On Regularization and Some Logistic Regression

Alexander G. Ororbia II

Introduction to Machine Learning

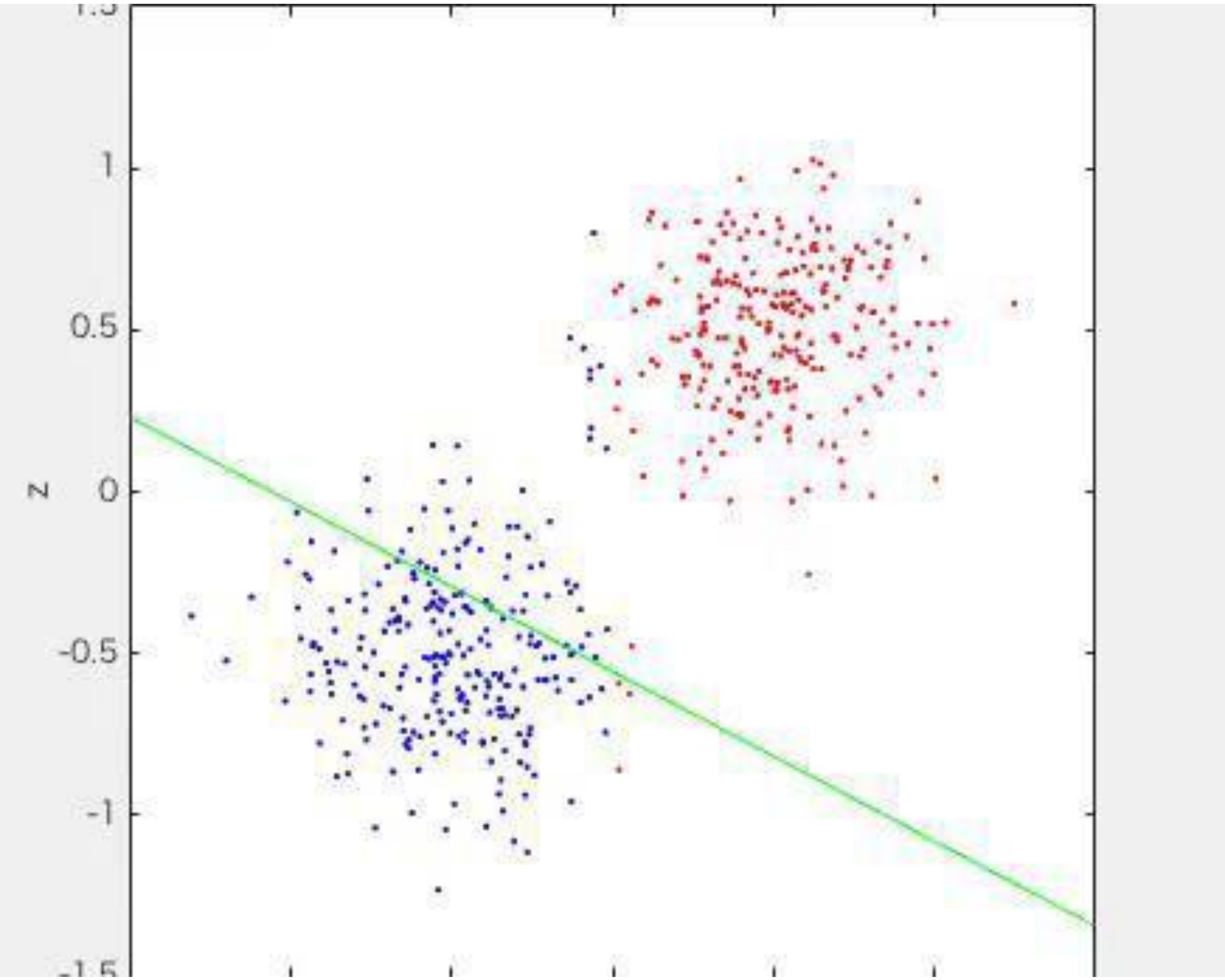
CSCI-635

10/2/2023

Special Thanks: Sargur N. Srihari, Andrew Ng, Carlos Guestrin

Machine Learning Algorithms

	Supervised Learning	**Unsupervised Learning**
Discrete	Classification	Clustering
Continuous	$\boxed{\text{Regression}}$	Dimensionality reduction


1.5
1
0.5
z
0
-0.5
-1
-1.5

Limiting Model Capacity

- Regularization has been used for decades prior to advent of deep learning
- Linear- and logistic-regression allow simple, straightforward and effective regularization strategies
  - Adding a parameter norm penalty $\Omega(\theta)$ to the objective function J :

$$\boxed{\tilde{J}(\boldsymbol{\theta}; X, y) = J(\boldsymbol{\theta}; X, y) + \alpha\Omega(\boldsymbol{\theta})}$$

    - where $\alpha\varepsilon[0,\theta)$ is a hyperparameter that weight the relative contribution of the norm penalty term Ω
      - Setting α to 0 results in no regularization. Larger values correspond to more regularization

Gradient of Regularized Objective

- Objective function (with no bias parameter)

$$\boxed{\tilde{J}(w;X,y) = \frac{\alpha}{2} w^T w + J(w;X,y)}$$

- Corresponding parameter gradient

$$\boxed{\nabla_w \tilde{J}(w;X,y) = \alpha w + \nabla_w J(w;X,y)}$$

- To perform single gradient step, perform update:

$$\boxed{\boldsymbol{w} \leftarrow \boldsymbol{w} - \varepsilon\left(\alpha \boldsymbol{w} + \nabla_{\boldsymbol{w}} J\left(\boldsymbol{w}; X, \boldsymbol{y}\right)\right)}$$

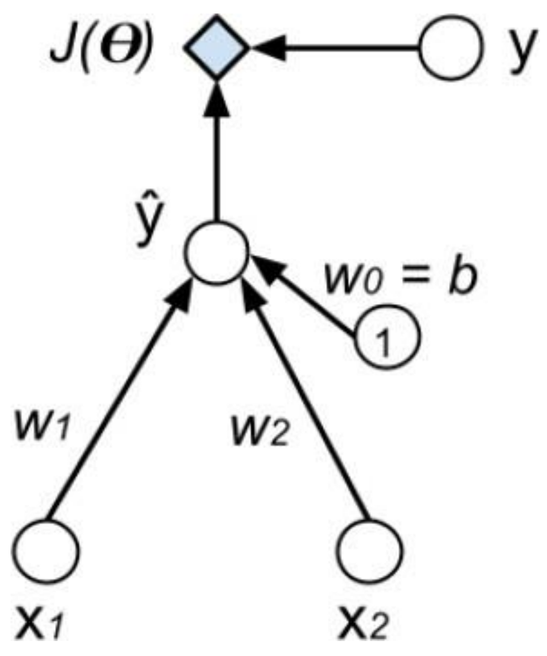
- Written another way, the update is

$$\boxed{\boldsymbol{w} \leftarrow (1 - \varepsilon\alpha)\boldsymbol{w} - \varepsilon\nabla_{\boldsymbol{w}} J\left(\boldsymbol{w}; X, \boldsymbol{y}\right)}$$

  – We have modified learning rule to shrink $\boldsymbol{w}$ by constant factor 1-$\varepsilon\alpha$ at each step

Multivariate Regressor Architecture

$$f_{\Theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$ **Hypothesis!**



Cost:

$$\mathcal{J}(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}(f_{\Theta}(x^i) - y^i)^2 + \frac{\beta}{2m}\sum_{j=1}^{n}\theta_j^2$$

Derivative/Update:

$$\frac{\partial \mathcal{J}(\Theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(f_{\Theta}(x^i) - y^i)x_j^i + \frac{\beta}{m}\theta_j$$

Optimizer:

$$\theta_j = \theta_j - \alpha\frac{\partial \mathcal{J}(\Theta)}{\partial \theta_j} = \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(f_{\Theta}(x^i) - y^i)x_j^i,\ j = 0, 1, 2, \cdots, n.$$

Machine Learning Algorithms

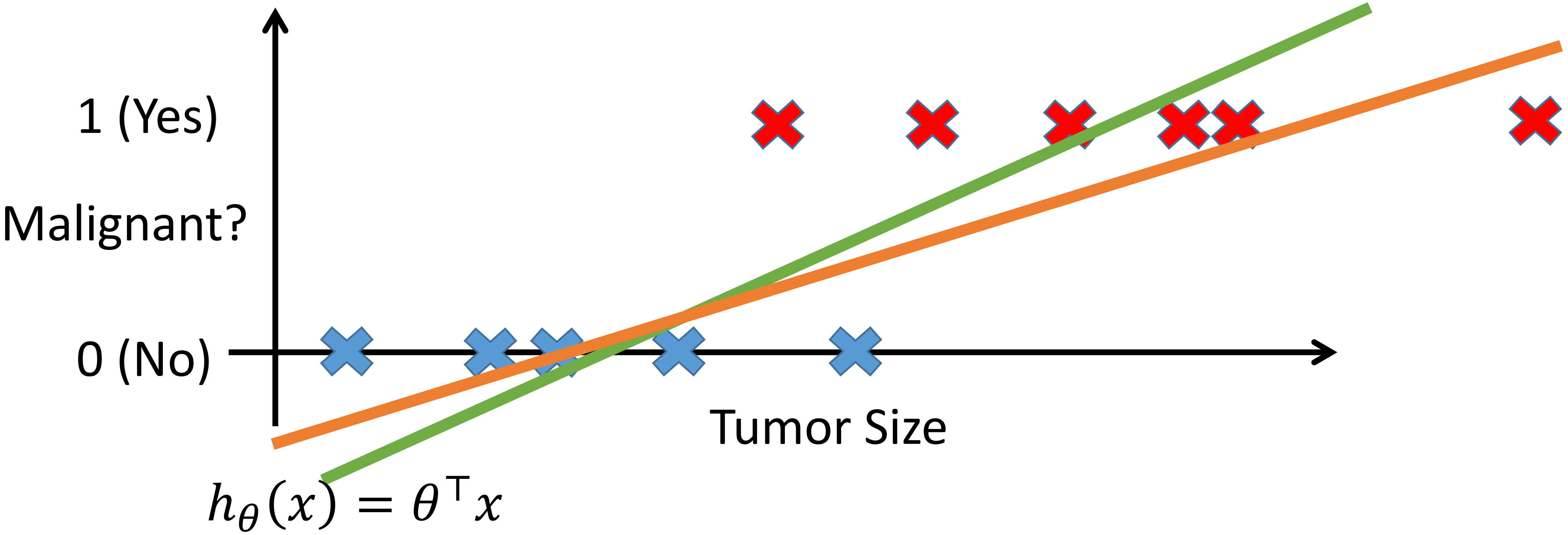
	Supervised Learning	**Unsupervised Learning**
Discrete	$\boxed{\text{Classification}}$	Clustering
Continuous	$\boxed{\text{Regression}}$	Dimensionality reduction

Logistic Regression

- Hypothesis representation
- Cost function
- Logistic regression with gradient descent
- Regularization
- Multi-class classification

Logistic Regression

- **Hypothesis representation**
- Cost function
- Logistic regression with gradient descent
- Regularization
- Multi-class classification



- 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 "$y = 0$"

Why use Logistic Regression?

- There are many important research topics for which the dependent variable is "limited"
- For example: whether or not a person smokes/drinks/skips class/takes advanced mathematics
  - For these, outcome is not continuous or distributed normally
    - Example: Are mothers who have high school education less likely to have children with IEP’s (individualized plans, indicating cognitive or emotional disabilities)
- Binary logistic regression is a type of regression analysis where dependent variable is a dummy variable:
  coded 0 (*negative class*: did not smoke) or 1 (*positive class*: did smoke)

Classification: $y = 1$ or $y = 0$

$h_\theta(x) = \theta^\top x$ (from linear regression)
can be > 1 or < 0

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

Logistic regression is actually for **classification**

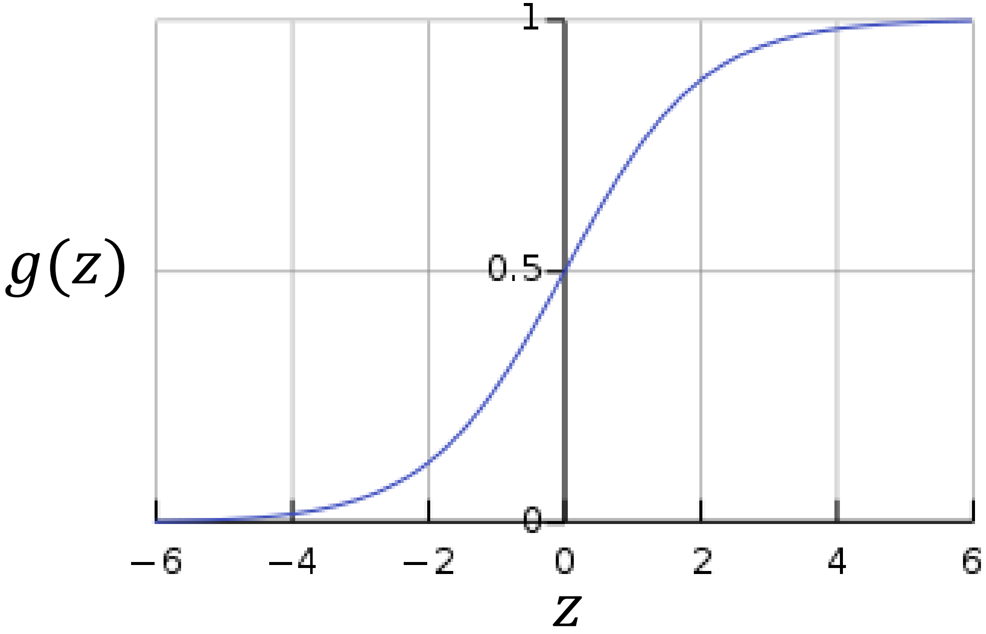
Hypothesis Representation

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

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

- $h_\theta(x) = g(\theta^\top x)$,

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



- Sigmoid function
- Logistic (link) function

Interpretation of Hypothesis Output

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

- Example: If $x = \begin{bmatrix} x_0 \\ \mathrm{x}_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \mathrm{tumorSize} \end{bmatrix}$
- $h_\theta(x) = 0.7$

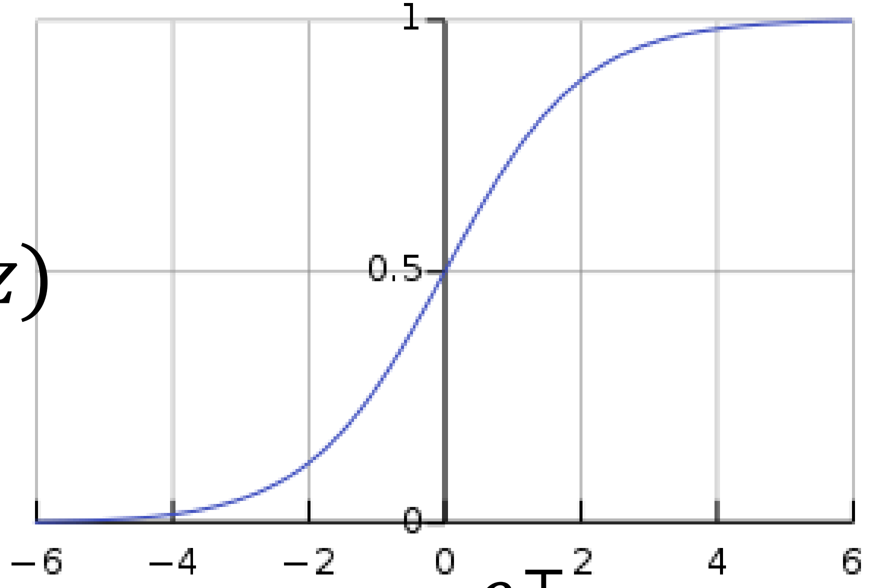
- Tell patient that 70% chance of tumor being malignant

Logistic Regression

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

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

Log-linear model



$g(z)$

$z = \theta^\top x$

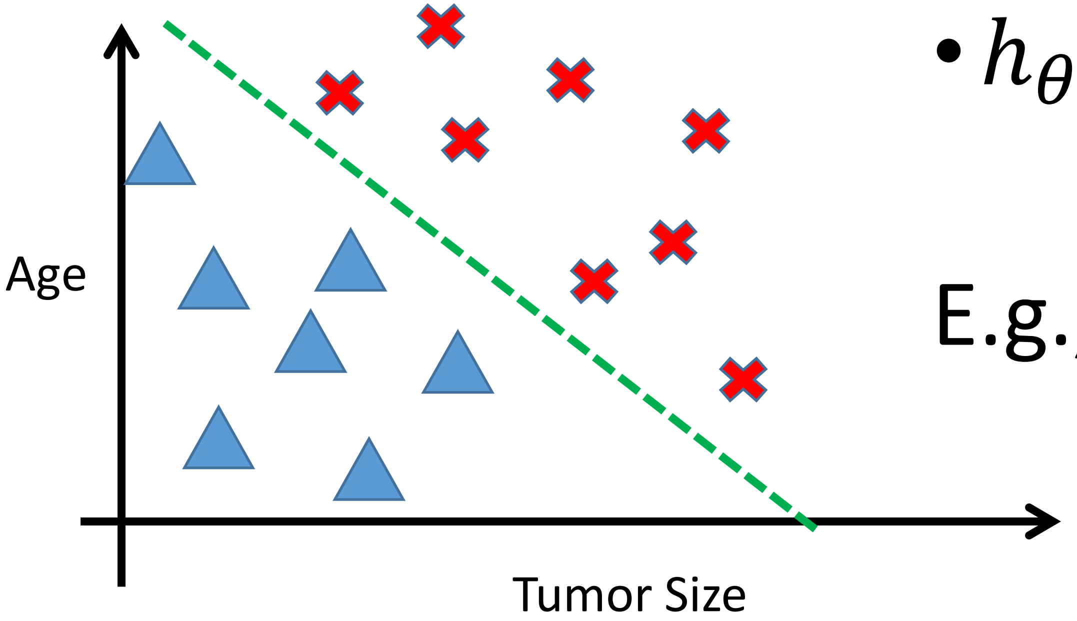
Suppose we predict "y = 1" if $h_\theta(x) \geq 0.5$

$$z = \theta^\top x \geq 0$$

we predict "y = 0" if $h_\theta(x) < 0.5$

$$z = \theta^\top x < 0$$

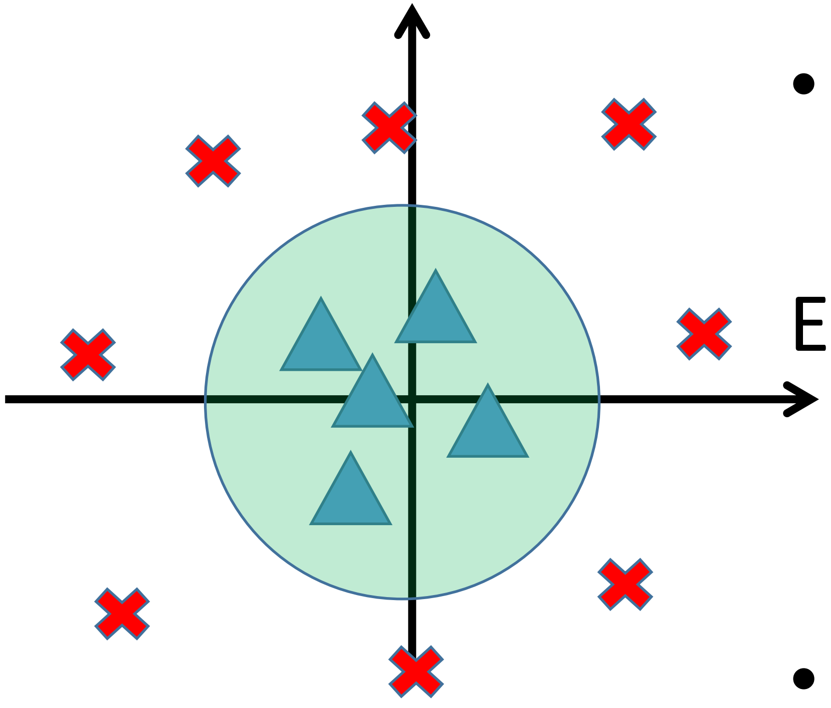
Decision Boundary



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

E.g., $\theta_0 = -3,\ \theta_1 = 1,\ \ \theta_2 = 1$

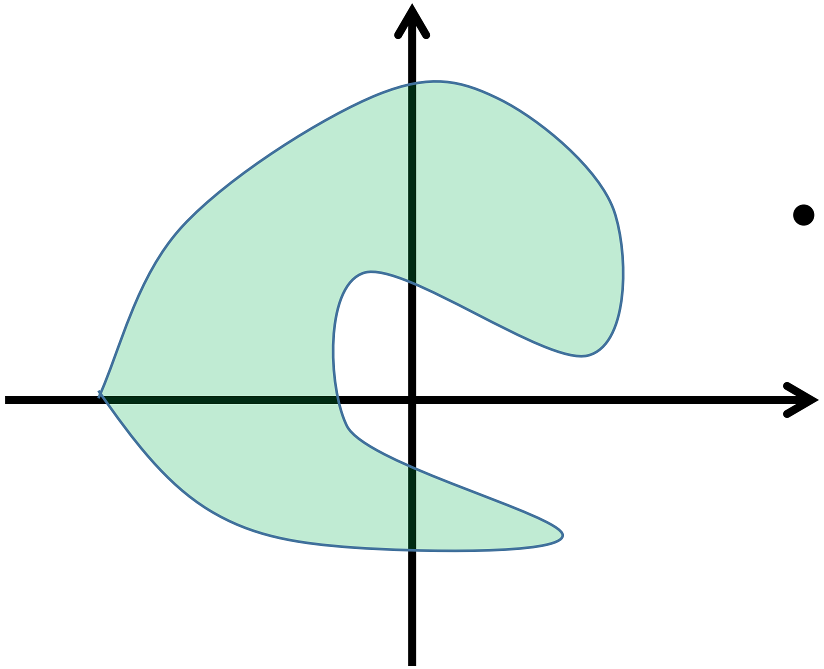
- Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$



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

E.g., $\theta_0 = -1, \theta_1 = 0, \theta_2 = 0, \theta_3 = 1, \theta_4 = 1$

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



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

Where Does the Form Come From?

- Logistic regression hypothesis representation

$$h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}}$$

- Consider learning f: $X \to Y$, where
  - X is a vector of real-valued features $[X_1, \cdots, X_n]^\top$
  - Y is Boolean
  - Assume all X_i are ***conditionally independent*** given Y
  - Model $P(X_i|Y = y_k)$ as Gaussian $N(\mu_{ik}, \sigma_i)$
  - Model $P(Y)$ as Bernoulli π

Where Does the Form Come From?

- Logistic regression hypothesis representation

$$h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}}$$

- Consider learning f: $X \rightarrow Y$, where
  - X is a vector of real-valued features $[X_1, \cdots, X_n]^\top$
  - Y is Boolean
  - Assume all X_i are ***conditionally independent*** given Y
  - Model $P(X_i|Y = y_k)$ as Gaussian $N(\mu_{ik}, \sigma_i)$
  - Model $P(Y)$ as Bernoulli π

What is $P(Y|X_1, X_2, \cdots, X_n)$?

Questions?