CS 307 Lecture 6

Logistic Regression
~~Linear~~ Classification : Training Set $= \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

with an add^n 1 ← $x^{(i)} \in \mathbb{R}^{d+1}$, $y^{(i)} \in \{-1,+1\}$

Goal: Compute $h_\theta : \mathbb{R}^{d+1} \to \{-1,+1\}$ s.t.

$$h_\theta(x) = \begin{cases} +1 & \text{if } \theta^T x > 0 \\ -1 & \text{o.w.} \end{cases}$$

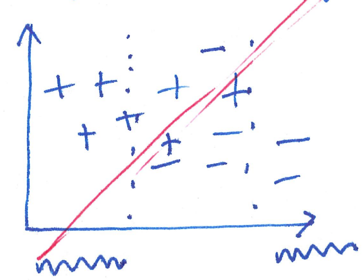
where $J(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta, x^{(i)}, y^{(i)})$ is minimized

$$l_{log}(\theta, x^{(i)}, y^{(i)}) = \ln\left(1 + e^{-y^{(i)} \theta^T x^{(i)}}\right)$$

Revisiting Logistic Regression

Training Set $= \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^{d+1}$, $y^{(i)} \in \{0,1\}$

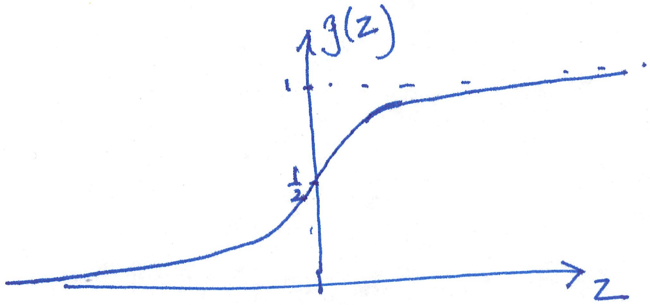
Think of the hypothesis as outputing the probability of the output being 1.



i.e
$$P[y=1 \mid x;\theta] = h_\theta(x).$$
$$P[y=0 \mid x;\theta] = 1 - h_\theta(x)$$

$$h_\theta(x) = g(\theta^T x) \in [0,1]$$

$$g(z) = \frac{1}{1+e^{-z}}$$] sigmoid, logistic fn.



$$\begin{bmatrix} P[y=1 \mid x;\theta] = h_\theta(x) \\ P[y=0 \mid x;\theta] = 1-h_\theta(x) \end{bmatrix} \quad \text{where} \quad h_\theta(x) = g(\theta^T x), \quad g(z) = \frac{1}{1+e^{-z}}.$$

(a) Given h_θ what is the output on an input x?

Output 1 iff $h_\theta(x) = P[y=1 \mid x;\theta] > 0.5$.

$$\frac{1}{1+e^{-\theta^T x}} > \frac{1}{2} \Rightarrow e^{-\theta^T x} < 1 \Rightarrow \theta^T x > 0$$

$a^x < 1$ $(a>1)$
$\Rightarrow x < 0$

(b) How do we find h_θ?

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

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

$$\underbrace{P[\{y^{(i)}\}_{i=1}^n \mid \{x^{(i)}\}_{i=1}^n ;\theta]} = \prod_{i=1}^n P[y^{(i)} \mid x^{(i)};\theta]$$
$$= \prod_{i=1}^n h_\theta(x^{(i)})^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

$$\underbrace{L(\theta)}_{\text{Likelihood}} = \prod_{i=1}^n h_\theta(x^{(i)})^{y(i)} (1-h_\theta(x^{(i)}))^{(1-y^{(i)})}$$

Pick θ s.t. $L(\theta)$ is maximized.

Maximum Likelihood Principle

$$\log L(\theta) = \frac{1}{n}\sum_{i=1}^n y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))$$

New Goal: Maximize log of likelihood

Minimize $-$(log of likelihood).

$$l_{nll}(\theta, x, y) = -[y \log h_\theta(x) + (1-y)\log(1-h_\theta(x))]$$

Minimize $J(\theta) = \frac{1}{n}\sum_{i=1}^n l_{nll}(\theta, x^{(i)}, y^{(i)})$

Training set : $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

Find θ s.t. ~~maxize~~ minimize

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_{nll}(\theta, x^{(i)}, y^{(i)}) \quad \text{where}$$

$$l_{nll}(\theta, x, y) = -\left[y \log h_\theta(x) + (1-y)\log(1-h_\theta(x))\right]$$

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

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

Run Gradient Descent.

Need compute $\nabla_\theta J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta l_{nll}(\theta, x^{(i)}, y^{(i)})$

$$\frac{d}{dz} g(z) = \frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right) = \frac{-1}{(1+e^{-z})^2}\left[e^{-z}(-1)\right]$$

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

$$= g(z)(1-g(z))$$

$$\frac{\partial}{\partial\theta_j} l_{nll}(\theta, x, y) = -\frac{\partial}{\partial\theta_j}\left[y \log h_\theta(x) + (1-y)\log(1-h_\theta(x))\right]$$

$$= -\left[y\frac{1}{h_\theta(x)} \cdot \frac{\partial h_\theta(x)}{\partial\theta_j} + \frac{(1-y)}{1-h_\theta(x)} \cdot \frac{\partial(1-h_\theta(x))}{\partial\theta_j}\right]$$

$$\boxed{\frac{d}{dx} h(y(x)) = \frac{d}{d(y(x))} h(y(x)) \cdot \frac{d}{dx} y(x)}$$

$$= -\frac{\partial h_\theta(x)}{\partial\theta_j}$$

$$= -\left[\frac{y}{h_\theta(x)} - \frac{(1-y)}{1-h_\theta(x)}\right]\frac{\partial h_\theta(x)}{\partial\theta_j} \qquad (h_\theta(x) = g(\theta^T x))$$

$$\theta^T x = \theta_0 x_0 + \theta_1 x_1 + \dots \theta_j x_j$$

$$\frac{\partial}{\partial\theta_j}\theta^T x = x_j$$

$$\cancel{= \left[\frac{y}{h_\theta(x)} \quad \frac{(1-y)}{}\right]}$$

$$= -\left[\frac{y}{g(\theta^T x)} - \frac{(1-y)}{1-g(\theta^T x)}\right] g(\theta^T x)(1-g(\theta^T x))\frac{\partial\theta^T x}{\partial\theta_j}$$

$$= -\left[y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right]x_j$$

$$= -\left[y - \cancel{yg(\theta^T x)} - g(\theta^T x) + \cancel{y\,g(\theta^T x)}\right]x_j = \left[g(\theta^T x) - y\right]x_j$$

$$= \left[h_\theta(x) - y\right]x_j$$

Update Rule for gradient descent for logistic Regression

$$\theta = \theta - \alpha \nabla_\theta J(\theta)$$

$$\begin{aligned}
\theta_j &= \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\
&= \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(\underbrace{\frac{\partial}{\partial \theta_j} l_{nll}(\theta, x^{(i)}, y^{(i)})} \right) \\
&= \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}
\end{aligned}$$

For linear regression: $h_\theta(x^{(i)}) = \theta^T x^{(i)}$

For logistic regression: $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$