Logistic Regression 1
Andrew Ng. Coursera

Agenda
→ log reg Intro
→ Sigmoid functions
→ Gradient Descent
→ loss function
→ Gradient/ Optimization
→ Implementation

Recap

Linear reg ⟶ Regression : linear model

① Hypothesis : $h(x) = w_0 + w_1 x$
② $J_{cost} = \frac{1}{2m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2$
③ $\frac{\partial J}{\partial w_0}$ & $\frac{\partial J}{\partial w_1}$
④ Low = min $J = \frac{1}{2}(y^{(i)} - y^{(i)})^2$

Logistic Regression

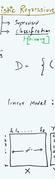
→ Supervised
→ Classification Algo.
→ [0, 1]

Binary classification

$$D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \; ; x^{(i)} \in \mathbb{R}^n, \; y^{(i)} \in \{0, 1\}$$

$y^{(i)} \in \mathbb{R}$

⇒ linear model : hypothesis $\boxed{h(x) = w^T x + w_0}$



Separating hyperplane

$$\hat{y} = [w^T x + w_0]$$

$$y^{(i)} = \begin{cases} 1 & w^T x + w_0 > 0 \\ 0 & w^T x + w_0 < 0 \end{cases}$$

$$\hat{y} = \begin{cases} 1 & w^T x + w_0 \geq 0 \\ 0 & w^T x + w_0 < 0 \end{cases}$$

$\hat{y} \in \{0, 1\}$

Step function

$\hat{y}^{(i)}$ not smooth.

$\hat{y}^{(i)} \in \mathbb{R}$

$\hat{y}^{(i)} \geq 0.5 \Rightarrow 1$

$\hat{y}^{(i)} < 0.5 \Rightarrow 0$

$$y^{(i)} \xrightarrow{\text{hypothesis}} \boxed{h(x^{(i)}) = w^T x + w_0} \xrightarrow{\text{step function}} \hat{y}^{(i)} \rightarrow \{0, 1\}$$

$$y^{(i)} \xrightarrow{\text{hypothesis}} \boxed{h(x^{(i)}) = w^T x + w_0} \xrightarrow{\text{sigmoid}} \hat{y}^{(i)} \rightarrow \{0, 1\}$$

step function

Logistic function.

Sigmoid function.

$x^{(i)} \to \infty \quad \sigma(x^{(i)}) \to 1$

$x^{(i)} \to -\infty \quad \sigma(x^{(i)}) \to 0$

$x^{(i)} \to 0 \quad \sigma(x^{(i)}) \to 0.5$

$$\hat{y} = \sigma(x^{(i)}) = (0, 1) \; \& \; [0, 1]$$

$$\hat{y}^{(i)} = 0.5 \xrightarrow{\sigma(x)=0} x$$

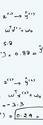
$\boxed{\hat{y} \in (0, 1)}$

$\sigma(x^{(i)})$ $\boxed{\hat{y}^{(i)} = P(y^{(i)} = 1 | x^{(i)})}$ ← most of Sigmoid can be thought as Probability Conditional Probability.

$$P(y^{(i)} = 0 | x^{(i)}) = 1 - \hat{y}^{(i)}$$

$$x^{(i)} \xrightarrow[\text{in dataset}]{w, w_0} \underset{(-\infty, \infty)}{w_1^T x + w_0} \xrightarrow{\sigma(x)} \underset{(0, 1)}{\hat{y}^{(i)}} \quad P(y^{(i)} = 1 | x^{(i)})$$

$\hat{y}^{(i)} \to w_1^T x + w_0$



$\sigma(w_1^T x + w_0)$ = Sigmoid($w_1^T x$) = $\frac{1}{1 + e^{-w_1^T x}}$

① Sometimes binary.
② Speaking for. (0, ∞) → (0, 1)
③ Probabilities of 1
$\hat{y} = P(y^{(i)} = 1 | x^{(i)})$
④ Monotonic function. (non decreasing)

$x_1 \leq x_2$
$\sigma(x_1) \leq \sigma(x_2)$

Gradient Descent.

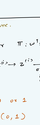
$w^{(i)} \to w^{(i)} - \eta \frac{\partial J}{\partial w^{(i)}}$

$w^{(i)} = w^{(i)} - \eta \, \Delta w^{(i)}$

$w^{(i)} = 0$

$\sigma(w^{(i)}) = 0.98 \Rightarrow \hat{y}^{(i)}$

$\sigma(w^{(i)}) = \boxed{0.98 < \hat{y}^{(i)}}$

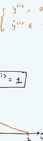


x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}

y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10}

$\sigma(x^{(i)}) > 0.5$

$\sigma(x^{(i)}) > 0.5$


$\hat{y}^{(i)}$
$\hat{y}^{(i)}$
$\hat{y}^{(i)}$