Logistic Regression

DATA MINING

LI ZENG

Contact: L.Zeng@tilburguniversity.edu

Learning tasks

- Prediction
  - To fit a shape that gets as close to the data as possible

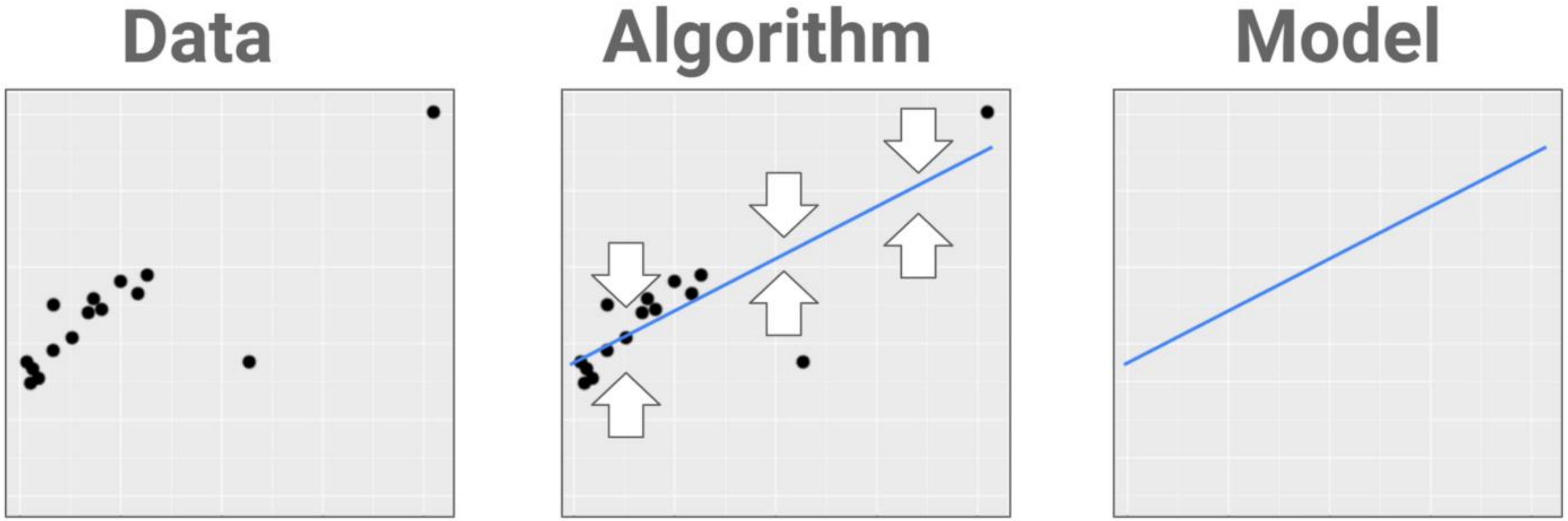


Image source: towardsdatascience.com/

Learning tasks

- Classification
  - To separate the data into several classes
  - When the output/response is categorical

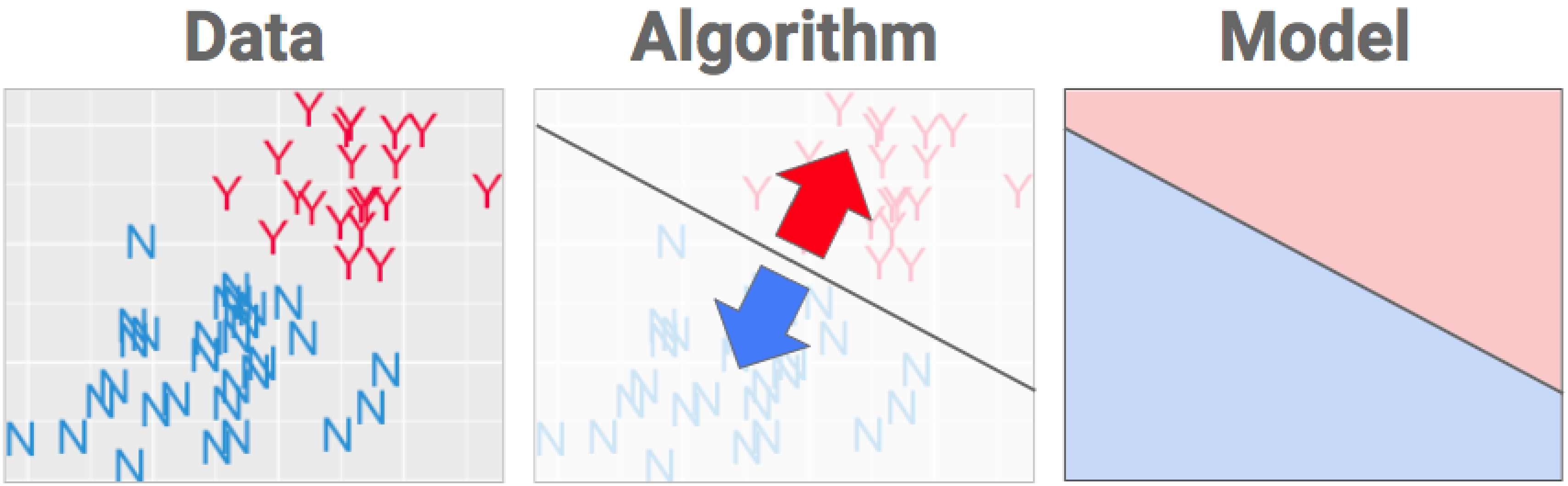


Image source: towardsdatascience.com/

Classification

- Classification
  - Malware classification
  - Email spam or not
  - Customer churn prediction
  - Tumor detection

$Y = \{0,1\}$ – binary classification

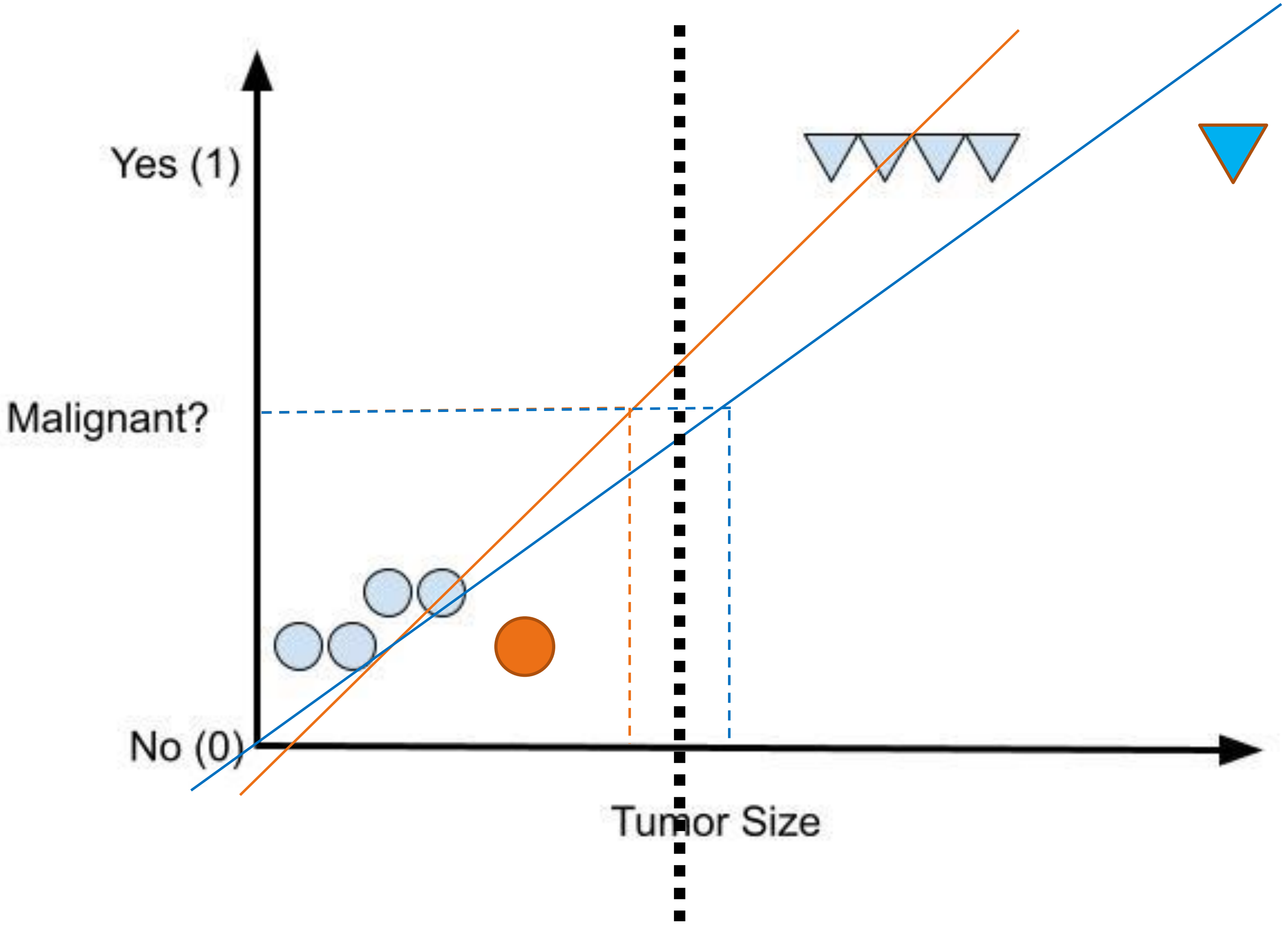
  - Object classification – apples, oranges, bananas, etc.

$Y = \{0,1,2,3, \ldots\}$ – multiple classification

Example: Tumor Classification

Threshold classifier output:

$h_\theta(x) = 0.5$



$$y = \begin{cases} 1, h_\theta(x) < 0.5 \\ 0, h_\theta(x) \geq 0.5 \end{cases}$$

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

Logistic Regression

To have: $0 \leq h_\theta(x) \leq 1$

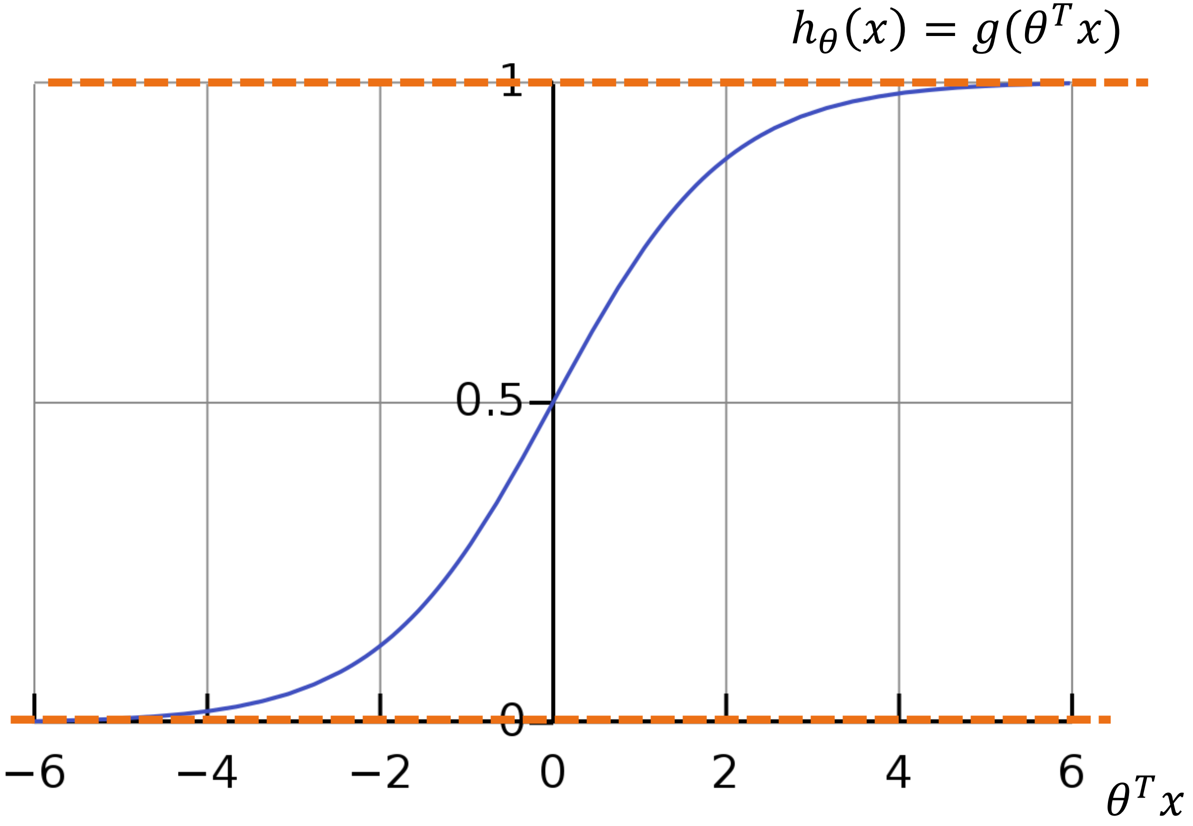
$$h_\theta(x) = g(\theta^T x)$$

$$g\ (z) = \frac{1}{1+e^{-z}},\ z = \theta^T x = \beta_0 + \beta_i x_i$$

Sigmoid/logistic function

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

or $\frac{e^z}{1+e^z}$



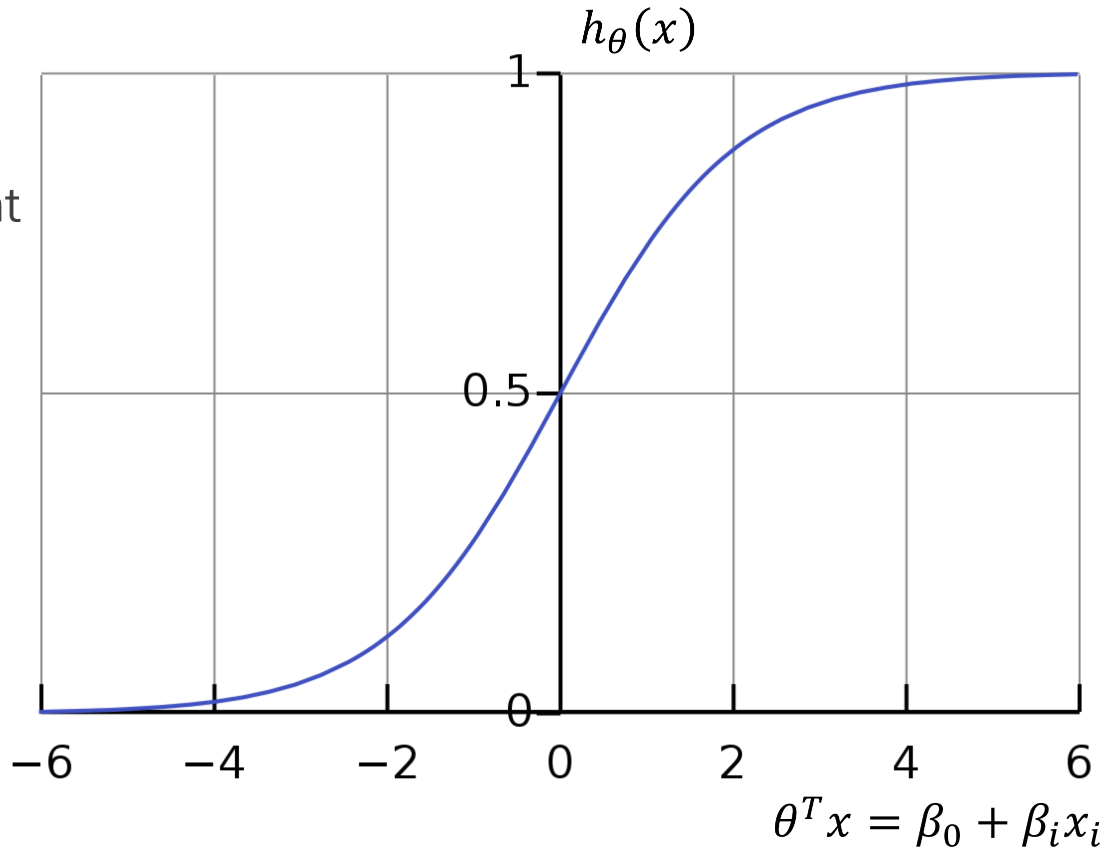
Interpretation of Hypothesis Output

- $h_\theta(x)$ =estimated probability that y=1 on input x
- In our previous tumor detection example,
- If $h_\theta(x)$=0.7 → 70% chance of tumor being malignant

$$x = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 \\ tumor\ size \end{pmatrix}$$

- Probability that y=1, given x, parameterized by θ

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



Logit function

A back-ward transformation of the sigmoid function to calculate the linear scores starting from the prediction probabilities.

The logit function maps probabilities from the range (0,1) to the entire real number range $(-\infty,\infty)$

$$logit\left(h_\theta(x)\right) = \log\left(\frac{g(\theta^T x)}{1 - g(\theta^T x)}\right) = \theta^T x$$

$$logit\left(p_{y=1}\right) = \log\left(\frac{p_{y=1}}{1 - p_{y=1}}\right) = \log\left(\frac{p_{y=1}}{p_{y=0}}\right)$$

Interpretation of coefficients

- The coefficient in logistic regression is the expected change in log odds of having the outcome per unit change in x
- The intercept is the **expected log-odds ratio** in favor of Class 1 over Class 0 when all the features are equal to zero (if exponentiated, it is the **expected odds-ratio**)
- Increasing the predictor by 1 unit (or going from one level to the next – categorical variables) multiples the odds of having the outcome by e^{θ}

	Coefficient	Standard Error	P-value
Intercept	-1.93	0.13	<0.001
Smoking	0.38	0.17	0.03

Interpretation of coefficients

- $\theta = 0.38, e^{0.38}$=1.46 - odds ratio that associates smoking to the risk of heart disease
- The smoking group has a 1.46 times the odds of the non-smoking group of having heart disease
- Alternatively, the smoking group has 46% more odds of having heart disease than the non-smoking group

	Coefficient	Standard Error	P-value
Intercept	-1.93	0.13	<0.001
Smoking	0.38	0.17	0.03

Interpretation of coefficients

- For negative coefficients?
- $If\ \theta = -0.38, e^{-0.38}$=0.68
- Smoking is associated with a 32% (1-0.68) reduction in the relative risk of heart disease

	Coefficient	Standard Error	P-value
Intercept	-1.93	0.13	<0.001
Smoking	0.38	0.17	0.03

Interpretation of intercept

- The intercept is -1.93
- It should be interpreted assuming a value of 0 for all the predictors in the model.
- If we put back the value to the logistic model, we get 0.13
- The probability that a non-smoker will have a heart disease in the next 10 years is 0.13
- Without calculating this probability, can we learn something from the sign of intercept?
  - If the intercept is negative:
  - If the intercept is positive:
  - If the intercept is equal to 0:

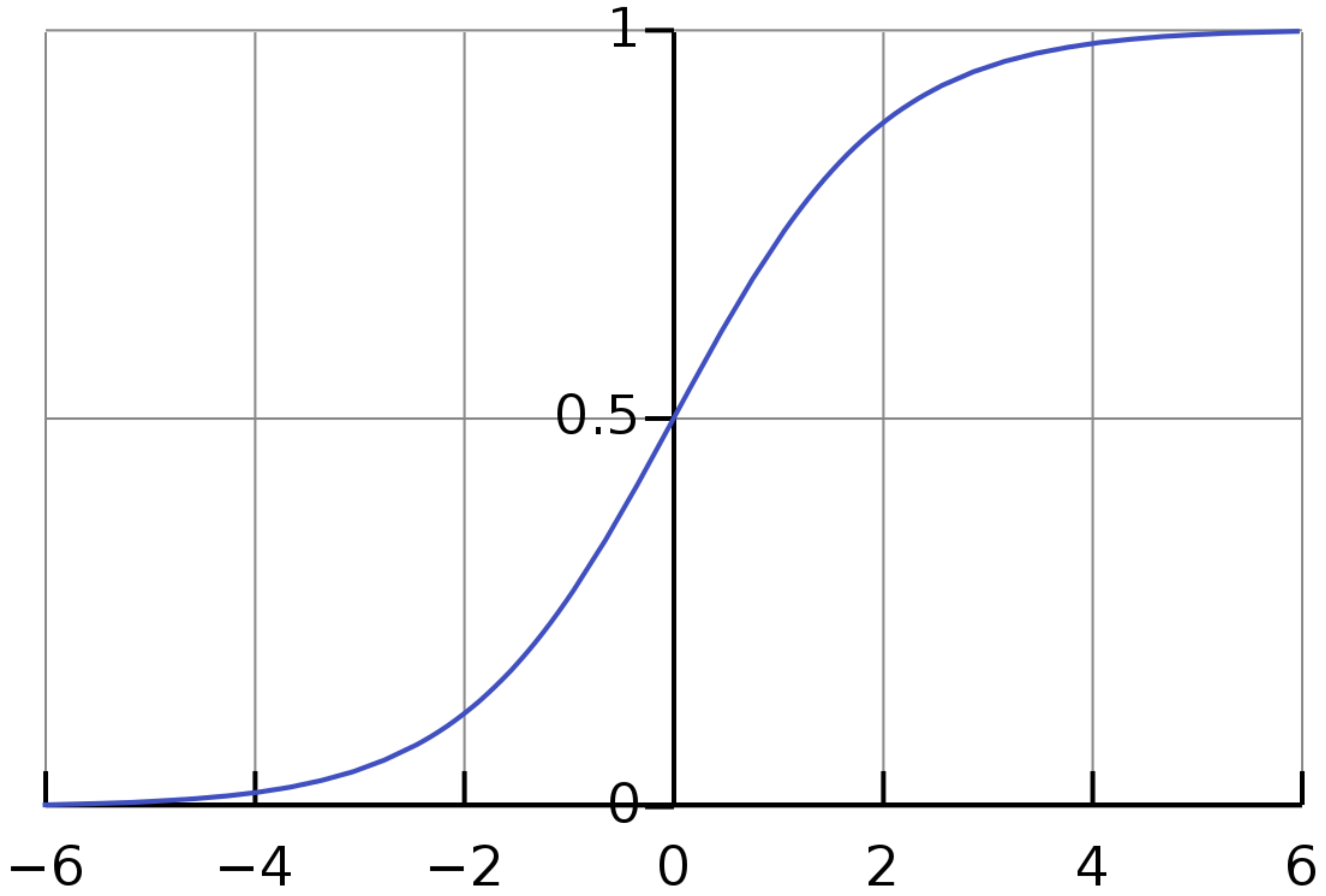
Interpretation of SE

- $SE = 0.17, e^{(\theta \pm 2SE)} = e^{(0.38 \pm 2\times 0.17)} = [1.04, 2.05]$
- We are 95% confident that smokers have on average 4% to 105% more odds of having heart disease than non-smokers

Decision Boundary

Predict y = 1:

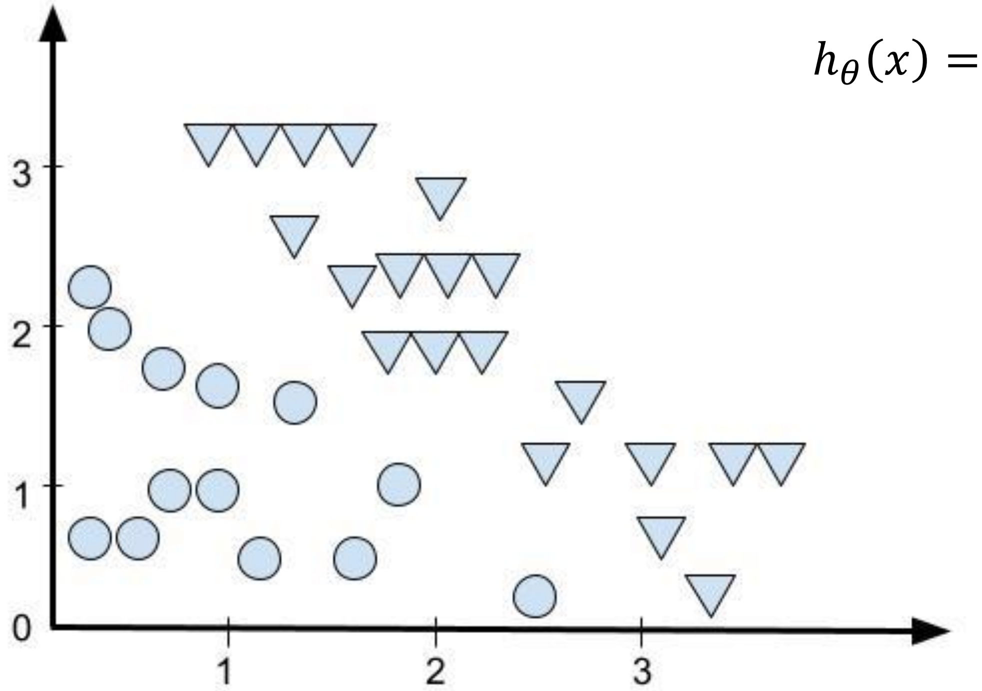
Predict y = 0:



Decision Boundary

Predict y = 1:

Predict y = 0:



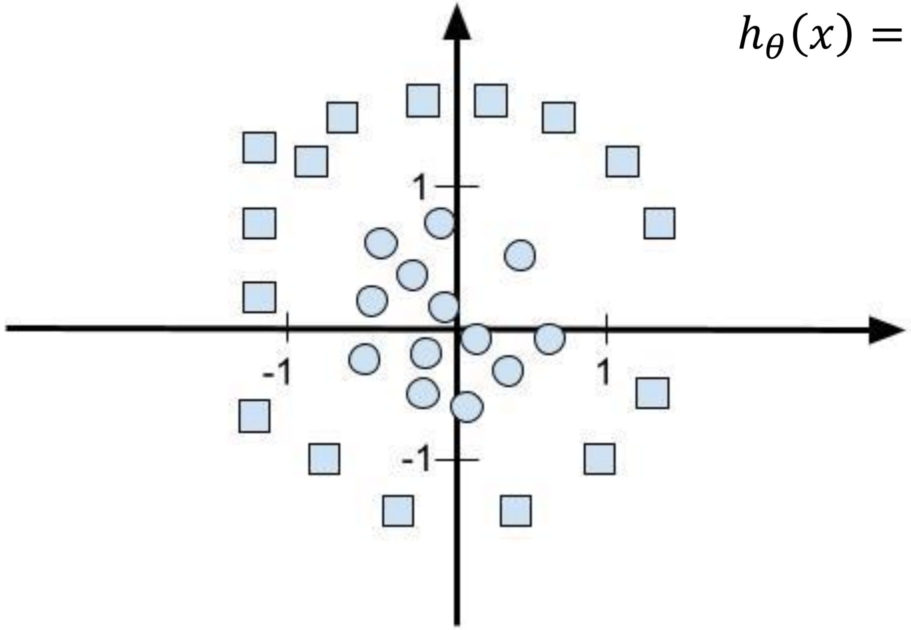
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

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

Decision Boundary

Predict y = 1:

Predict y = 0:



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

$$\theta = [-1,0,0,1,1]^T$$

Cost Function

- Recall:
- For linear regression, we try to minimize:

$$\min_{\beta} \sum_{i}^{n} [y_i - \hat{y}_i(\beta)]^2$$

Nice convex property when use optimization techniques (e.g. gradient descent) for parameters

- For logistic regression:

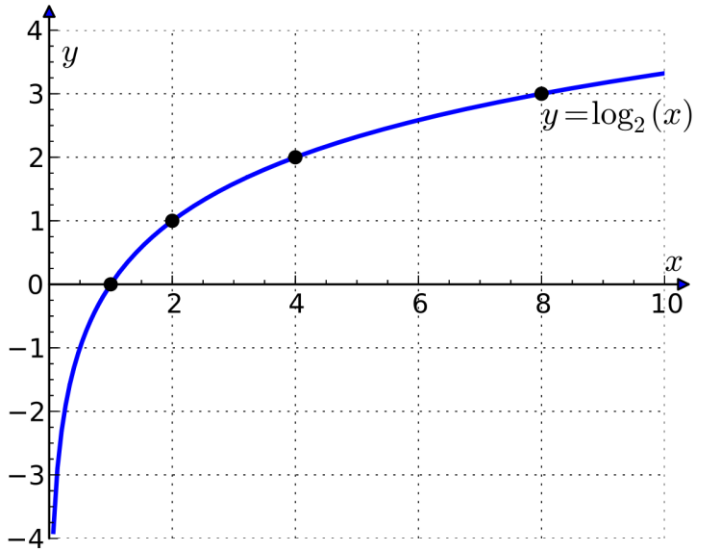
$$[(h_\vartheta(x) - y)]^2$$

Cost Function

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

If y=1

$$h_\vartheta(x) = 1, Cost = 0$$
$$\text{As } h_\vartheta(x) \to 0, Cost \to \infty$$

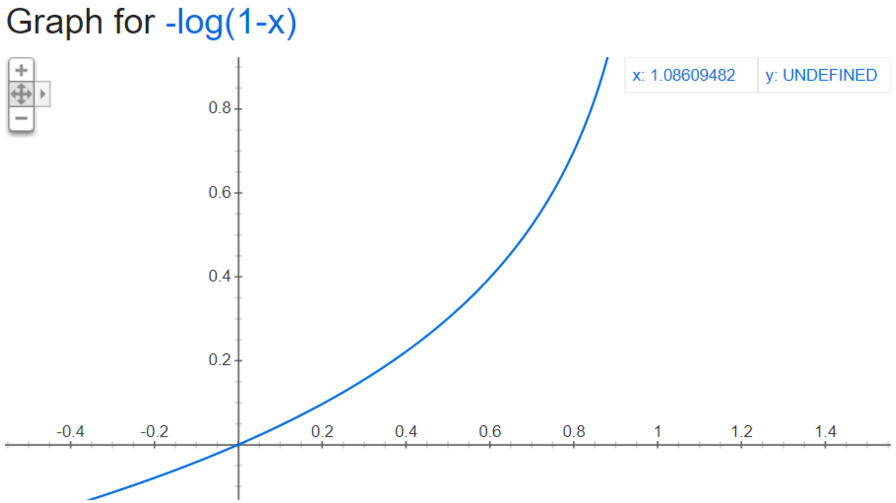


Cost Function

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

If y=0

$$h_\vartheta(x) = 0, Cost = 0$$
$$\text{As } h_\vartheta(x) \rightarrow 1, Cost \rightarrow \infty$$



Cost Function

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

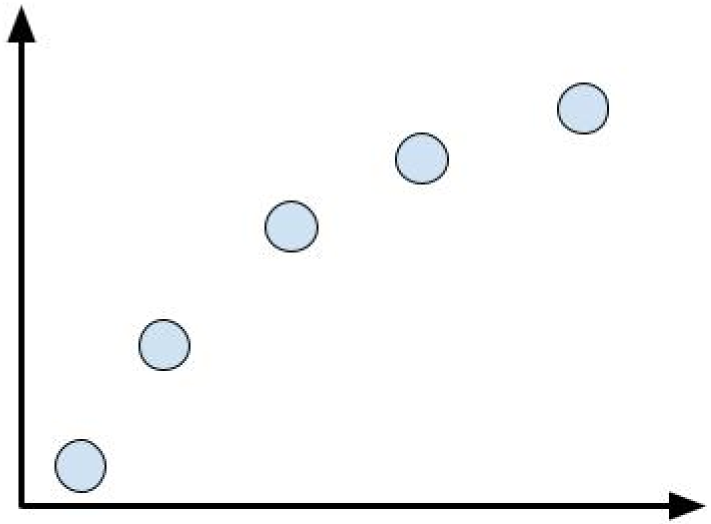
$$Cost(h_\vartheta(x), y) = -y\log\big(h_\vartheta(x)\big) - (1 - y)\log\big(1 - h_\vartheta(x)\big)$$

Overfitting?

- Reduce number of features
  - Manually select which features to keep
  - Model selection
- Regularization
  - Keep all the features, but reduce magnitude of parameters θ_j
  - Works well when we have a lot of features, each of which contributes a bit to the predictions

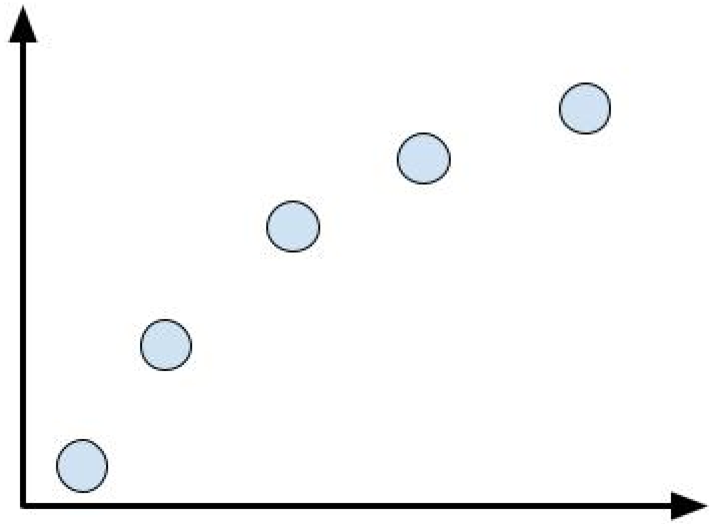
Regularization

$$\min_{\beta} \sum_{i}^{n} [y_i - \hat{y}_i(\beta)]^2 + 1000\theta_3^2 + 1000\theta_4^2$$



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

We try to penalize and make $\boldsymbol{\theta_3}$ and $\boldsymbol{\theta_4}$ small



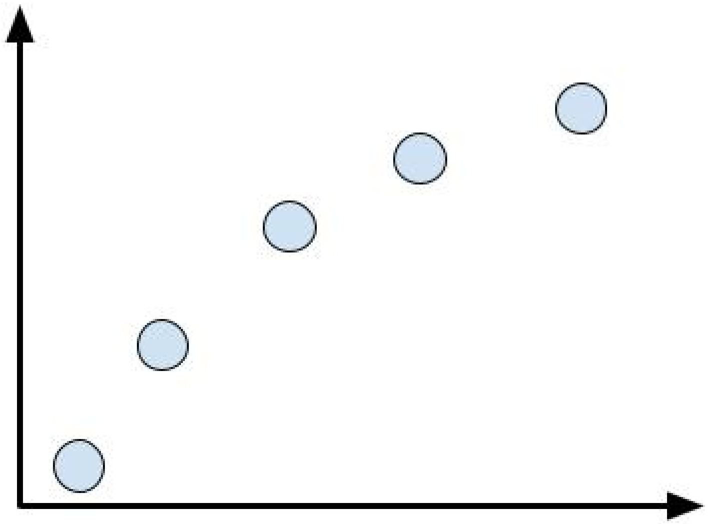
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Regularization

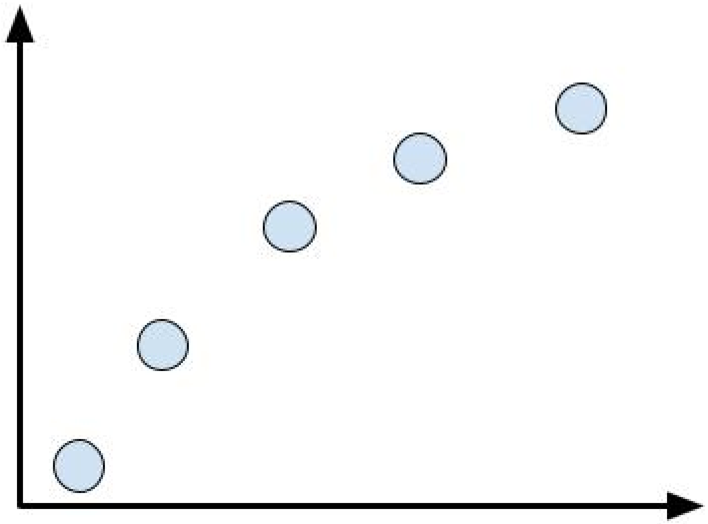
$$\min_{\beta} \sum_{i}^{n} [y_i - \hat{y}_i(\beta)]^2 + \lambda \sum \beta^2$$

Small values for parameters:

- Simpler hypothesis
- Less prone to overfitting



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

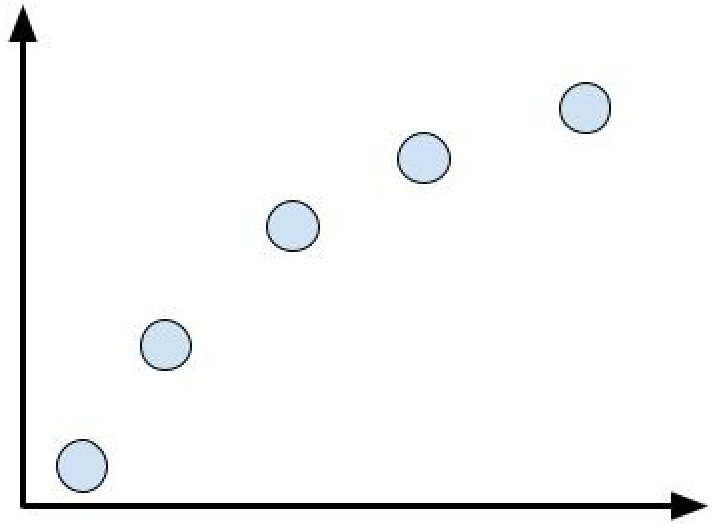


$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Regularization

What if λ is set to an extremely large value?

$$\min_{\beta} \sum_{i}^{n} [y_i - \widehat{y}_i(\beta)]^2 + \lambda \sum \beta^2$$



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

Ridge

- Ridge regression simply adds the $l_2$norm of the regression coefficients

$$\min_{\beta} \sum_{i}^{n} [y_i - \widehat{y}_i(\beta)]^2 + \lambda \sum \beta^2$$

- Ridge regression works well for stabilizing predictions and coefficients estimates.
  - All variables stay in the model and are shrunk proportionately
  - Coefficients lose interpretability

LASSO

- The least absolute shrinkage and selection operator (LASSO) regularizes and implements automatic variable selection

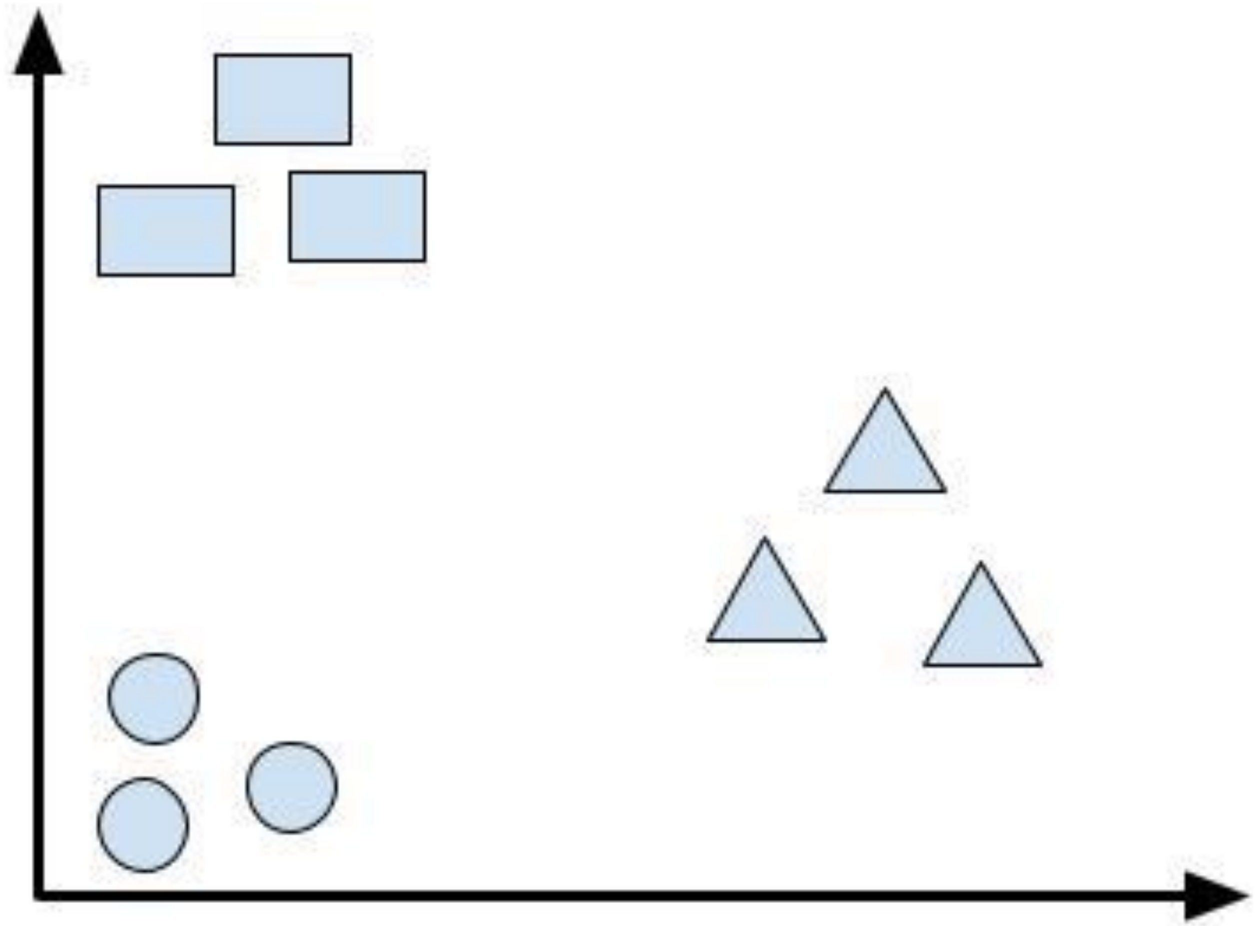
$$\min_{\beta} \sum_{i}^{n} [y_i - \widehat{y}_i(\beta)]^2 + \lambda \sum |\beta|$$

- In the LASSO solution, the coefficients can be set to zero. Therefore, LASSO may also induce sparsity.

Multiclass Classification

1. Train a logistic regression classifier for each class i to predict the probability that y=I
2. On a new input x, pick the class i with the maximum probability

1. ONE-VS-ALL (ONE-VS-REST)

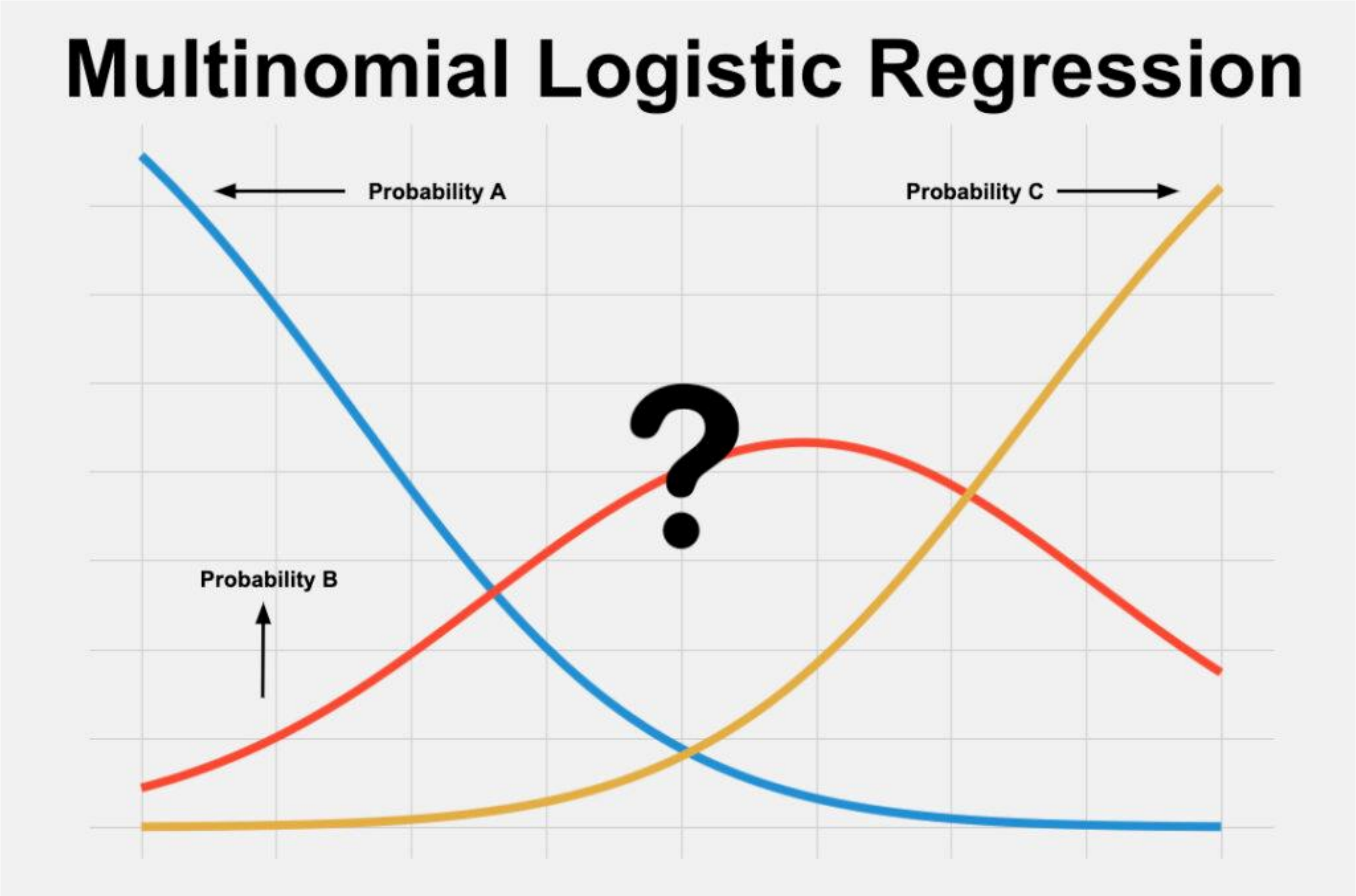


Multiclass Classification

Assumption:

1. Linearity

2. MULTINOMIAL LOGISTIC REGRESSION

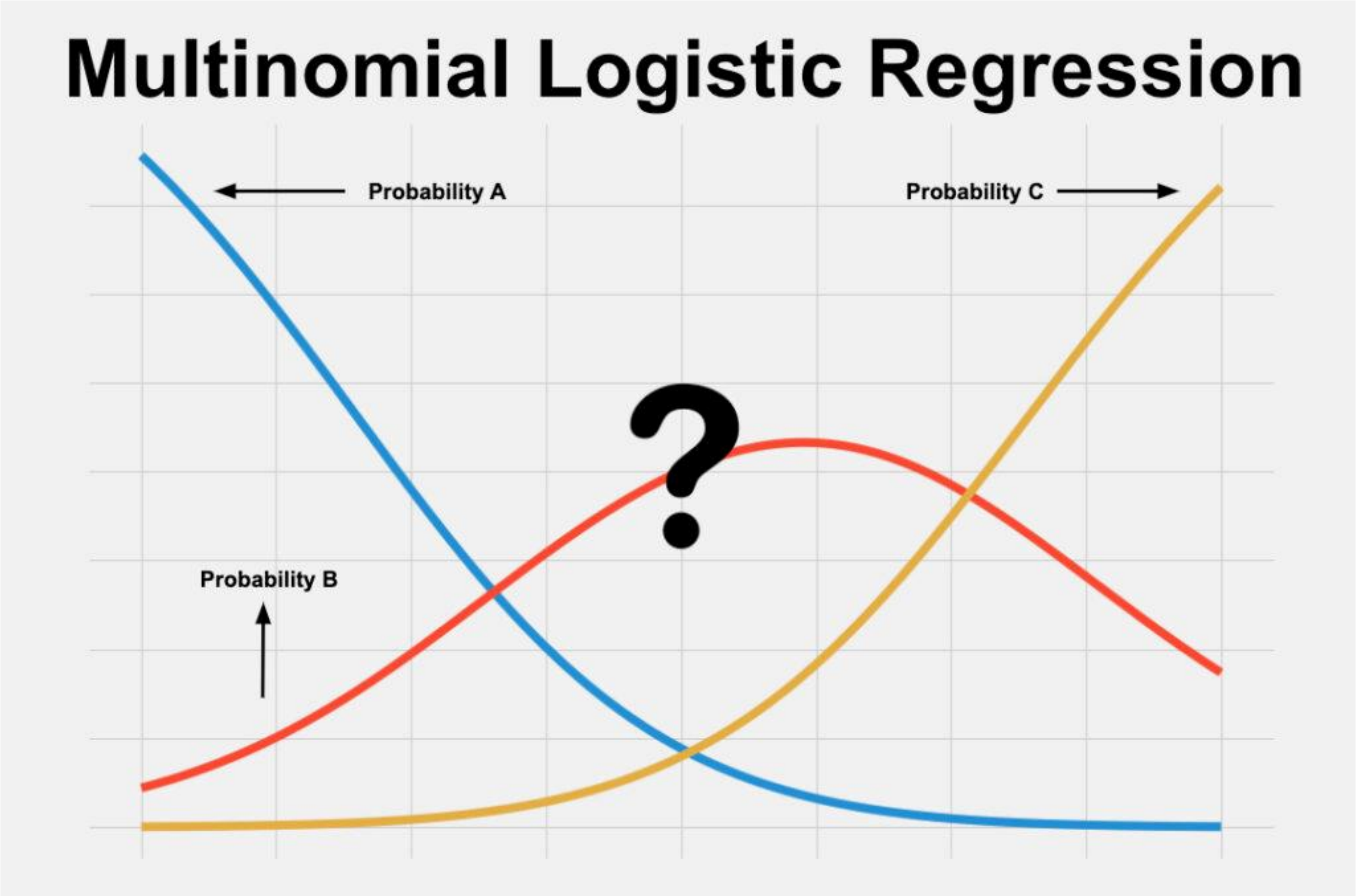


Multiclass Classification

Assumption:

1. Linearity
2. No Outliers

2. MULTINOMIAL LOGISTIC REGRESSION

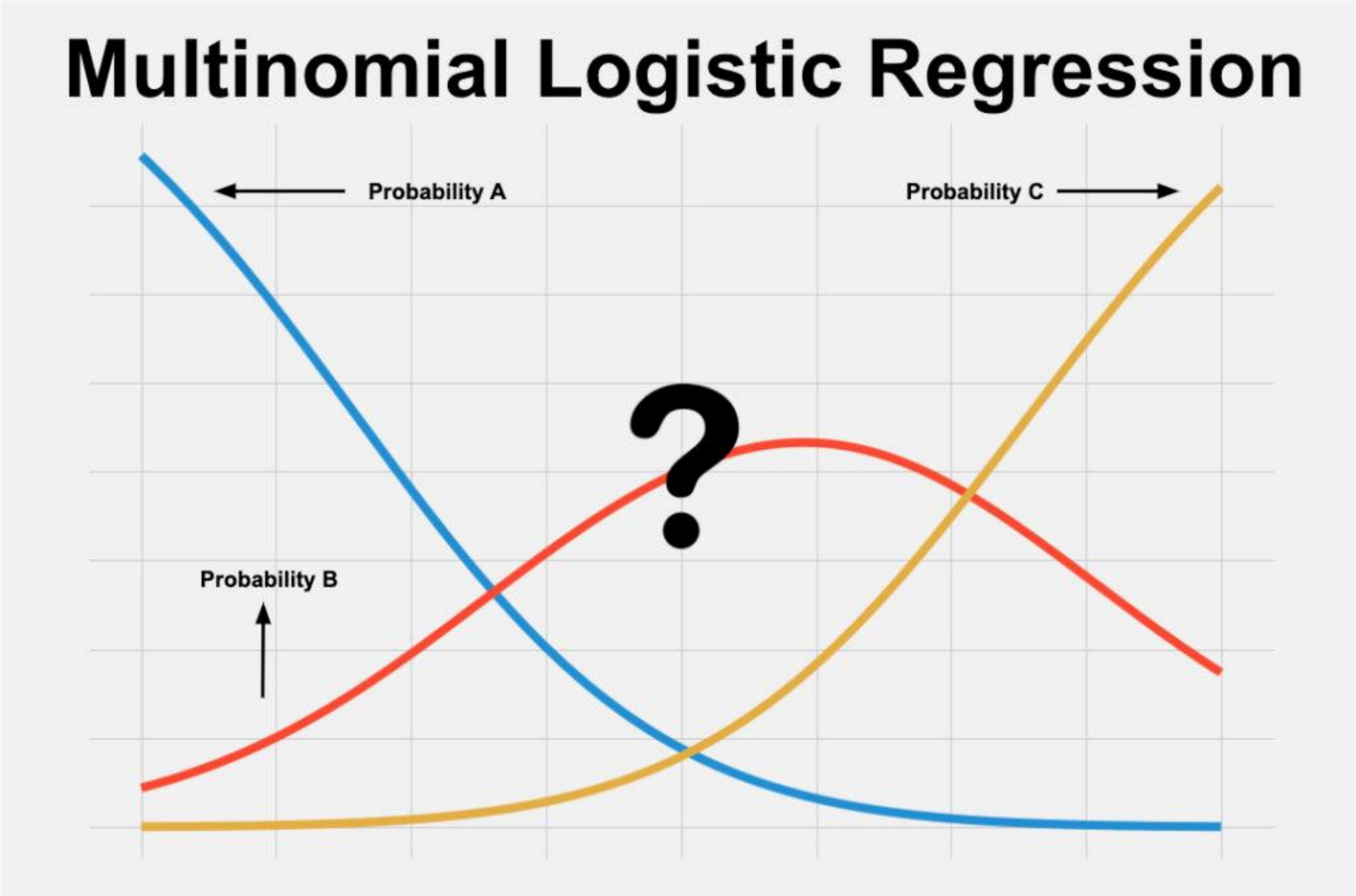


Multiclass Classification

Assumption:

1. Linearity
2. No Outliers
3. Independence

2. MULTINOMIAL LOGISTIC REGRESSION



Multiclass Classification

Assumption:

1. Linearity
2. No Outliers
3. Independence
4. No Multicollinearity

2. MULTINOMIAL LOGISTIC REGRESSION