Aplicaciones en Internet

Classification Algorithms

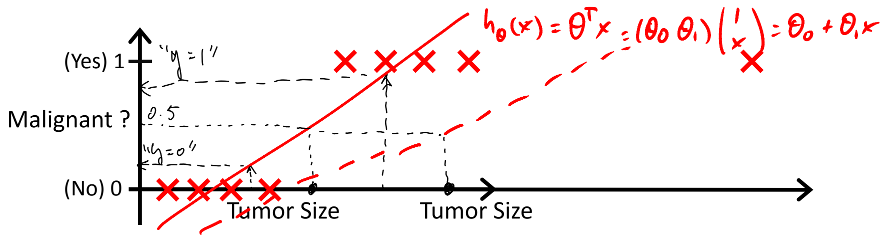
Logistic Regression

Classification

Email: Spam / Not Spam?
Online Transactions: Fraudulent (Yes / No)?
Tumor: Malignant / Benign ?

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

0: "Negative Class" (e.g., benign tumor)
1: "Positive Class" (e.g., malignant tumor)



Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

because $y \in \{0, 1\}$
we want
$0 \leq h_\theta(x) \leq 1$

Classification: y = 0 or 1

$h_\theta(x)$ can be > 1 or < 0

Logistic Regression: $0 \leq h_\theta(x) \leq 1$

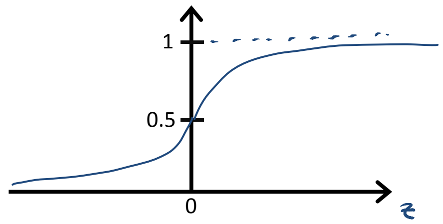
Logistic Regression Model

Want $0 \le h_\theta(x) \le 1$

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

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

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



Sigmoid function
Logistic function

$\lim_{z \to -\infty} g(z) = 0 \qquad \lim_{z \to \infty} g(z) = 1$

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that y = 1 on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$$h_\theta(x) = 0.7$$

Tell patient that 70% chance of tumor being malignant

$h_\theta(x) = P(y=1 \mid x;\theta) \longrightarrow$ "probability that y = 1, given x, parameterized by θ"

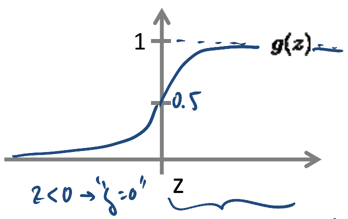
because $y \in \{0,1\} \longrightarrow$

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

Logistic regression

parameter vector
feature vector

$$h_\theta(x) = g(\theta^T x)$$
$$g(z) = \frac{1}{1+e^{-z}}$$



Suppose predict "$y = 1$" if $h_\theta(x) \ge 0.5 \rightarrow g(z) \ge 0.5$

$\rightarrow z \ge 0 \rightarrow \theta^T x \ge 0$

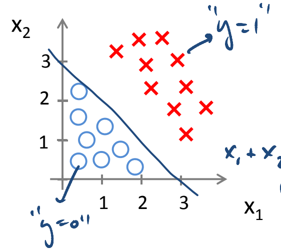
predict "$y = 0$" if $h_\theta(x) < 0.5 \rightarrow g(z) < 0.5$

$\rightarrow z < 0 \rightarrow \theta^T x < 0$

DECISION BOUNDARY $\theta^T x = 0$

Decision Boundary

$$h_\theta(x) = g(\theta^T x) = g\left((\theta_0, \theta_1, \theta_2)\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}\right)$$



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$x_1 + x_2 = 3 \rightarrow -3 + x_1 + x_2 = 0 \rightarrow (-3, 1, 1)\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = 0$

DECISION BOUNDARY

$\theta = \begin{pmatrix} -3 \\ 1 \\ 1 \end{pmatrix}$

Predict "$y = 1$" if $-3 + x_1 + x_2 \ge 0$

Predict "$y = 0$" if $-3 + x_1 + x_2 < 0$

Suppose we use ~~linear~~ logistic regression with two characteristics x_1 y x_2 and get θ_0=5, θ_1=-1, θ_2=0. Which of these figures shows the decision boundary?

$\theta = \begin{pmatrix} 5 \\ -1 \\ 0 \end{pmatrix}$ $\theta^T x = 0$

$(5 \ -1 \ 0)\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = 0$

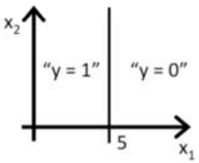
$5 - x_1 = 0$

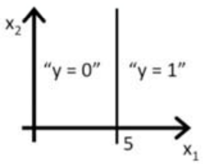
$x_1 = 5$

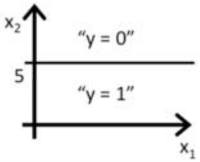
"y=1" if $\theta^T x \geq 0$

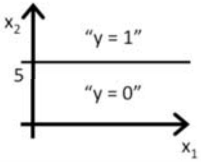
$5 - x_1 \geq 0$

$5 \geq x_1$

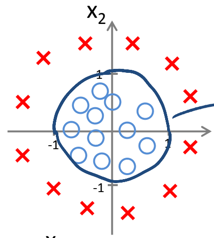
X 







Non-linear decision boundaries



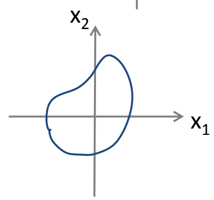
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$x_1^2 + x_2^2 = 1$

$-1 + x_1^2 + x_2^2 = 0$

$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \ldots)$$

Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

m examples $\quad x \in \begin{bmatrix} x_0 \\ x_1 \\ \ldots \\ x_n \end{bmatrix} \quad x_0 = 1, y \in \{0, 1\}$

x_1^2

$x_1 x_2$

$x_1 x_2^2$

$\vdots$

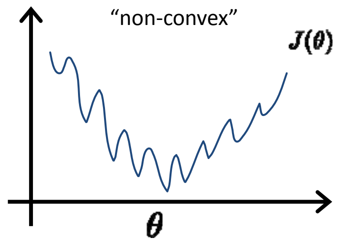
$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

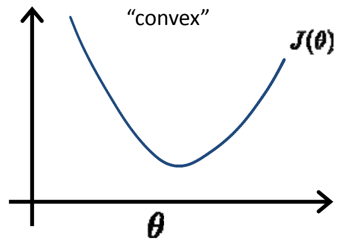
How to choose parameters θ ?

Cost function

Linear regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

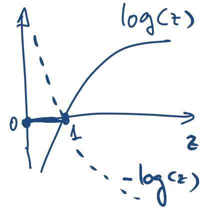
$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2} \left(\underbrace{h_\theta(x^{(i)})}_{\frac{1}{1+e^{-\theta^T x}}} - y^{(i)}\right)^2$$

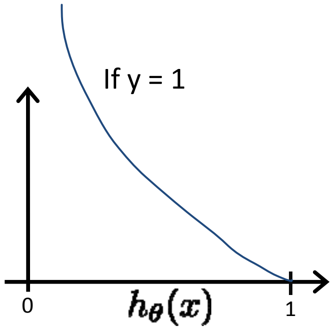




Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



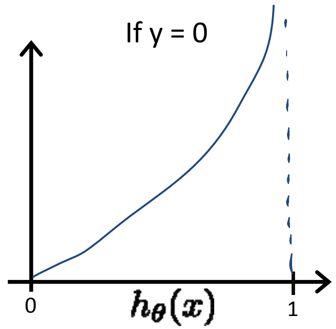


Cost $= 0$ if $y = 1, h_\theta(x) = 1$
But as $h_\theta(x) \to 0$ $\to h_\theta(x) = P(y=1|x;\theta)$
$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



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

this cost is known as the
$-\log$ likelihood

Logistic regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note: $y = 0$ or 1 always

Logistic regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m}\Big[\sum_{i=1}^{m} \underbrace{y^{(i)} \log h_\theta(x^{(i)})}_{\text{cost with } y^{(i)}=1} + \underbrace{(1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right)}_{\text{cost with } y^{(i)}=0}\Big]$$

To fit parameters θ:

$$\min_\theta J(\theta)$$

To make a prediction given new x:

Output $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ $\to \begin{cases} \text{if } h_\theta(x) \geq 0.5 \to \theta^T x \geq 0 \to y = 1 \\ \text{if } h_\theta(x) < 0.5 \to \theta^T x < 0 \to y = 0 \end{cases}$

Gradient Descent

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

Want $\min_\theta J(\theta)$:

Repeat {

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

} (simultaneously update all θ_j)

Gradient Descent

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

Want $\min_\theta J(\theta)$:

in logistic regression $h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x}}$

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

Repeat {

$$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} + \underbrace{\frac{\lambda}{m}\theta_j}_{\text{for regularization}}\right]$$

} (simultaneously update all θ_j)

Algorithm looks identical to linear regression!

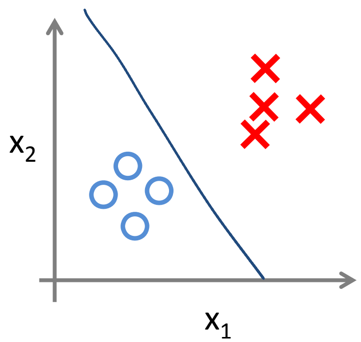
Multiclass classification

Email foldering/tagging: Work, Friends, Family, Hobby

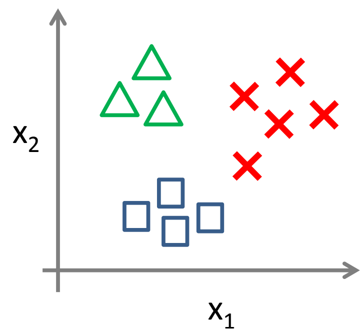
Medical diagrams: Not ill, Cold, Flu

Weather: Sunny, Cloudy, Rain, Snow

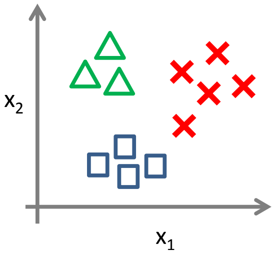
Binary classification:

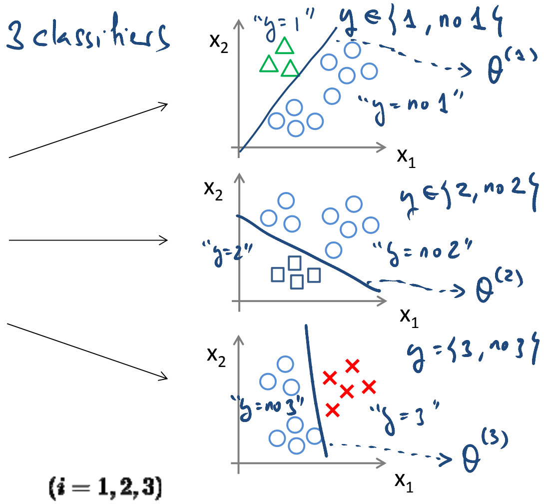


Multi-class classification:



One-vs-all (one-vs-rest):





Class 1: △
Class 2: □
Class 3: ✕

$h_\theta^{(i)}(x) = P(y = i | x; \theta) \qquad (i = 1, 2, 3)$

One-vs-all

Train a logistic regression classifier $h_\theta^{(i)}(x)$ for each class i to predict the probability that $y = i$.

$h_\theta^{(1)}(x) = h_{\theta^{(1)}}(x) = P(y=1 | x; \theta^{(1)})$, $h_{\theta^{(2)}}(x) = P(y=2 | x; \theta^{(2)})$ ...

On a new input x, to make a prediction, pick the class i that maximizes

$$\max_i h_\theta^{(i)}(x) \longrightarrow \max_i P(y=i | x; \theta^{(i)})$$

$$\hookrightarrow \max_i \left((\theta^{(i)})^T x \right)$$

Suppose we have a classification problem with k classes. Using the 1-vs-all method, how many logistical sorters will we have to train?

o k-1

o k

o k+1

o Approximately $\log_2(k)$