Week 02 - Part 2

Review:

- Supervised Learning
  - discrete y_n:
  - Continuous y_n:

Today:

- We study a specific type of regression.

  Linear Regression

- Least squares Solution.

Linear Regression

Training Set:

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

Define the augmented form. It makes life easier!

Criterion:

Goal:

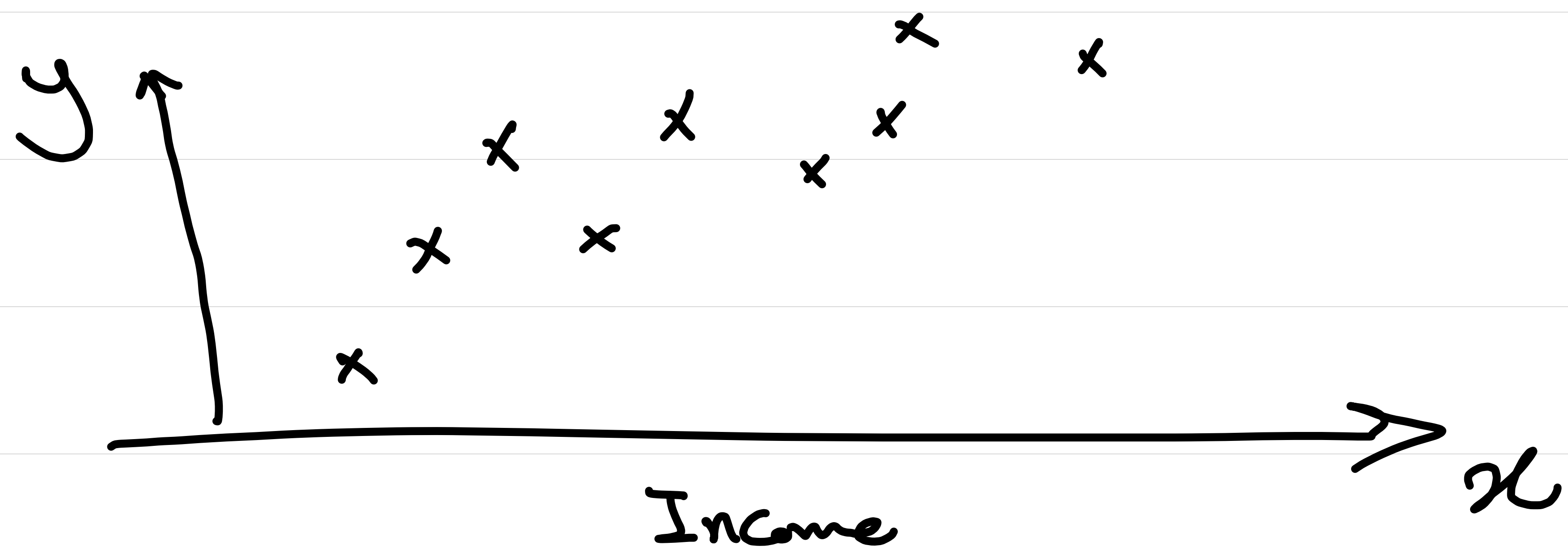
E.g.: The bank wants to set a proper credit limit for each customer.

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:

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

2) Target vector:

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

3) Weight vector:

4) Model:

5) Error :

Q) When is $E_{in}(\underline{w}) = 0$?

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

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

- 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

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

- Similar to derivative, gradient points in the direction of steepest increase.
- Let's see a $d=1$ example

Detour: Basic Gradients Everyone must Know

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

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

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

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

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

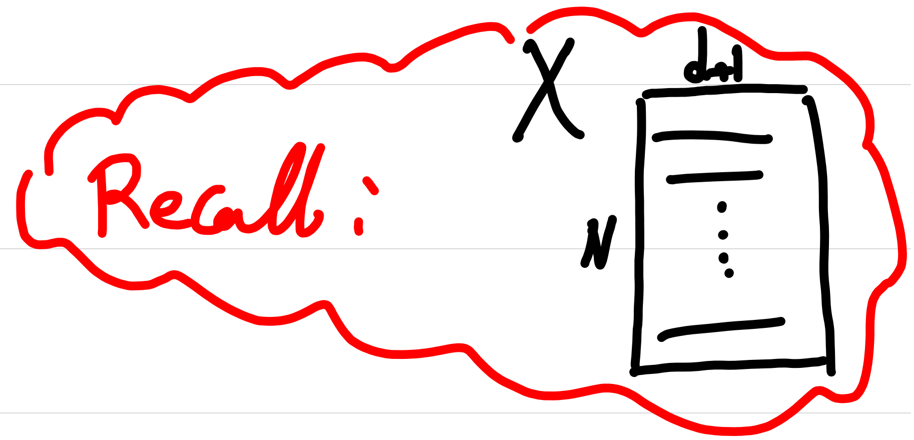
- Let's get back to the problem we had
- We want to find the minimum of

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

- Hence, we most find a $\underline{w}$ such that $\nabla_{\underline{w}} f(\underline{w}) = 0$
- Let's find $\nabla_{\underline{w}} f(\underline{w})$:

Least square Solution

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



- Thus,

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

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

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

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

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

1. Observe that $X^+ X =$

But, $X X^+ =$

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:

Summary:

- Least square solution: $\underline{w}_{ls} =$

- Prediction by $\underline{w}_{ls}$: $\hat{y}_{ls} =$