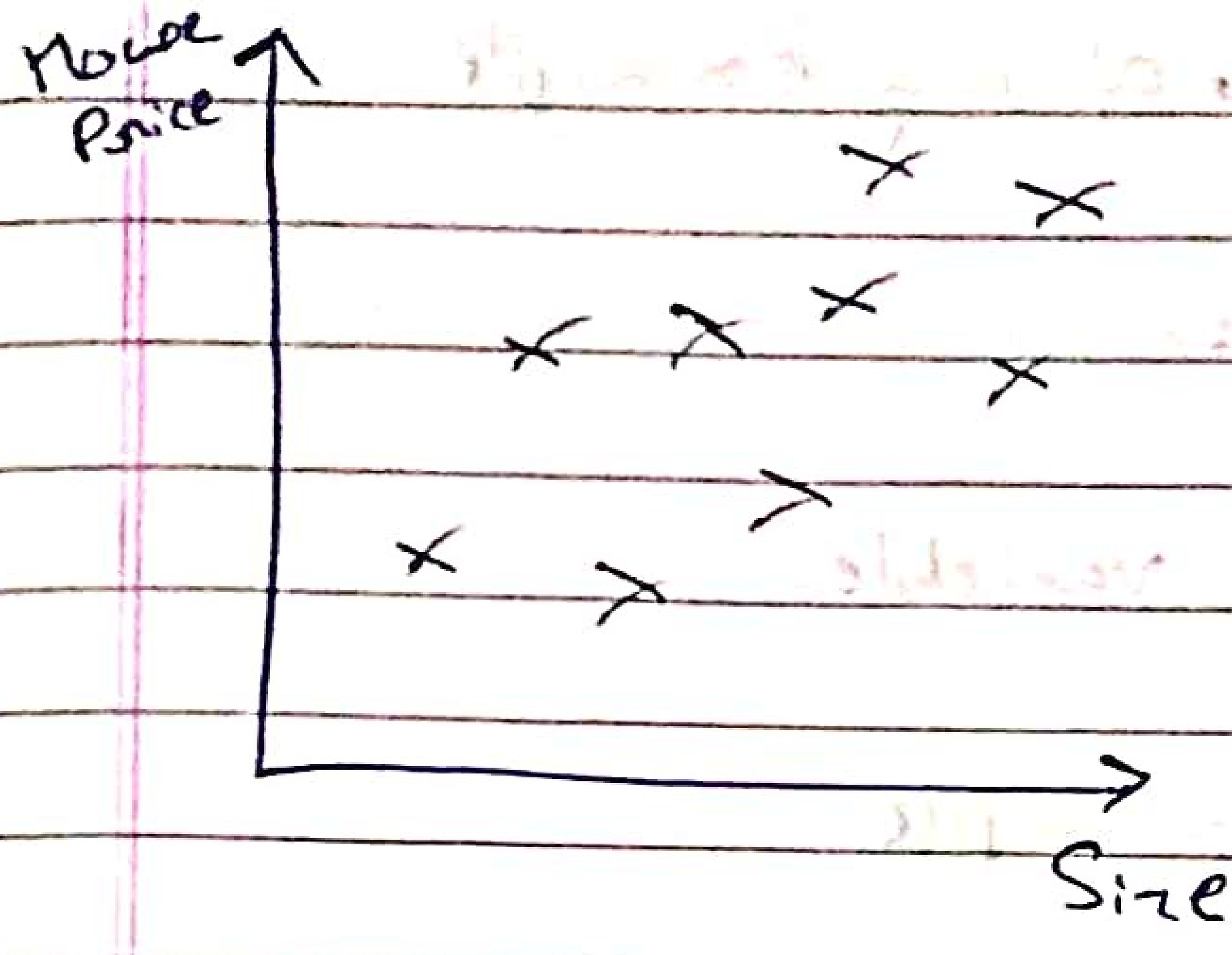
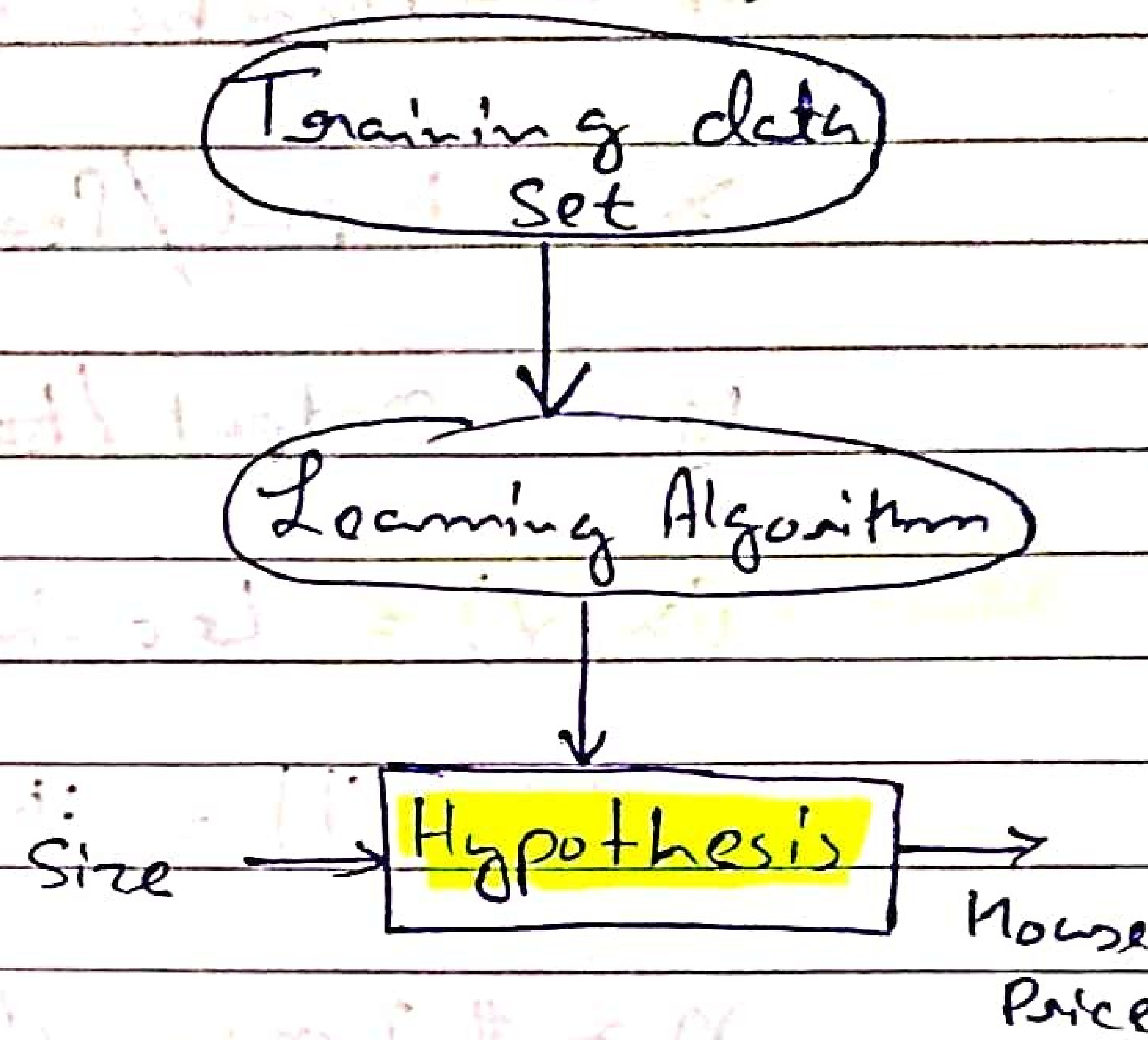
2

Linear Regression





⇒ How to represent h?

$$h(x) = \theta_0 + \theta_1 X$$ {for Linear regression}

⇒ If more than 1 input features:

{Technically it is not linear, it is affine function}

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h(x) = \sum_{j=0}^{2} \theta_j x_j \quad \text{where } x_0 = 1$$

Let $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$ $\qquad x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$

Parameters

$$\boxed{h(x) = \theta^T X}$$

Date

Page

m = # Number of training example

x = input / features

y = Output / target variable

(x, y) = training example

$(x^{(i)}, y^{(i)})$ = i^{th} training example

n = # Number of features.

⇒ Choose θ such $h(x) \approx y$ $\forall$ training example

$h_\theta(x) \equiv h(x)$

⇒ Linear regression $\cancel{J = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$

$$\boxed{J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$$

Cost function

Objective: $\min_{\theta} J(\theta)$

Date
Page

★ Gradient descent (Batch)

⇒ Start with some θ

⇒ Keep changing θ to reduce $J(\theta)$

$$\theta_j := \theta_j - \alpha \frac{\delta}{\delta \theta_j} J(\theta)$$ → Learning Rate

~~$\theta_j - \alpha (h_\theta(x) - y) x_j$~~

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

⇒ Repete until convergence for all $j = 0, 1 \dots n$

⇒ Batch gradient descent

★ Stochastic gradient descent

Repete {

for $i = 1$ to m {

$$\theta := \theta - \alpha (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

}

}

⇒ Taking one training example at a time.

⇒ Works well with large dataset.

Date
Page

★ <u>Normal equation</u> (Least Square Solution)

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)}\right)^2$$

Let

$$X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(m)T} \end{bmatrix} \quad \& \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

$(m \times n+1)$, $(n+1 \times 1)$

⇒ Then,

$$\begin{bmatrix} h(x^{(1)}) \\ h(x^{(2)}) \\ \vdots \\ h(x^{(m)}) \end{bmatrix} = X\theta$$

⇒ Let

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

⇒ Then,

$$J(\theta) = \frac{1}{2}(X\theta - y)^T (X\theta - y)$$

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - y)^T (X\theta - y)$$

$$= \frac{\nabla_\theta}{2}\left(\theta^T X^T X \theta - 2 y^T X \theta + y^T y\right)$$

$$= X^T X \theta - X^T y = 0$$

$$\Rightarrow \boxed{\theta = (X^T X)^{-1} X^T y}$$