Logistic Regression: Classification

Email: Spam/Not?

Online Transactions: Fraudulent (Y/N)

Tumor: Malig/Benign?

$y \in \{0, 1\}$ — 0: "Negative Class", 1: "Positive Class" (absence of / presence of)

multiclass $y \in \{0, 1, 2, 3\}$

binary

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

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

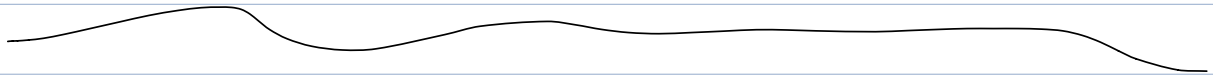
If $h_\theta(x) < 0.5$, "$y = 0$"

Classification: $y = 0$ or 1

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

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

(Classification)

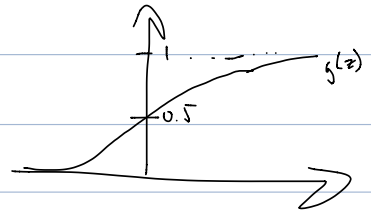


Want $0 \leq h_\theta(x) \leq 1$

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

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

$g(z) = \frac{1}{1 + e^{-z}}$ → sigmoid/logistic function



parameters θ

Interpretation of Hypothesis Output

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

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

$y = 0$ or 1

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

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

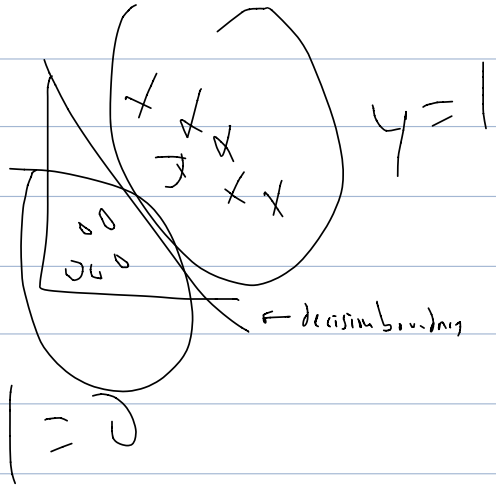
Decision Boundary

$g(z) \geq 0.5$ when $z \geq 0$

$h_\theta(x) = g(\theta^T x) \geq 0.5$ when $\theta^T x \geq 0$

$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$ → predict "$y=1$" if $\underbrace{-3 + x_1 + x_2}_{\theta^T x} \geq 0$

$x_1 + x_2 \geq 3$



Logistic Regression: Cost Function

Training Set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$

m examples

$x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \mathbb{R}^{n+1}$ $x_0 = 1, y \in \{0,1\}$

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

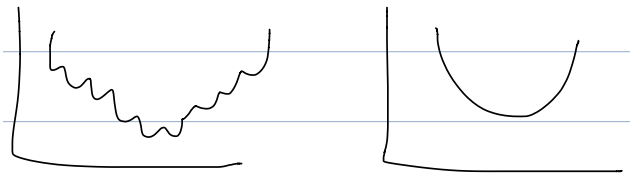
How to choose parameters θ?

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

$\text{Cost}(h_\theta(x^{(i)}), y)$

$\text{Cost}(h_\theta(x), y) = \frac{1}{2}(h_\theta(x) - y)^2$

"non-convex" "convex"



Logistic regression cost function

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

Simplified Cost Function & gradient descent

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

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

$$\text{Cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y)\log(1 - h_\theta(x))$$

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

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

To fit parameters θ:

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

To make a prediction given new x:

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

Gradient descent

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

Repeat {

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

}

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

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

Repeat {

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

}

Algorithm looks identical to linear regression!

Advanced Optimization

- Gradient descent
- Conjugate gradient
- BFGS
- L-BFGS

Multi-class classification

ex: Email folders/tagging: Work, friends, family, Lobby

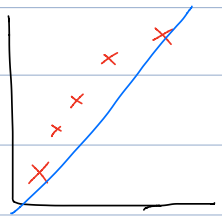
ex2: medical diagnosis: not ill, cold, flu

ex3: Weather: sunny, cloudy, rain, snow
y = 1 2 3 4

One vs all (one vs rest)

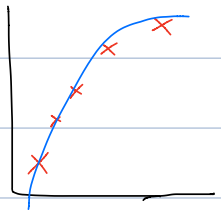
$h_\theta^{(i)}(x) = P(y=i \mid x; \theta)$ $(i = 1, 2, 3)$

The Problem of Overfitting



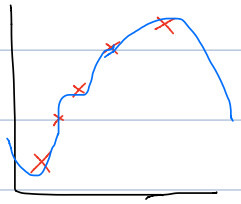
$\theta_0 + \theta_1 x$

"underfit" "high bias"



$\theta_0 + \theta_1 x + \theta_2 x^2$

"just right"



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"overfit" "high variance"

Addressing overfitting

options:

1. Reduce # of features
   - manually select
   - Model selection algorithm
2. Regularization
   - keep all features, but reduce magnitude/values of parameters θ_j

---

Regularization Cost Function

Small values for parameters $\theta_0, \theta_1, \ldots, \theta_n$

- "Simple" hypothesis

- Less prone to overfitting

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

(regularization parameter)

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

Regularized Linear Regression

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

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

Repeat {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}$$

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

$\left(\frac{\partial}{\partial \theta_j} J(\theta)\right)$ regularized

$$\theta_j := \theta_j\underbrace{\left(1 - \alpha\frac{\lambda}{m}\right)}_{<1} - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

}

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}\right)^{-1} X^T y$$

Regularized Logistic Regression

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

Repeat {

$$\rightarrow \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_0^{(i)}$$

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

}

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