Logistic Regression

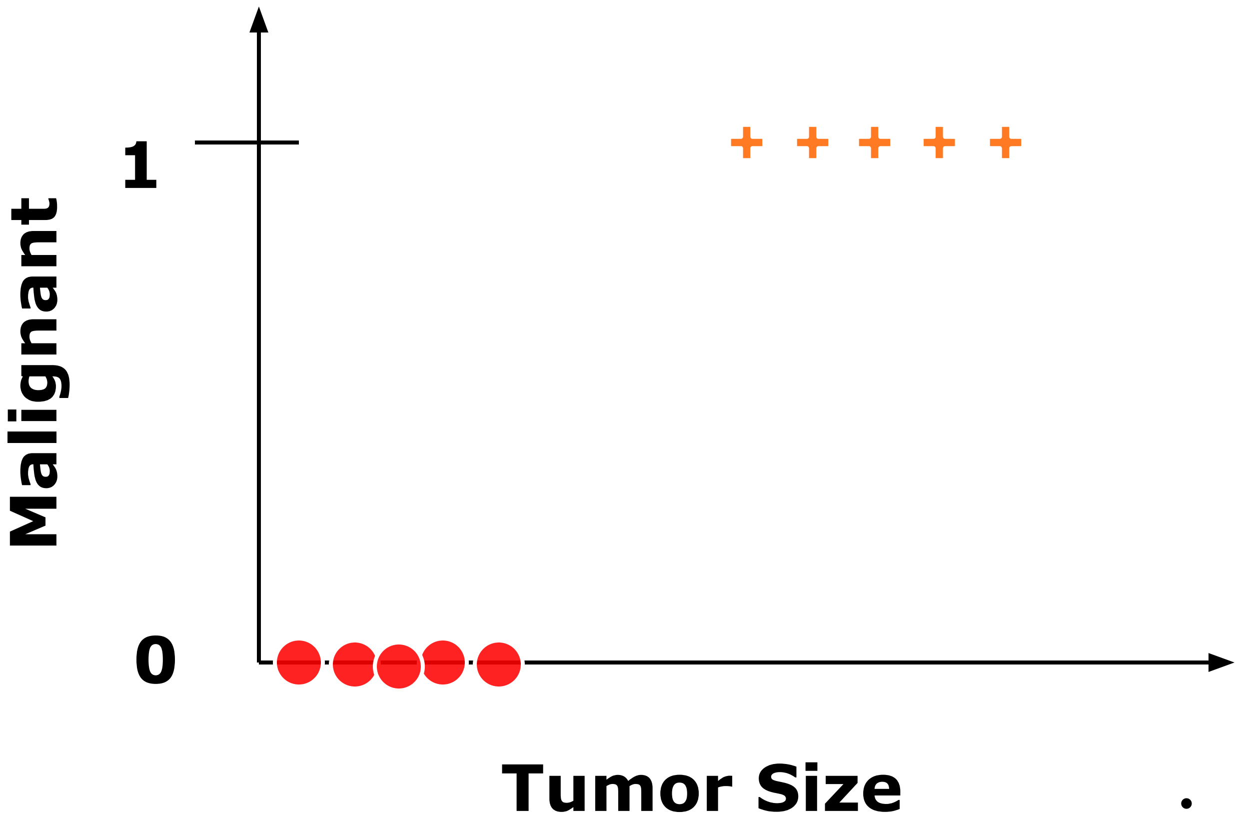
Agenda

- Regression for Binary Variables
- Logistic Regression Hypothesis function
- Decision Boundary
- Cost Function
- Gradient Descent
- Overfitting Problem
- Regularized Logistic Regression

Regression for Binary Variables

- Examples
  - Email: Spam/Not Spam?
  - Cancer: Malignant or Benign?
  - Fraud Detection
  - Loan Defaulters
- Variable takes binary values {0,1}
  - 0: negative class
  - 1: positive class

Why not Linear Regression?



- Doesn’t do well with outliers
- Can assume values well beyond 0 and 1

Logistic Regression

- Classification Algorithm
- Y is discrete/binary
- Hypothesis Function

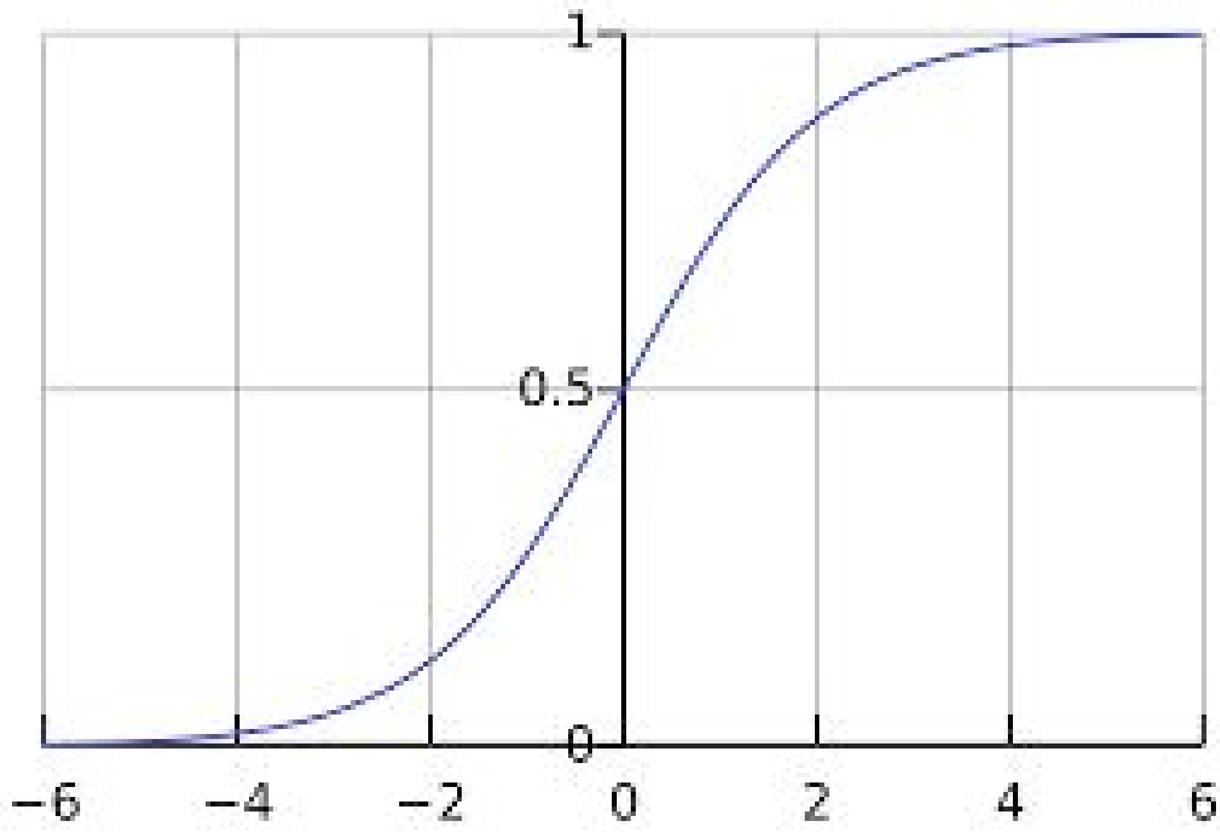
$0 <= h_\theta(x) <= 1$

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

$g(z) = {}^{1}/_{1+e^{-z}}$ | Sigmoid/Logistic Function

$h_\theta(x) = {}^{1}/_{1+e^{-\theta_T X}}$

Plotting Sigmoid Function



Interpretation

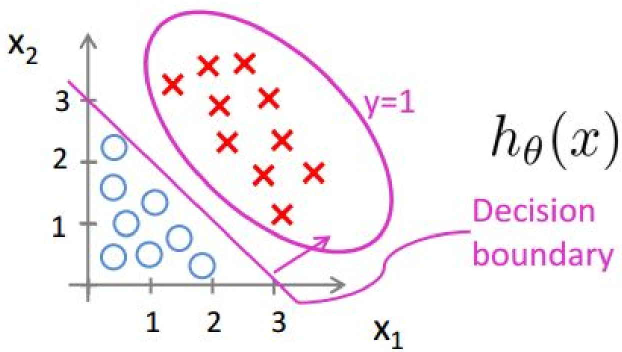
- $h_\theta(x)$ computed will be the probability that y=1
- Example:
  - If x=(x0,x1) = (1,tumor size)
  - $h_\theta(x) = 0.7$

  70% chance that tumor is malignant!

  $h_\theta(x) = p(y=1|x;\theta)$

Decision Boundary

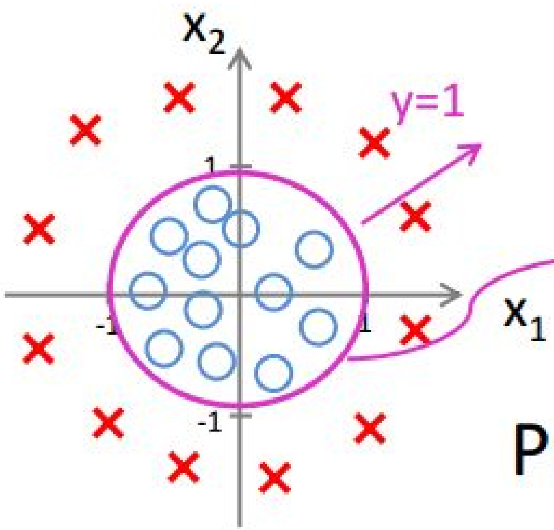
- Suppose predict y=1 if $h_\theta(x) >= 0.5$
- $h_\theta(x) = g(\theta^T X) >= 0.5$
- $\theta^T X >= 0$



$$h_\theta(x) = g(\underset{\underset{-3}{\shortparallel}}{\theta_0} + \underset{\underset{1}{\shortparallel}}{\theta_1} x_1 + \underset{\underset{1}{\shortparallel}}{\theta_2} x_2)$$

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

Non linear Decision Boundary



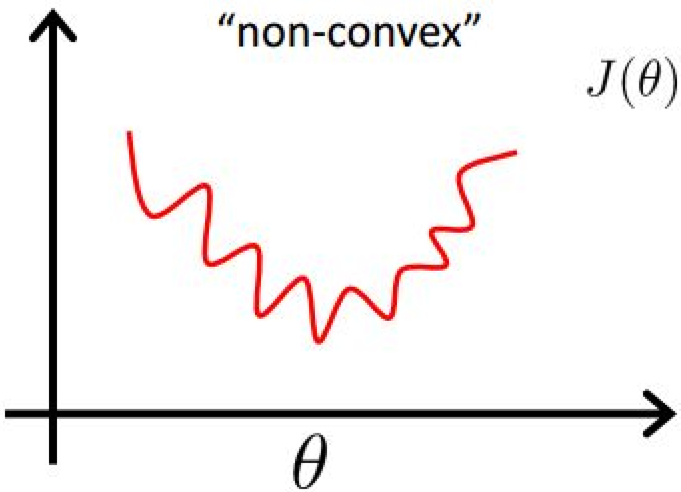
$$h_\theta(x) = g(\overset{-1}{\theta_0} + \overset{0}{\theta_1} x_1 + \overset{0}{\theta_2} x_2 + \underset{1}{\theta_3} x_1^2 + \underset{1}{\theta_4} x_2^2)$$

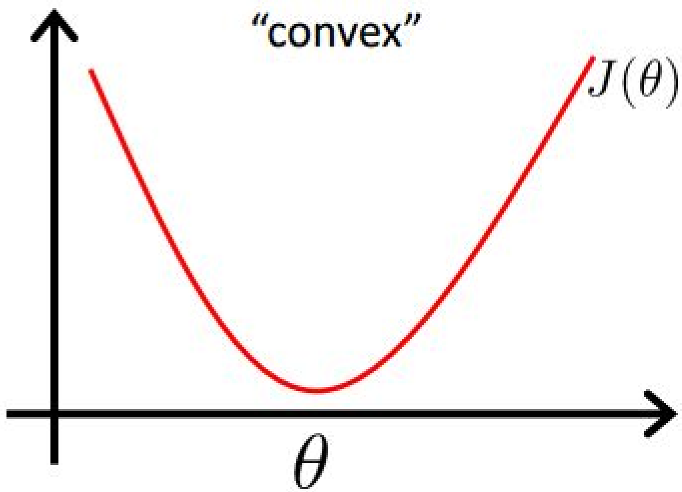
Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

Cost Function

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

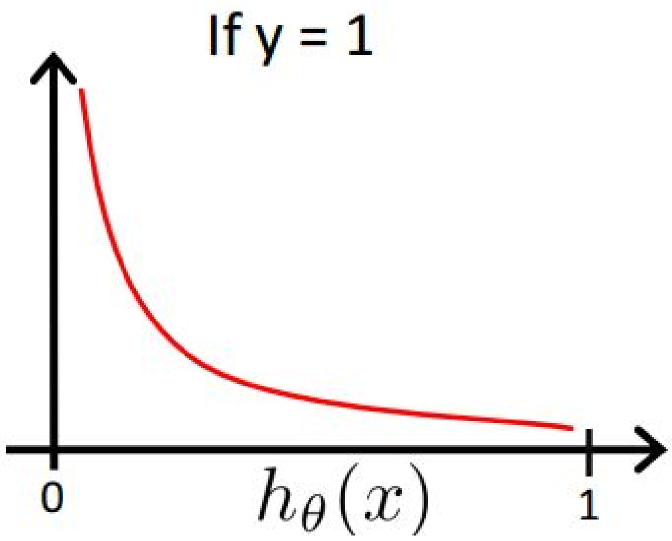
$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$





Cost Function Contd.

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



Cost $= 0$ if $y = 1, h_\theta(x) = 1$
But as $h_\theta(x) \rightarrow 0$
$Cost \rightarrow \infty$

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.

Cost Function Contd.

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

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

Gradient Descent

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

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

Repeat $\{$

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

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

(simultaneously update all θ_j)

$\}$

Advanced Optimization

Optimization algorithms:

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

Advantages:

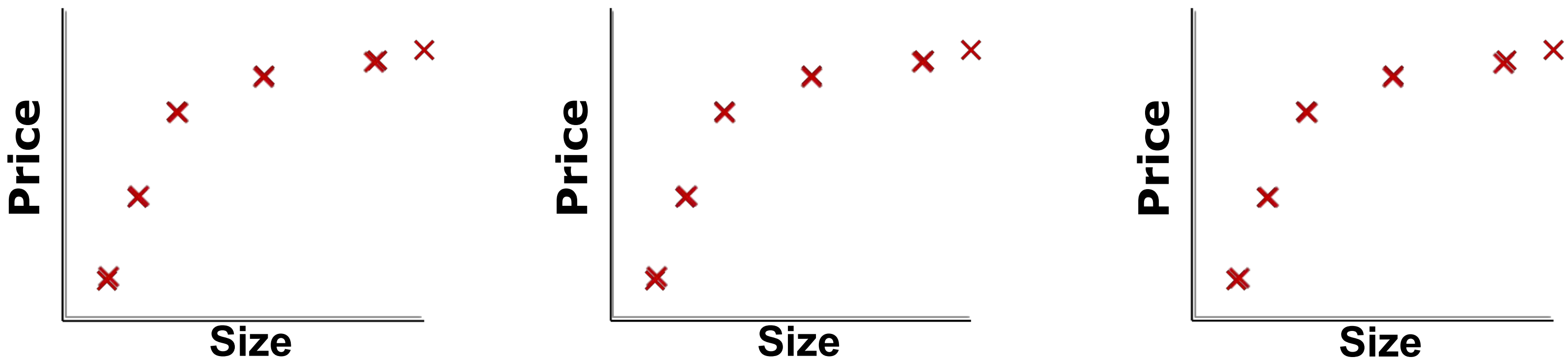
- No need to manually pick α
- Often faster than gradient descent.

Disadvantages:

- More complex

Overfitting Problem

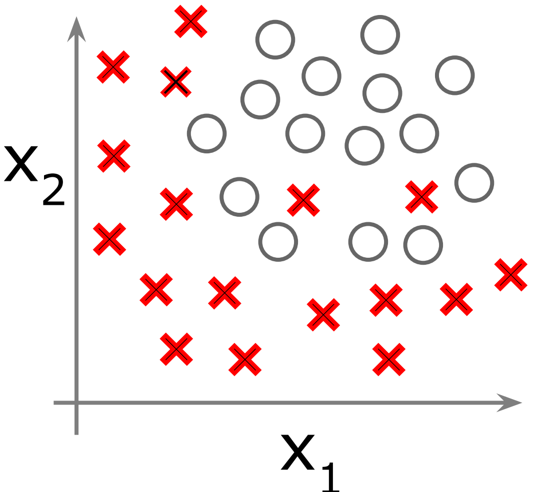
Example: Linear regression (housing prices)



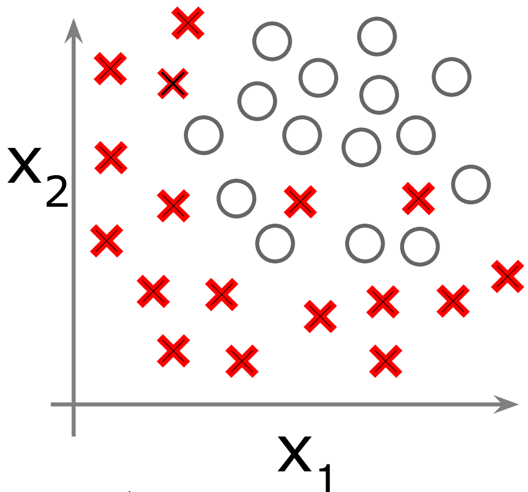
$\theta_0 + \theta_1 x$ $\qquad$ $\theta_0 + \theta_1 x + \theta_2 x^2$ $\qquad$ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \approx 0$), but fail to generalize to new examples (predict prices on new examples).

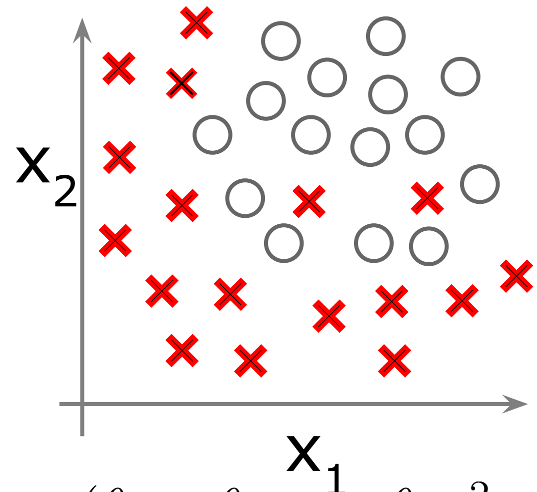
Example Logistic Regression



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



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$