Logistic Regression for Classification:

DS 4400
Ehsan Elhamifar

Can we build a classification scheme that models and optimizes $p(y|x)$ without the need to model relationships among different features/attributes?

$\longrightarrow$ Discriminative Modeling $p(y|x)$ ✓

Intuition from linear regression $f(x) = w^T \phi(x)$ (optimizing for w]

Can we learn a function g $X \to y$

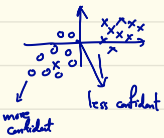
So that if $w^T \phi(x) > 0 \longrightarrow g(x) = 1$
$w^T \phi(x) < 0 \longrightarrow g(x) = 0$

$\hookrightarrow$ Separating hyperplane in the X space, dividing into two spaces / classes

We want to come up with a probabilistic model for $p(y|x)$

In the above, if $p(y=1|x) = \begin{cases} 1 & , w^T\phi(x) > 0 \\ 0 & , w^T\phi(x) < 0 \end{cases}$

$p(y_1 \dots y_n | x_1, \dots, x_n) = \prod p(x_i | y_i) = 0$ $\longleftarrow$ A single mistake will make the dataset to have a zero probability



We need to have confidence score about $p(y=1|x)$ farther from hyper plane then more confident in it being class 0 or 1 $\longrightarrow$ use exponential

$$\begin{cases} p(y=1 \mid x) \propto e^{\frac{1}{2} w^T \phi(x)} = \frac{1}{Z} e^{\frac{1}{2} w^T \phi(x)} \\ p(y=0 \mid x) \propto e^{-\frac{1}{2} w^T \phi(x)} \quad \text{to have a symmetric classifier} \quad = \frac{1}{Z} e^{-\frac{1}{2} w^T \phi(x)} \end{cases}$$

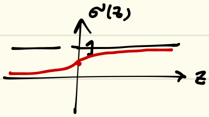
$$\to p(y=1 \mid x) + p(y=0 \mid x) = 1 \to \frac{1}{Z}\left[e^{\frac{1}{2} w^T \phi(x)} + e^{-\frac{1}{2} w^T \phi(x)} \right] = 1$$

$$\to \underbrace{Z}_{\text{normalizing constant}} = e^{\frac{1}{2} w^T \phi(x)} + e^{-\frac{1}{2} w^T \phi(x)}$$

$$\to p(y=1 \mid x) = \frac{e^{\frac{1}{2} w^T \phi(x)}}{e^{\frac{1}{2} w^T \phi(x)} + e^{-\frac{1}{2} w^T \phi(x)}} = \frac{1}{1 + e^{-w^T \phi(x)}} = \sigma(w^T \phi(x))$$

↳ logistic function

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



squashing $w^T\phi(x)$ (linear reg) to $[0,1]$

* what is data ?

$$D = \{ (x^1, y^1), \; , (x^N, y^N) \} \quad \text{where } x^1 = \begin{pmatrix} x^1_1 \\ x^1_2 \end{pmatrix}, \; y^1 \in \{0, 1, \; , L-1\}$$

↳ categorical / continuous

categorical $x^1 = \{0, 1, 2\}$ → sunny, cloudy, rainy

$x^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ "one hot" encoding all but 1 are zeros

* Parameters to learn ?

weight vectors $w \in \mathbb{R}^{|\phi(x)|}$ which model $p(y=1 \mid x) = \frac{1}{1 + e^{-\phi(x)^T w}}$

* Cost function? conditional log-likelihood

$$l(w) = \sum_{i=1}^{N} \log p(y^i | x^i, w)$$

$$p(y^i|x^i,w) = p(y^i=1|x^i,w)^{\mathbb{1}(y^i=1)} \times p(y^i=0|x^i,w)^{\mathbb{1}(y^i=0)}$$

$$l(w) = \sum_{i=1}^{N} \mathbb{1}(y^i=1) \log p(y^i=1|x^i,w) + \mathbb{1}(y^i=0) \log p(y^i=0|x^i,w)$$

$$= \sum_{i=1}^{N} \mathbb{1}(y^i=1) \log \frac{1}{1+e^{-\phi(x_i)^T w}} + \underbrace{\mathbb{1}(y^i=0)}_{1-\mathbb{1}(y^i=1)} \log\left(1 - \frac{1}{1+e^{-\phi(x_i)^T w}}\right)$$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{1+e^{+w^T\phi(x_i)}}\right) + \sum_{i=1}^{N} \mathbb{1}(y^i=1) \log \frac{\frac{1}{1+e^{-\phi(x_i)^T w}}}{\frac{1}{1+e^{\phi(x_i)^T w}}}$$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{1+e^{+w^T\phi(x_i)}}\right) + \sum_{i=1}^{N} \mathbb{1}(y^i=1) \log \frac{\frac{1}{1+e^{-\phi(x_i)^T w}}}{\frac{1}{1+e^{\phi(x_i)^T w}}}$$

$$= \sum_{i=1}^{N} \log \frac{1}{1+e^{w^T\phi(x_i)}} + \sum_{i=1}^{N} y^i \log \frac{1+e^{\phi(x_i)^T w}}{1+e^{-\phi(x_i)^T w}} \rightarrow e^{\phi(x_i)^T w}\left(1+e^{-\phi(x_i)^T w}\right)$$

$$= -\sum_{i=1}^{N} \log\left(1+e^{w^T\phi(x_i)}\right) + \sum_{i=1}^{N} y^i \phi(x_i)^T w$$

$$\Rightarrow l(w) = \sum_{i=1}^{N} \left[y^i \phi(x^i)^T w - \log\left(1+e^{w^T\phi(x_i)}\right) \right]$$

We want to maximize $l(w)$ or minimize $-l(w)$,

$$J(w) = -l(w) = \sum_{i=1}^{N} \left[-y^i \phi(x^i)^T w + \log\left(1+e^{w^T\phi(x_i)}\right) \right]$$

$$J(w) = -l(w) = \sum_{i=1}^{N} \left[-y^i \phi(x^i)^T w + \log\left(1+e^{w^T\phi(x_i)}\right) \right]$$

$$\frac{\partial J}{\partial w} = \sum_{i=1}^{N} -y^i \phi(x^i) + \frac{\partial \log\left(1+e^{w^T\phi(x_i)}\right)}{\partial w} \quad // \quad -\frac{\frac{\partial}{\partial w}\sigma(w^T\phi(x_i))}{\sigma(w^T\phi(x_i))}$$

$$\| \quad -\phi(x_i)\,\sigma(w^T\phi(x_i))\left(1-\sigma(w^T\phi(x_i))\right)$$

$$\log\left(1+e^{w^T\phi(x_i)}\right) = -\log\left(\sigma(-w^T\phi(x_i))\right)$$

$$\frac{\partial J}{\partial w} = \sum_{i} -\phi(x^i)\left[y^i - \underbrace{p(y^i=1 \mid x^i, w)}_{\frac{1}{1+e^{-\phi(x_i)^T w}}} \right]$$

No closed-form sol for $\frac{\partial J}{\partial w} = 0$!

Gradient descent

+ Initialize $w^{(0)} = 0$
+ At iteration k, update $w^{(k)} = w^{(k-1)} - \alpha \frac{\partial J}{\partial w}$

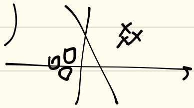
$$= w^{(k-1)} + \alpha \sum_i \phi(x^i)\left[y^i - p(y=1 \mid x^i, w) \right]$$

<u>Overfitting</u>

what is the decision boundary for logistic reg? where $p(y=1 \mid x, w) = \frac{1}{2}$

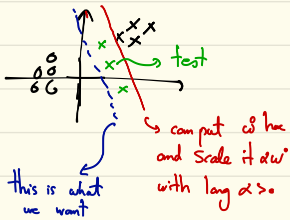
$$= \frac{1}{1+e^{-\phi(x)^T w}}$$

$\longrightarrow \phi^T(x) w = 0$ (scaling w does not change this)



$\longrightarrow$ scale w up by very large values to make $p(y=1 \mid x, w) \simeq 1$

overfitting



To prevent overfitting, do regularization on ω:

$$\min_{\omega} J(\omega) + \frac{\lambda}{2} \|\omega\|_2^2 \triangleq \bar{J}(\omega)$$

In G.D.

$$\frac{\partial \bar{J}}{\partial \omega} = \underbrace{\frac{\partial J}{\partial \omega}}_{\text{already computed.}} + \lambda \omega$$

$$= \sum_{i} -\phi(x^i)\left[y^i - \frac{1}{1+e^{-\phi(x_i)^T \omega}} \right] + \lambda \omega$$