Logistic Regression

Binary classification

Binary classification / Logistic Regression

$\boxed{\begin{array}{l} \text{Input} \\ x \in \mathbb{R}^n \end{array}}$ $x^{(i)} \in \mathbb{R}^n$, $x^{(m)} \in \mathbb{R}^n$

output $\in \mathbb{R}$

$y_{\text{hat}} = \text{output}$

$y_{\text{hat}} \in [-\infty, \infty]$ Continuous

$y_{\text{hat}} \in \{0, 1, 2, 3, 4\}$ Forgiveness

Classification

Strong

→ [Binary / Multiclass]

→ [Real / Int]

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

Weak / Clue

→ Clue: Yes, Clue: No

→ Clue: Yes, Clue: Maybe

→ multiple categories

→ Cancer, not cancer, Maybe

$y \in \{0, 1, \dots, k\}$

Binary classification

e.g. Picture Recognition / Tumor

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

Input

x_1: #Dogs ordering 🍕 not value

x_2: no. of calls

x_1 = Dogs

x_2 = Magnitude

$x_1 + x_2 + b = 0$ (Model)

$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $\theta^T x + b$

$\theta^T x$

Linear reg → $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i$ or $\theta^T x$

$\theta_0 = b$

$h_\theta(x) = \theta^T x \in \mathbb{R}$

Problem!

$\boxed{h_\theta(x) \text{ is real}}$

Fixed Upper number → 0 (-∞, 0)

→ 1 (0, ∞)

$\text{Prediction}(x; \theta) = \begin{cases} 0 & ; \theta^T x \geq 0 \\ 1 & ; \theta^T x < 0 \end{cases}$ →

→

Step function

$z = \theta^T x$

Not telling you the Probability.

$z \in (-\infty, \infty)$

$y \in (0, 1)$

$\boxed{\text{Squishy the values}}$

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

Sigmoid function

→ Impossible

→ Cannot (∞)

→ Probability, comparison

$\boxed{h_\theta(x)}$ → 1 Hypothesis

$z = \theta^T x$

$h_\theta(x) = \sigma(z)$

Sigmoid Curve

Now hypothesis $h_\theta(x) = \sigma(\theta^T x)$

$\boxed{h_\theta(x) \in (0, 1)}$

$\sigma(\theta^T x) \geq 0.5$; $\sigma(z) \geq 0.5$

$\sigma(\theta^T x) < 0.5$; $\sigma(z) < 0.5$

if $z \to \infty$ $\sigma(z) \to 1$

$z \to -\infty$ $\sigma(z) \to 0$

$z \to 0$ $\sigma(z) \to 0.5$

$\sigma(\theta^T x)$ gives Probability in which the class.

Hypothesis $h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$

Prediction function: $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$

⇒ Synthesis, Error Fn, Gradient, Gradient-Ascent

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

$\boxed{\text{Decision Surface / Boundary}}$

Likelihood Estimation & Loss function

$P(y = 1 \mid x; \theta) = h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$

Probability of 1

$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$

Probability Distribution of Outcomes

$P(y=1|\theta) = 0.7$

$P(y=0|\theta) = 1 - 0.7$

$= 0.3$

$P(y=1|x)$, $P(y=0|x)$

$\boxed{P(y \mid x; \theta) = [h_\theta(x)]^{y} [1 - h_\theta(x)]^{(1-y)}}$

$y = 1$ $P(y=1|x) = h_\theta(x)$

$y = 0$ $P(y=0|x) = 1 - h_\theta(x)$

$\ell(\theta) = P(y \mid x; \theta)$

[Likelihood]

$\ell(\theta) = \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta)$

Independent

$\ell(\theta) = \prod_{i=1}^{m} [h_\theta(x^{(i)})]^{y^{(i)}} [1 - h_\theta(x^{(i)})]^{(1-y^{(i)})}$

max $\ell(\theta)$

$\boxed{\log}$ monotonic function

$x_1 \leq x_2$

$\log(x_1) \leq \log(x_2)$

$\log(x^y) = y \log x$

take log

$\log(\ell(\theta)) = \frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))$

Maximize log likelihood

Global maxima

$f_1 = (x-5)^2$

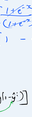
minimum

$f_2 = -(y-5)^2$

$-f$

$\underline{\text{Loss}} = -\log(\ell(\theta))$

Negative log likelihood



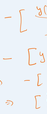
$-\ell(\theta)$

loss$(y, \hat{y})$

$L = \text{loss}(\theta) = -\frac{1}{m} \left[\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$

Neg log likelihood

$h_\theta(x) = \sigma(\theta^T x)$



Actual $y = 1$

$h_\theta(x^{(i)})$

Actual $y = 0$

$h_\theta(x^{(i)})$

Gradient

$\frac{\partial L}{\partial \theta_j}$?

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$

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

$\frac{d}{dx} e^{-x} \Rightarrow \frac{d(e^{-x})}{dx} \Rightarrow \frac{d}{dx}(1 + e^{-x})^{-1}$

$e^{-x} \cdot e^{-x}$

$\frac{dx}{dx} \cdot \frac{dx}{dx}$

$\Rightarrow \frac{d}{d(1 + e^{-x})}(1 + e^{-x})^{-1} \cdot \frac{d(1 + e^{-x})}{dx}$

$\Rightarrow \frac{-1}{(1 + e^{-x})^2} \cdot e^{-x} \cdot (-1)$

$\Rightarrow \frac{e^{-x}}{(1 + e^{-x})^2}$

$\Rightarrow \frac{1 - 1}{(1 + e^{-x})} \cdot \frac{(1 + e^{-x})}{(1 + e^{-x})}$

$\sigma(x) = \left[\frac{1 + e^{-x}}{(1 + e^{-x})} - \frac{1}{(1 + e^{-x})} \right]$

$\sigma(x) = \left[\frac{1 + e^{-x}}{(1 + e^{-x})} \right] - \left[\frac{1}{(1 + e^{-x})} \right]$

$\sigma(x) = [1 - \sigma(x)]$

$\boxed{\frac{d}{dx} \sigma(x) = \sigma(x) \cdot [1 - \sigma(x)]}$

$\frac{d}{d\theta_j} L \Rightarrow -\frac{1}{m} \left[\sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)}) \right]$ $\hat{y}^{(i)} = h_\theta(x^{(i)}) = \sigma(\theta^T x^{(i)})$

$= -\underbrace{\frac{d}{d\hat{y}} \left[y \cdot \log \hat{y} + (1-y) \log(1-\hat{y}) \right]}_{} \cdot \underbrace{\frac{d\hat{y}}{dz} \cdot \frac{dz}{d\theta_j}}_{}$

Sigmoid Linear reg.

$\Rightarrow -\left[\frac{y}{\hat{y}} - \frac{1-y}{(1-\hat{y})} \right] \cdot \hat{y} \cdot (1-\hat{y}) \cdot \frac{d(\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n)}{d\theta_j}$

x_j

$\Rightarrow -\left[\frac{y(1-\hat{y}) - \hat{y}(1-y)}{\hat{y}(1-\hat{y})} \right] \hat{y} \cdot (1-\hat{y}) \cdot x_j$

$\Rightarrow -\left[y - y\hat{y} - \hat{y} + \cancel{y\hat{y}} \right] \cdot x_j$

$\Rightarrow -\left[y - \hat{y} \right] x_j$

$\Rightarrow \left[\hat{y} - y \right] x_j$

$\boxed{\frac{\partial L}{\partial \theta_j} \Rightarrow \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)}}$

$\hat{y}^{(i)} = \theta^T x$

$\hat{y}^{(i)} = h_\theta(x^{(i)}) = \sigma(\theta^T x)$

Gradient Ascent

$\boxed{\begin{array}{l} \text{Separate weight coming from} \\ \theta_j = \theta_j - \alpha \frac{\partial L}{\partial \theta_j} \end{array}}$

① hypothesis function ⇒ $h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n)}}$

② Loss func: $L = -\left[\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)}) \right]$

③ Gradient: $\frac{\partial L}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \cdot x_j^{(i)}$ ✓

④ G.D $\theta_j = \theta_j - \alpha \frac{\partial L}{\partial \theta_j}$ ✓