 = \underline{w}^T \underline{x}_2 \\ \vdots \\ y_N = \underline{w}^T \underline{x}_N \end{cases}$$

This is a system of linear equation that we have to solve to get "pefect" $\underline{w}$.

\# of linear equations : N

\# of unknown parameters: $d+1$

- In practice, $N \gg d+1 \longrightarrow$ No solution
  "overdetermind" system of equation.

- Instead, we find a $\underline{w}$ that minimizes $E_{in}(\underline{w})$
  - The algorithm that we use is called least squares method.

- Want to minimize $E_{in}(\underline{w}) = \frac{1}{N} \| \underline{y} - \underline{\hat{y}} \|^2$

- define $f(\underline{w}) = \| \underline{y} - \underline{\hat{y}} \|^2 = \| \underline{y} - \underline{X}\, \underline{w} \|^2 = \sum_{n=1}^{N} \left(y_n - \underline{w}^T \underline{x}_n \right)^2$

$$= \sum_{n=1}^{N} \left(y_n - (w_0 + w_1 x_{n1} + w_2 x_{n2} + \dots + w_d x_{nd}) \right)^2$$

- This is a multivariate function.

- To minimize this, we need gradients.
  - Just like setting derivative to zero for univariate functions, we need to find a $\underline{w}$ for which the derivative w.r.t. all cordinates are zero.

Detour: Gradient Reminder

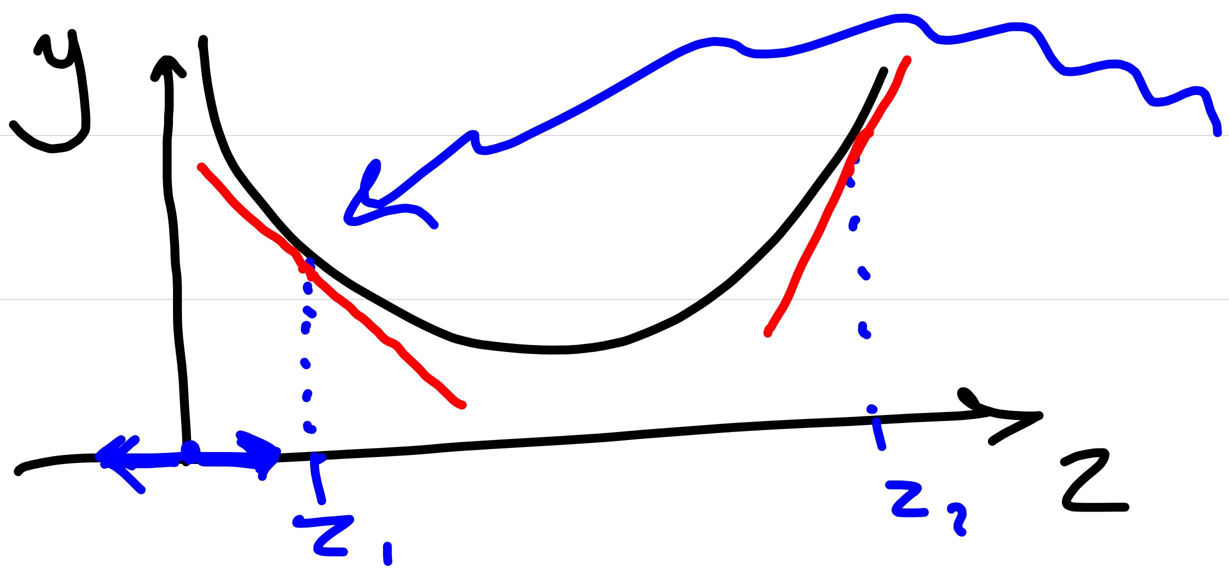
- Gradient of $g(\underline{z})$ w.r.t $\underline{z}$ is denoted by $\nabla_{\underline{z}}\, g(\underline{z})$ and defined as

$$\nabla_{\underline{z}}\, g(\underline{z}) = \begin{bmatrix} \partial g(\underline{z}) / \partial \underline{z}_1 \\ \partial g(\underline{z}) / \partial \underline{z}_2 \\ \vdots \\ \partial g(\underline{z}) / \partial \underline{z}_d \end{bmatrix}$$

dimensionality of $\nabla_{\underline{z}}$ is the same as that of $\underline{z}$.

- Similar to derivative, gradient points in the direction of steepest increase.

- Let's see a $d=1$ example



A negative derivative at z_1 indicates that the steepest increase direction is to the left

Detour: Basic Gradients Everyone must Know

- $\nabla_{\underline{w}}(\underline{w}^T \underline{x}_n) = \nabla_{\underline{w}}\left(\sum_{i=0}^{d} w_i x_{ni}\right)$

$$= \begin{bmatrix} \partial\left(\sum_{i=0}^{d} w_i x_{ni}\right)/\partial w_0 \\ \partial\left(\sum_{i=0}^{d} w_i x_{ni}\right)/\partial w_1 \\ \vdots \\ \partial\left(\sum_{i=0}^{d} w_i x_{ni}\right)/\partial w_d \end{bmatrix} = \begin{bmatrix} x_{n0} \\ x_{n1} \\ \vdots \\ x_{nd} \end{bmatrix} = \underline{x}_n$$

- $\nabla_{\underline{w}}(\underline{x}_n^T \underline{w}) = \underline{x}_n$

- $\nabla_{\underline{w}}(\underline{w}^T A \underline{w}) = 2A\underline{w}$

  → A is symmetric

- $\|\underline{a}\|^2 = \underline{a}^T \underline{a}$

- Let's get back to the problem we had
- We want to find the minimum of

$$\|\underline{y} - \underline{\hat{y}}\|^2 = \|\underline{y} - X\underline{w}\|^2 = f(\underline{w})$$

- Hence, we most find a $\underline{w}$ such that $\nabla_{\underline{w}} f(\underline{w}) = 0$
- Let's find $\nabla_{\underline{w}} f(\underline{w}) = \nabla_{\underline{w}} \|\underline{y} - X\underline{w}\|^2 = \nabla_{\underline{w}} \left((\underline{y} - X\underline{w})^T (\underline{y} - X\underline{w}) \right)$

$$= \nabla_{\underline{w}} \left((\underline{y}^T - \underline{w}^T X^T)(\underline{y} - X\underline{w}) \right) = \nabla_{\underline{w}} \left(\underline{y}^T \underline{y} - \underline{y}^T X \underline{w} - \underline{w}^T X^T \underline{y} + \underline{w}^T X^T X \underline{w} \right)$$

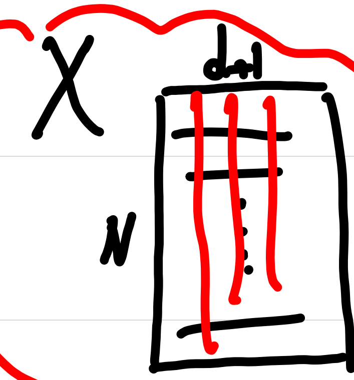
$$= 0 - X^T \underline{y} - X^T \underline{y} + 2X^T X \underline{w} = 2X^T (X\underline{w} - \underline{y})$$

Remark: $(AB)^T = B^T A^T$

Least square Solution

- The least square solution, $\underline{w}_{ls}$, is the weight vector such that $\nabla_{\underline{w}} f(\underline{w}_{ls}) = \underline{0}$.

Recall:



- Thus, $2X^T(X\underline{w}_{ls} - \underline{y}) = 0 \Rightarrow X^TX\underline{w}_{ls} = X^T\underline{y}$.

- Finding $\underline{w}_{ls}$ would have been so simple if we could multiply the two sides by $(X^TX)^{-1}$. But, what if X^TX is not invertible?

  - A reasonable assumption is X^TX is invertible, i.e., there are $(d+1)$ rows of X (i.e. $d+1$ datapoints) that are linearly independent.

- With this simplifying assumption, we have that $X^TX\underline{w}_{ls} = X^T\underline{y} \Rightarrow$

$$(X^TX)^{-1}(X^TX\underline{w}_{ls}) = (X^TX)^{-1}X^T\underline{y} \Rightarrow (X^TX)^{-1}(X^TX)\underline{w}_{ls} = (X^TX)^{-1}X^T\underline{y}$$

$$\Rightarrow I\,\underline{w}_{ls} = (X^TX)^{-1}X^T\underline{y} \Longrightarrow \underline{w}_{ls} = (X^TX)^{-1}X^T\underline{y}$$

- $\mathrm{Rank}(X) = d+1 \iff X^T X$ is invertible

- With that (reasonable) assumption, $\underline{w}_{LS} = \underbrace{(X^T X)^{-1} X^T}_{\text{(pseudo\_inverse of } X)} \underline{y}$

  - $X^{+} = (X^T X)^{-1} X^T$

  - $\underline{w}_{LS} = X^{+} \underline{y}$

Why is $X^+ = (X^T X)^{-1} X^T$ called pseudo-invers of X?

① Observe that $X^+ X = (X^T X)^{-1} X^T X = I$

But, $X X^+ = X (X^T X)^{-1} X^T \neq I$

Why is $X^+ = (X^T X)^{-1} X^T$ called Pseudo-invers of X ?

② Recall: Originally we had the system of equations $\underline{y} = X \underline{w}$ and wanted to solve it.

- To solve this equation system, we must find inverse of X so that $X^{-1} \underline{y} = X^{-1} X \underline{w} = I \underline{w} = \underline{w}$.

- But X is not invertible (It's not even a square matrix. Inverse is for square matrix)

- However, X^+ would do the trick: from ①

$$\underline{y} = X \underline{w} \Longrightarrow X^+ \underline{y} = X^+ X \underline{w} = I \underline{w} = \underline{w}$$

Summary:

- Least square solution: $\underline{w}_{ls} = (X^T X)^{-1} X^T \underline{y}$

- prediction by $\underline{w}_{ls}$: $\underline{\hat{y}}_{ls} = X \underline{w}_{ls} = X (X^T X)^{-1} X^T \underline{y}$