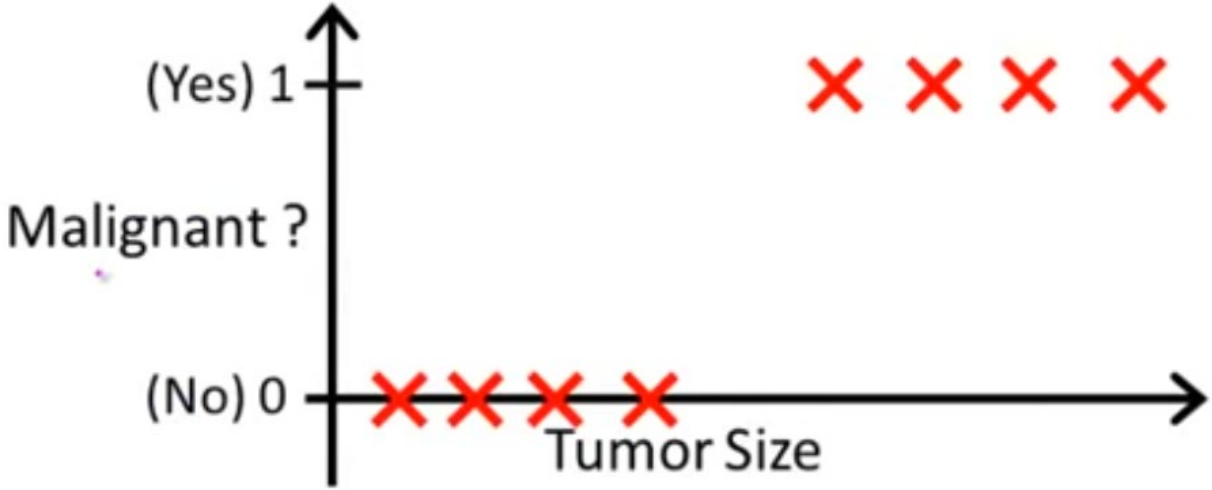
Introduction to Machine Learning

Logistic Regression

A characterization of machine learning problems

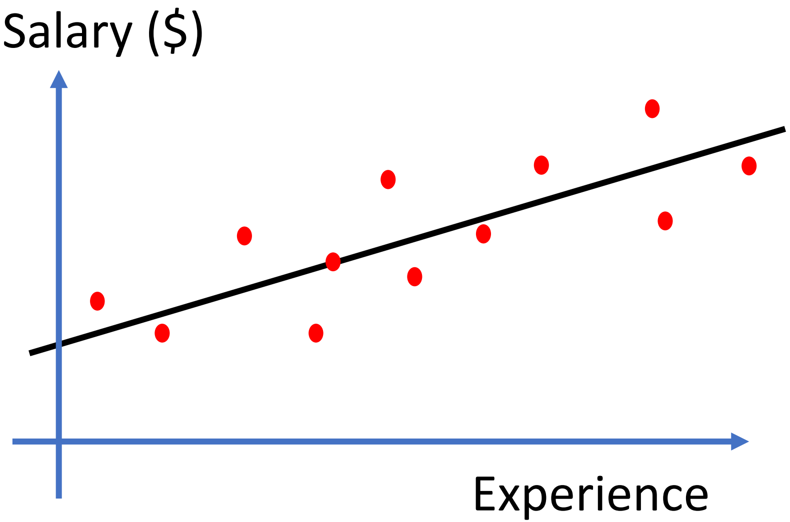
	Supervised	*Unsupervised*
Discrete	Classification *Logistic Regression*	Clustering
Continuous	Regression	Dimensionality reduction

Why linear regression may not make a good classifier

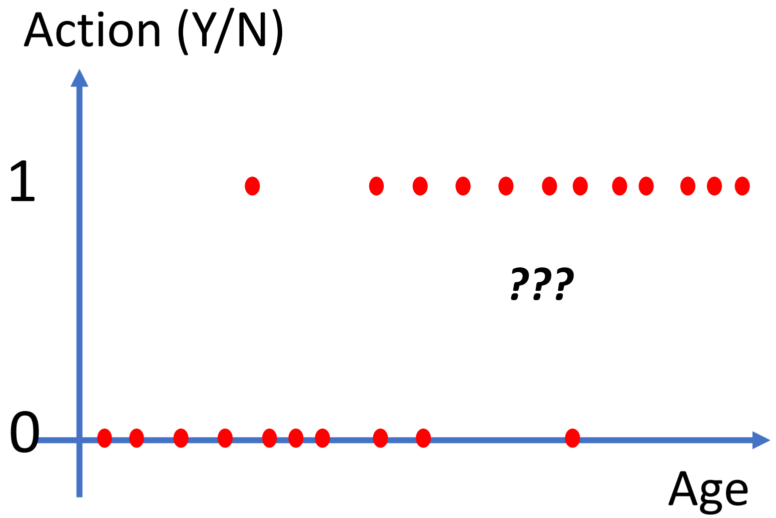


Another example

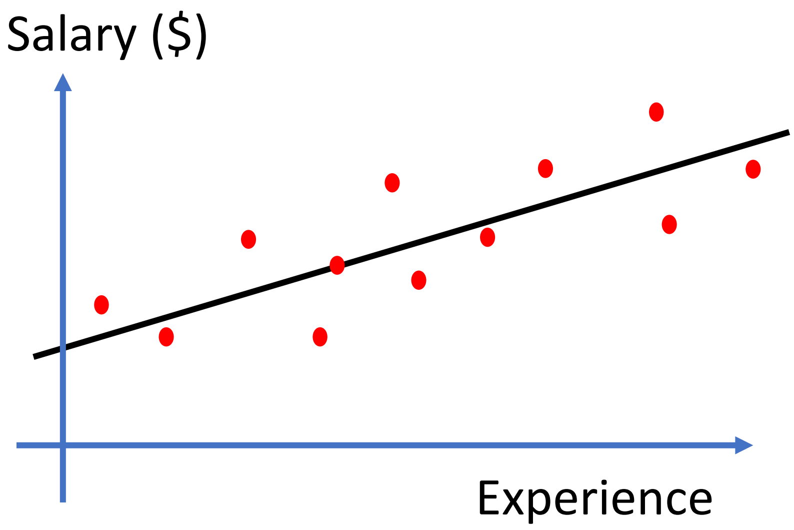
We know this:

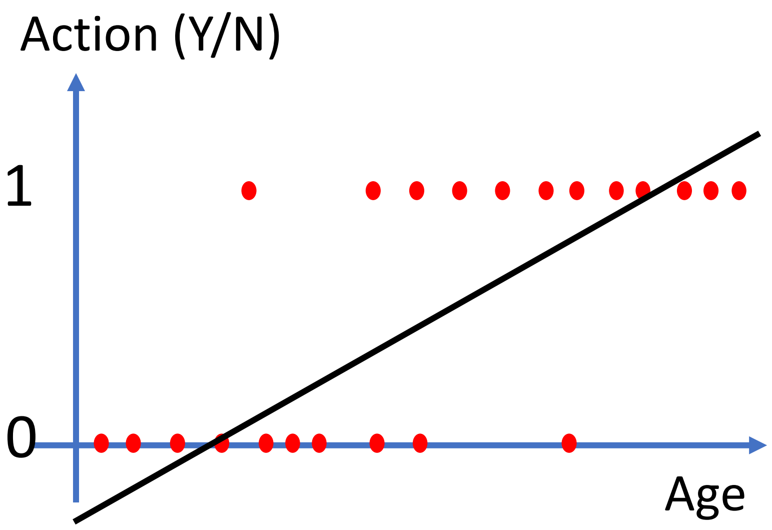


This is new:

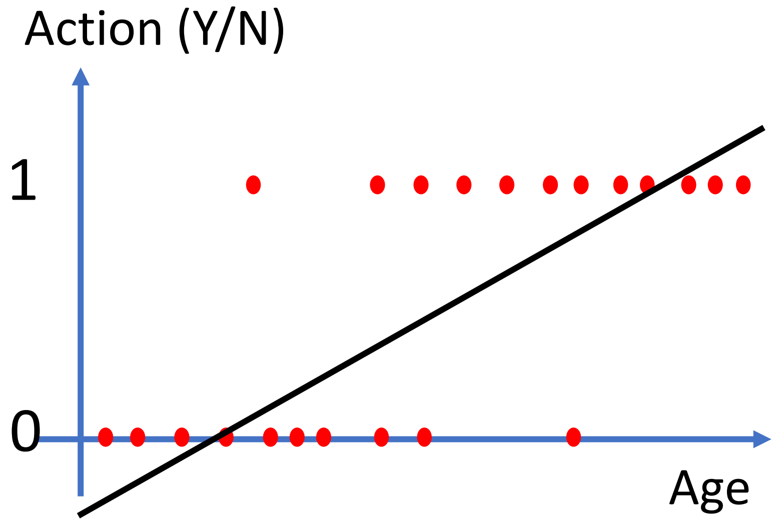


Another example

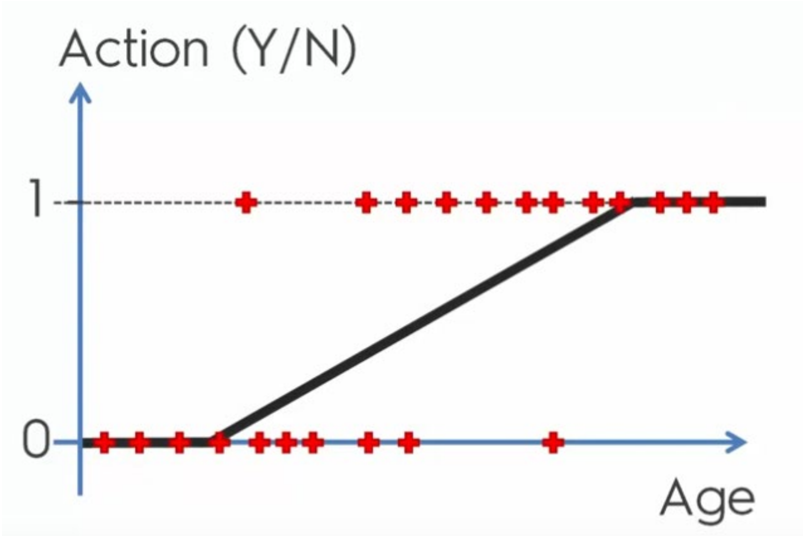


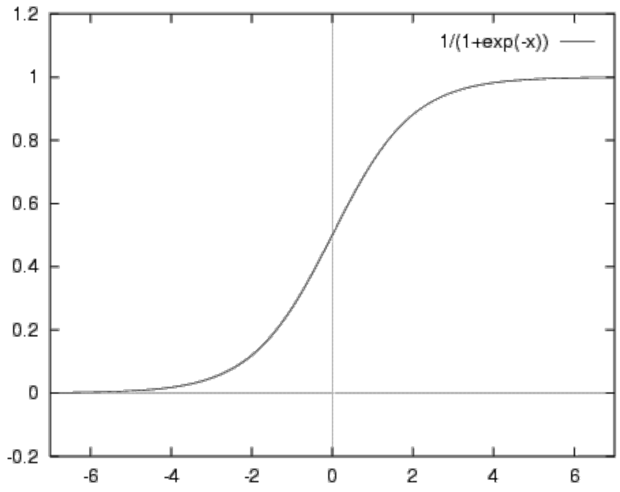


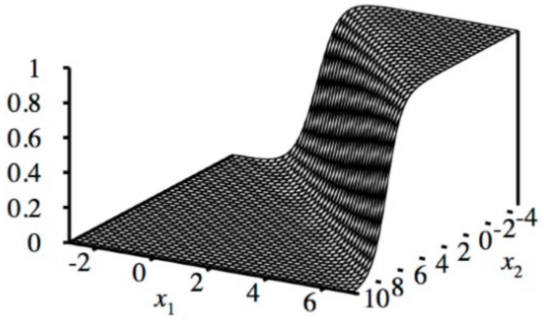
Another example

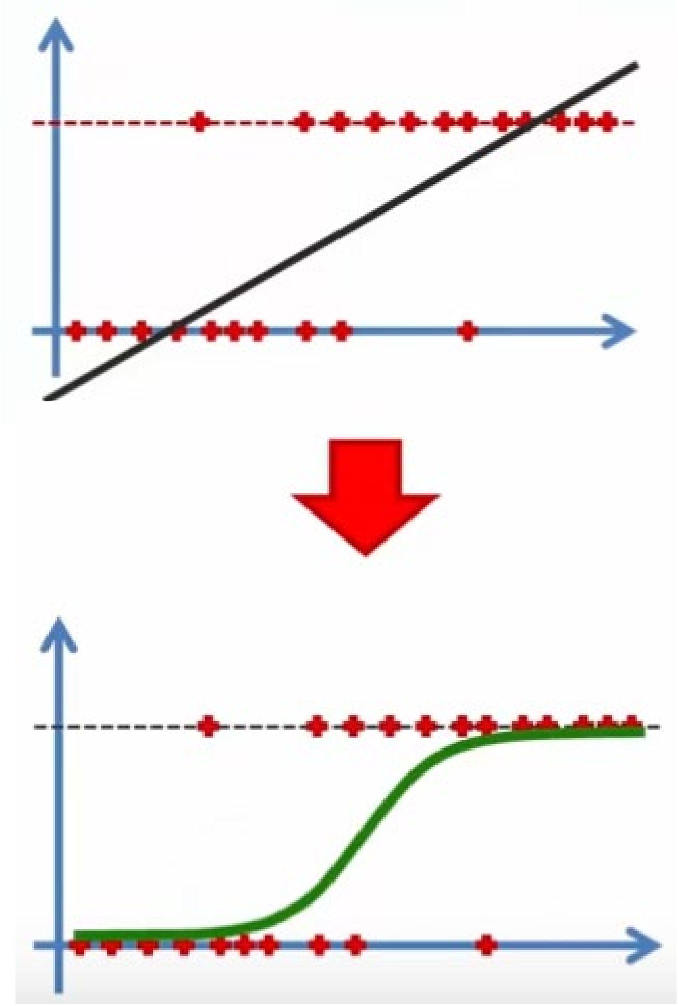


Another example









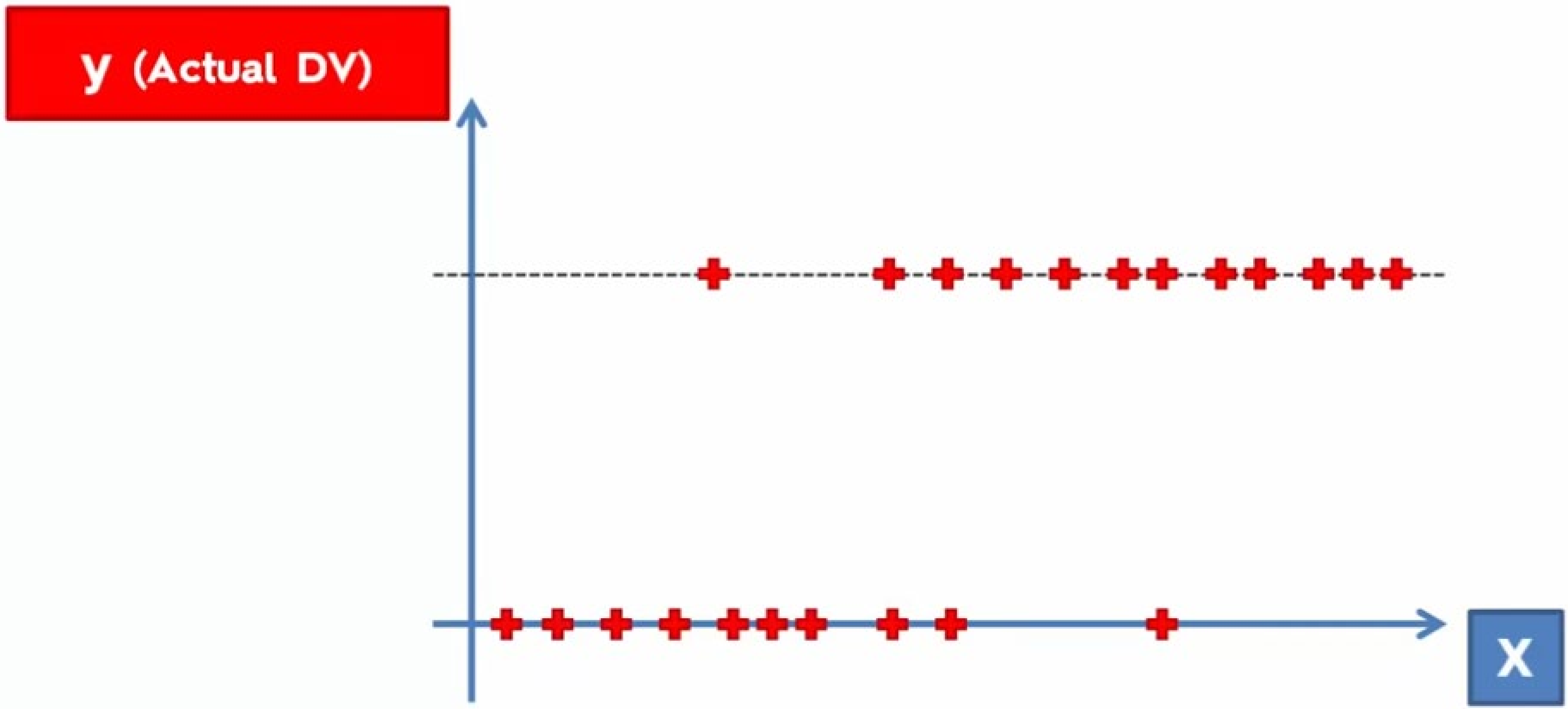
$$y = \theta_0 + \theta_1 x$$

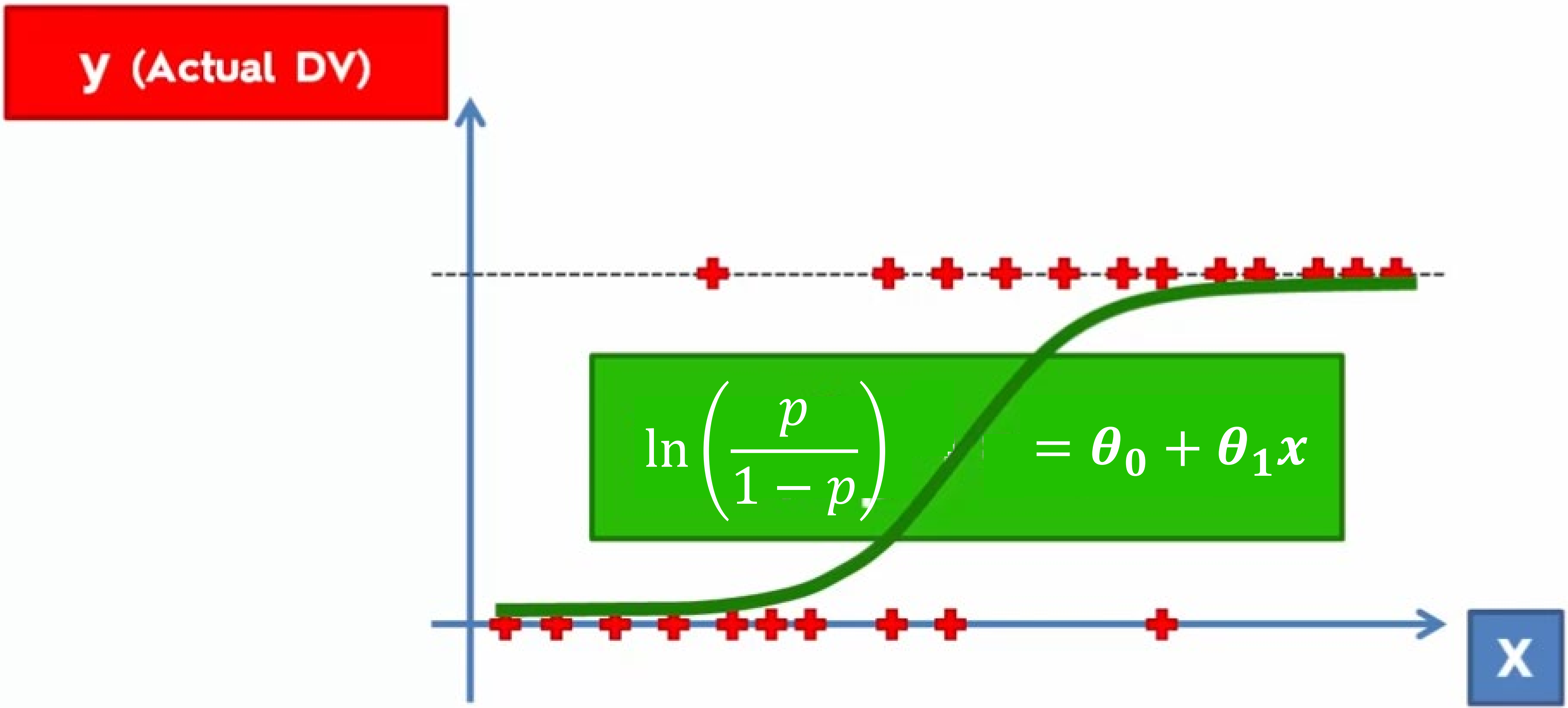
Sigmoid Function

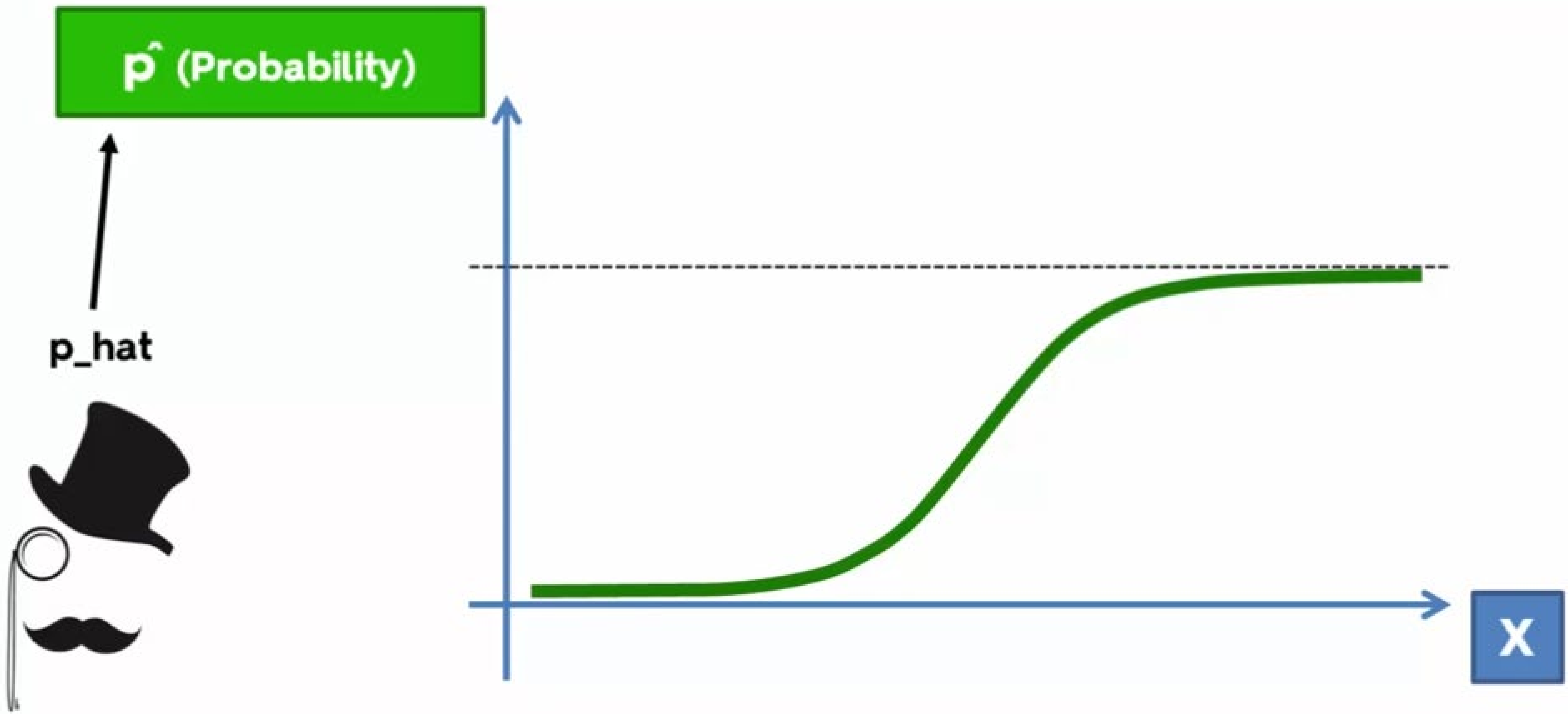
$$p = \frac{1}{1 + e^{-y}}$$

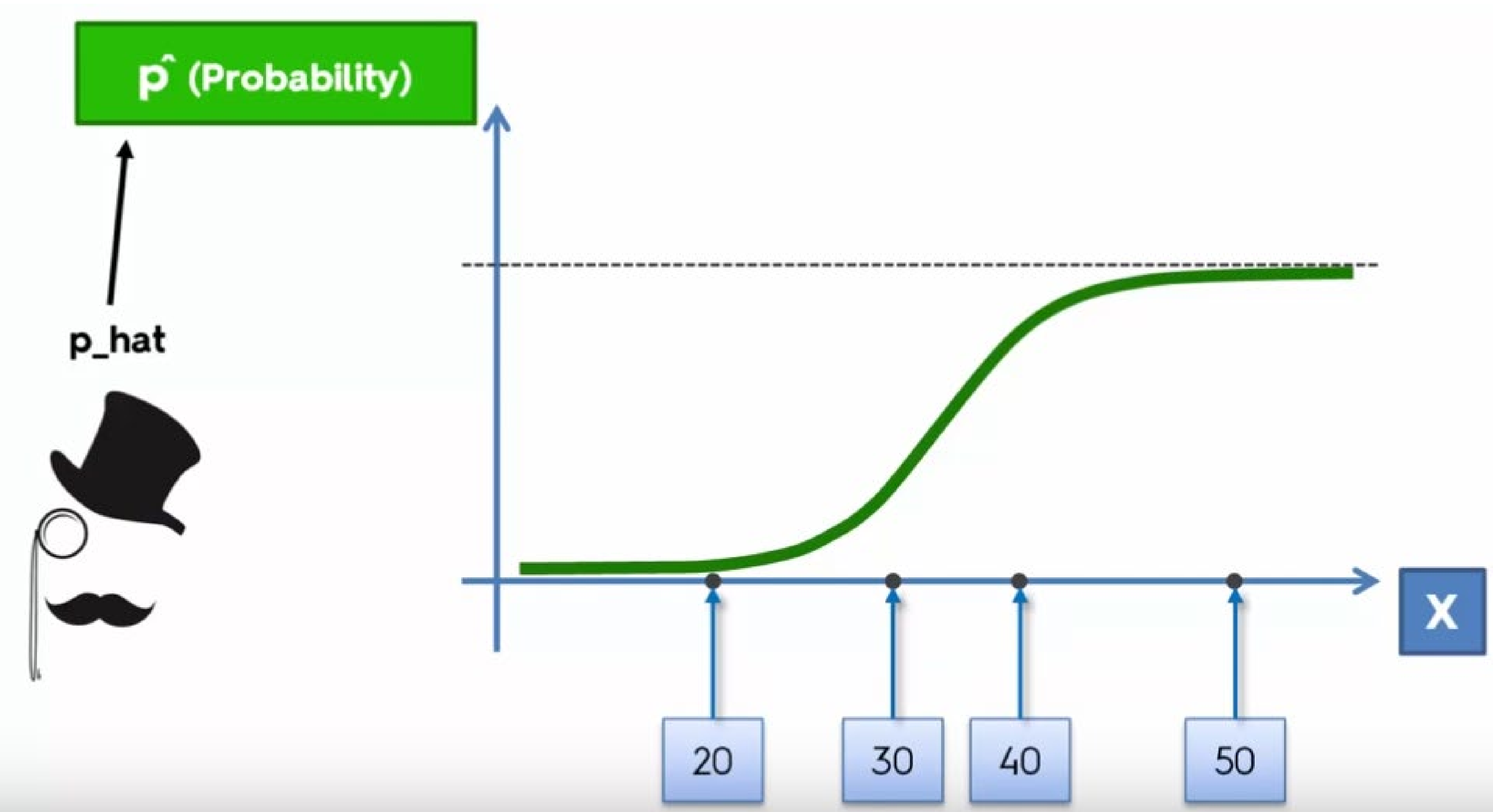
$$\ln\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x$$

WHAT JUST HAPPENED ???

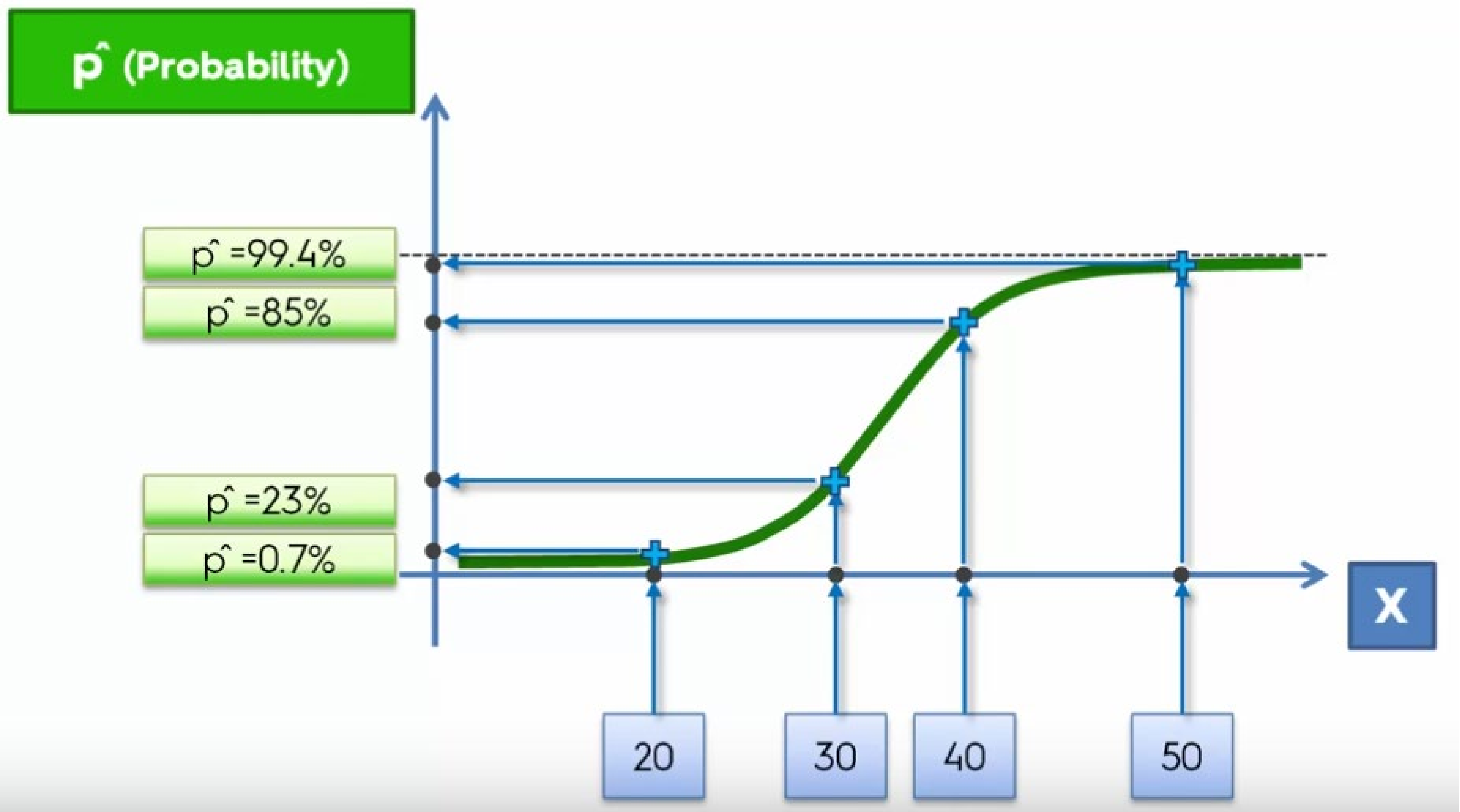

y (Actual DV)
X


y (Actual DV)
$\ln\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x$
x

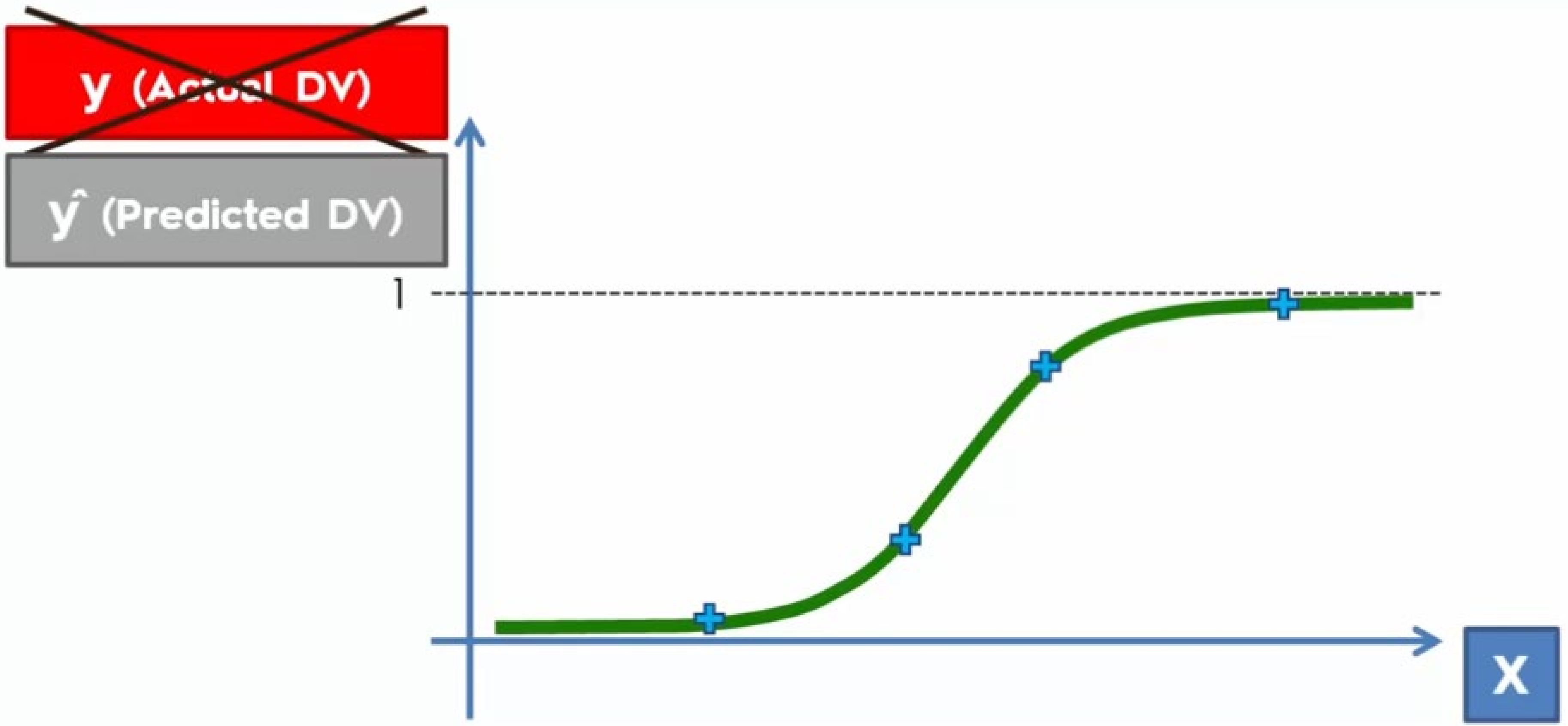

p̂ (Probability)
p_hat
X



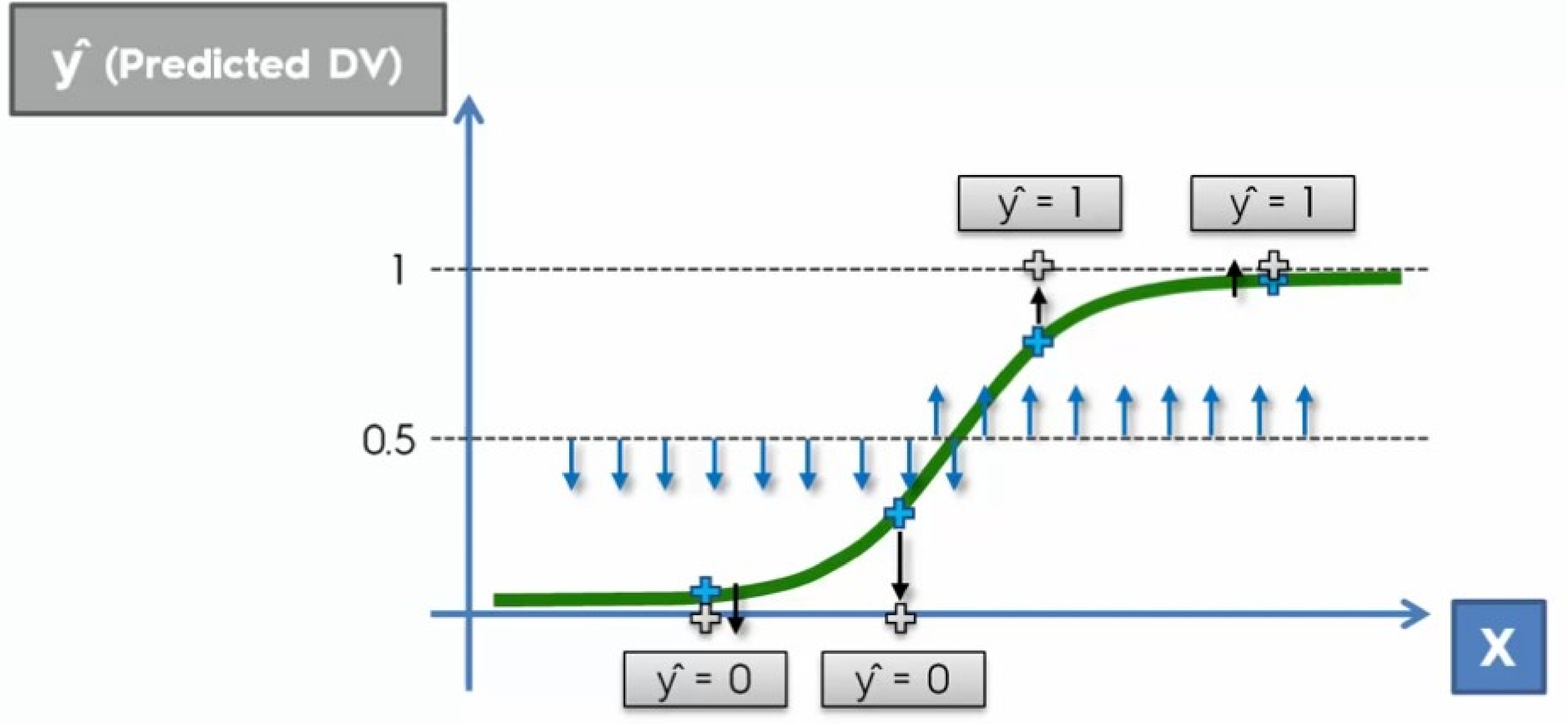
p̂ (Probability)
p_hat
x
20
30
40
50



p̂ (Probability)
p̂ =99.4%
p̂ =85%
p̂ =23%
p̂ =0.7%
X
20
30
40
50



y (Actual DV)
ŷ (Predicted DV)
1
X

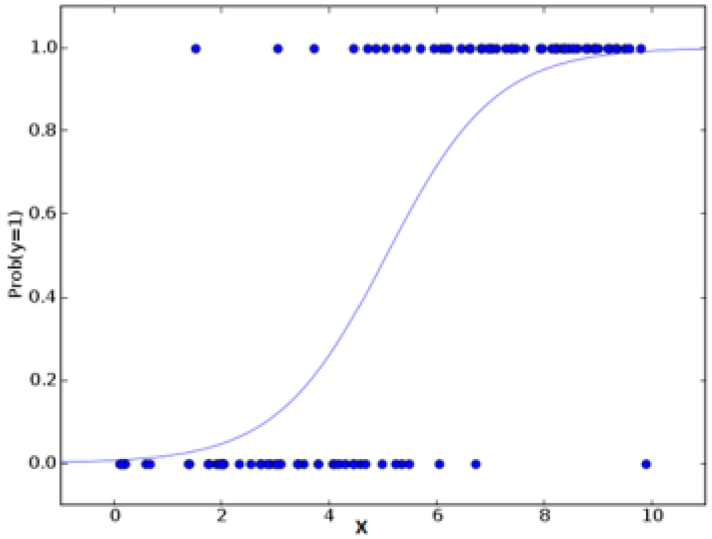

ŷ (Predicted DV)
ŷ = 1
ŷ = 1
1
0.5
ŷ = 0
ŷ = 0
X

Estimating parameters

- Calculate prediction based on coefficients
- Adjust coefficients based on prediction error (loss function)

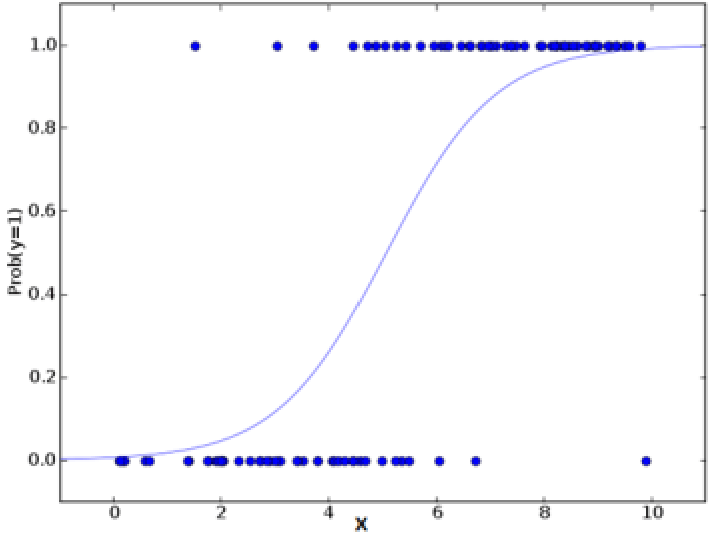
Sigmoid function

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$



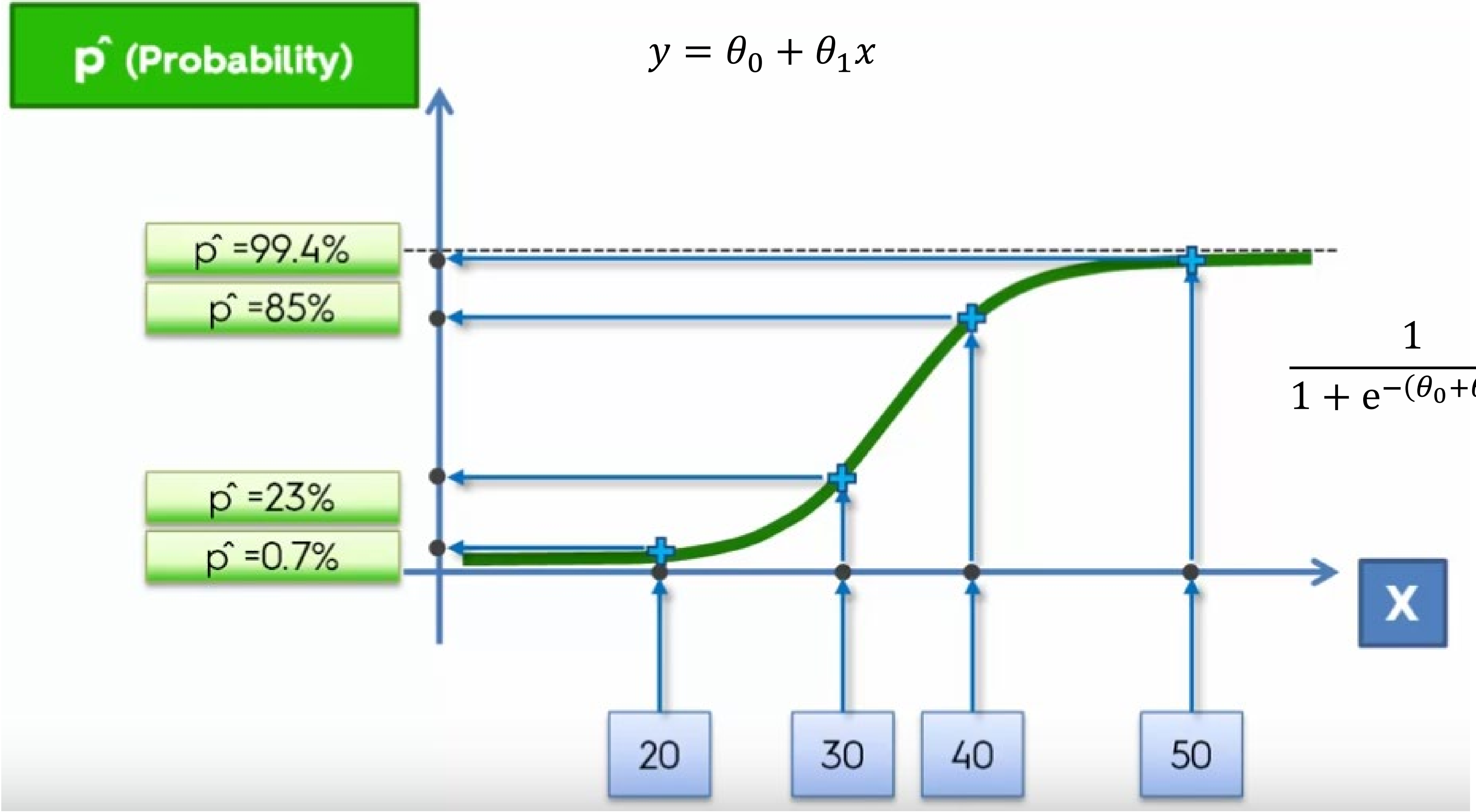
Sigmoid function

$$y = \theta_0 + \theta_1 x$$



$$\frac{1}{1 + \mathrm{e}^{-(\theta_0 + \theta_1 x)}}$$

$$y = \theta_0 + \theta_1 x$$



$$\frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

Linear Regression Cost Function

- $J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(h(x_i) - y_i)^2$

Simplify Logistic Regression Cost Function

- $Cost(h(x), y) = \begin{cases} -\log(h(x)) & if\ y = 1 \\ -\log(1 - h(x)) & if\ y = 0 \end{cases}$
- Since y = 0 or 1 always, can rewrite
- *Cost(h(x), y) = -y log(h(x)) – ((1-y) log(1-h(x))*

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} Cost(h(x_i), y_i)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} y_i \log(h(x_i)) + (1 - y_i)\log(1 - h(x_i))$$

- To fit parameters θ
  - $\min_\theta J(\theta)$
- To make prediction given new x:
  - Output $h(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$
- Just need to find θs
  - Use gradient descent

Gradient Descent

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(h(x_i)) + (1 - y_i)\log(1 - h(x_i))$$

- Want $\min_\theta J(\theta)$
- Repeat

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Gradient Descent

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(h(x_i)) + (1 - yi)\log(1 - h(x_i))$$

- Want $\min_{\theta} J(\theta)$
- Repeat

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Let's try this out