High Level Overview
of
Data Science & Machine Learning

Hrishikesh Bhatkhande

19-Dec-2017

Agenda

1	*Data, Big Data & Data Analytics*
2	*Components of Data Science*
3	*Machine Learning Overview*
4	*Linear & Logistic Regression and another Classifier algorithm*
5	*Appendix*

About Me

Hrishikesh Bhatkhande, PMP®

Contact #: +91 7028184981
hrishikesh.bhatkhande@gmail.com

LinkedIn - https://www.linkedin.com/in/hrishikesh-bhatkhande-pmp-341b8421
Kaggle - https://www.kaggle.com/hrishikesh312
GitHub - https://github.com/Hrishikesh312/

Career So Far

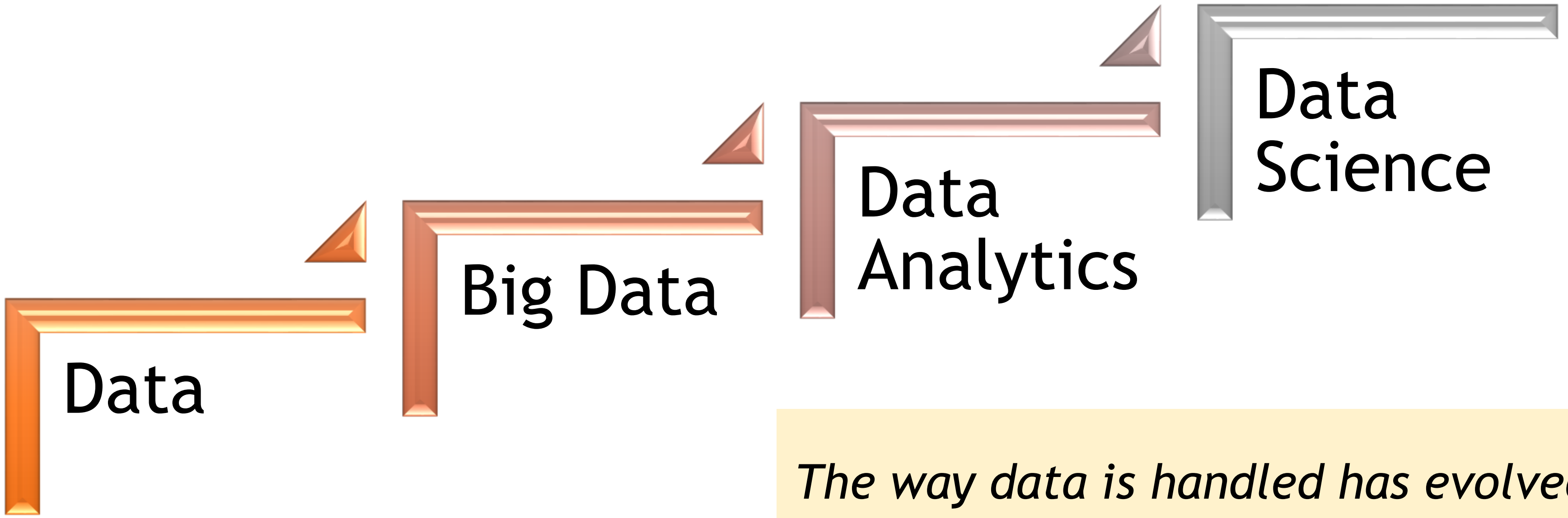
- ✓ IT Professional with 13+ years of experience Delivering and Managing IT projects primarily in Healthcare domain working in India and USA for Infosys Ltd and Syntel Pvt Ltd.
- ✓ Have managed multi-million dollar projects as part of well-known Healthcare programs in USA such as Healthcare-Reform (Obamacare), COB initiative
- ✓ Primary expertise in end-to-end Project Management, Delivery Management & Account Management

Data Science

- ✓ A Data Science enthusiast trained in R, Python, Statistics, Machine Learning, NLP, Hadoop & Spark frameworks aspiring to contribute in the field of Data Science
- ✓ Recently completed a Post Graduate Program (PGP) in Data Science, Business Analytics & Big Data at Aegis School of Data Science in association with IBM
- ✓ Prize Winner at Techgig Data Science competition - Bagged a consolation prize among ~3000 entries

Data, Big Data & Data Analytics

Data, Big Data & Data Analytics



The way data is handled has evolved with the increase in ***Scale*** *and* ***Complexity*** *of data*

- *As of 2016, 90% of the data in the world until then had been created in the last two years alone, at 2.5 quintillion bytes of data a day!*
- *It was expected that 2017 would create more data than ever before*

Data, Big Data & Data Analytics

Data	Big Data
❑ Data maintained in multiple separate systems and multiple formats ❑ Generally structured Data (column and rows structure) ❑ Typically data comes from traditional sources ❑ Data-warehousing for analytics and reporting	❑ Huge amount of data ❑ Typically doesn't follow any structure ❑ Mostly raw and non-transformed data with no roadmap ❑ Data also comes from non-traditional sources such as social media ❑ Characterized by the V's of Big Data

Data, Big Data & Data Analytics

Characteristics of Big Data

Volume	Variety	Velocity	Veracity	Validity
Scale / Amount	Different Types and Sources	How fast it changes	Reliability & Truthfulness	Accuracy & Correctness

Variability	Vulnerability	Volatility	Visualization	Value
Inconsistencies & Multiple Dimensions	Security Concerns	Irrelevancy & Outdatedness	Ease of viewing	Potential of Data for the Business

Data, Big Data & Data Analytics

Data Analytics

- ***1. Descriptive***
  What happened?
- ***2. Diagnostic***
  Why did it happen?
- ***3. Predictive***
  What is likely to happen?
- ***4. Prescriptive***
  What should I do about it?
- ***5. Cognitive***
  What did I learn?

Data, Big Data & Data Analytics

Examples of Types of Data Analytics - in Cricket		
Descriptive	What happened	A Bowler's pitch-map A Batsman's Wagon Wheel Graphical view of Batsman's scores in each test innings
Diagnostic	Why did it happen?	A bowler with an overall economy rate of 4.2 and 1.1 wickets every match had figures of 10-0-80-0 in a particular one-day match, because – -- he was a left arm spinner and had no assistance from the pitch -- the batting team had 7 left-handers
Predictive	What is likely to happen	Bowler B is likely to concede 1.7 runs more per delivery if he tries to bowl a Yorker and misses the length compared to if he sticks to a normal line and length
Prescriptive	What should we do about it?	In order for Team X to win the match – -- Team X should score in excess of 325 runs batting first -- At least one of the top 3 batsmen in Team X should score 80+ runs

Components of Data Science

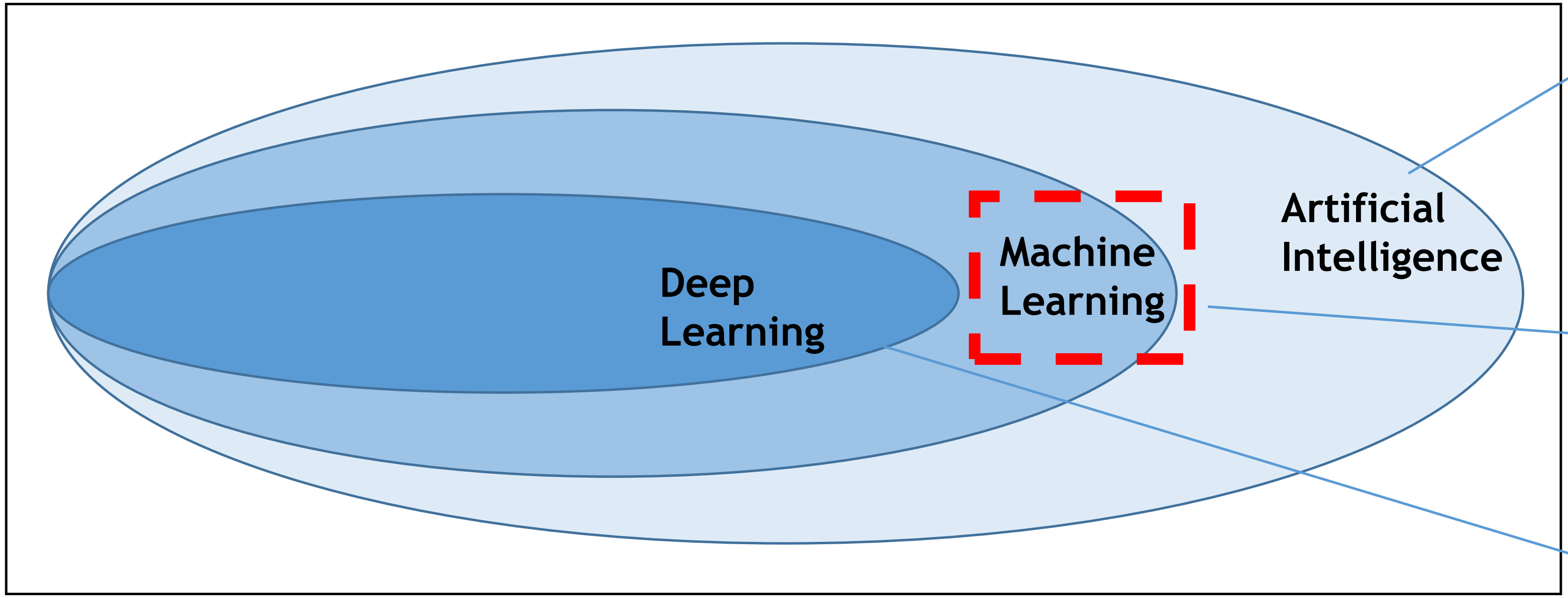
Components of Data Science

- Domain Knowledge
- Statistics & Machine Learning
- Technology & Computer Science
- Data Engineering
- Data Visualization

Machine Learning Overview

Machine Learning Overview

Machine Learning:
Ability of computers to learn without being explicitly programmed



Intelligent systems based on programmed rules as well as self-learning algorithms

Self Learning algorithms that improve with experience

Multilayered neural networks implemented on vast amounts of data

Machine Learning Overview

Types of Machine Learning

Supervised

Input & Output both known
(Output is labeled)

Regression

Output is continuous

Classification

Output is a set of discrete categories

Unsupervised

Input known, Output unknown
(Output is not labeled)

Reinforcement

Perform and be rewarded;
Use Reward signal to improve

Machine Learning Overview

Examples of Machine Learning Algorithms

- Supervised Regression algorithms –
  - Regression (Simple Linear, Multiple Linear, Polynomial)

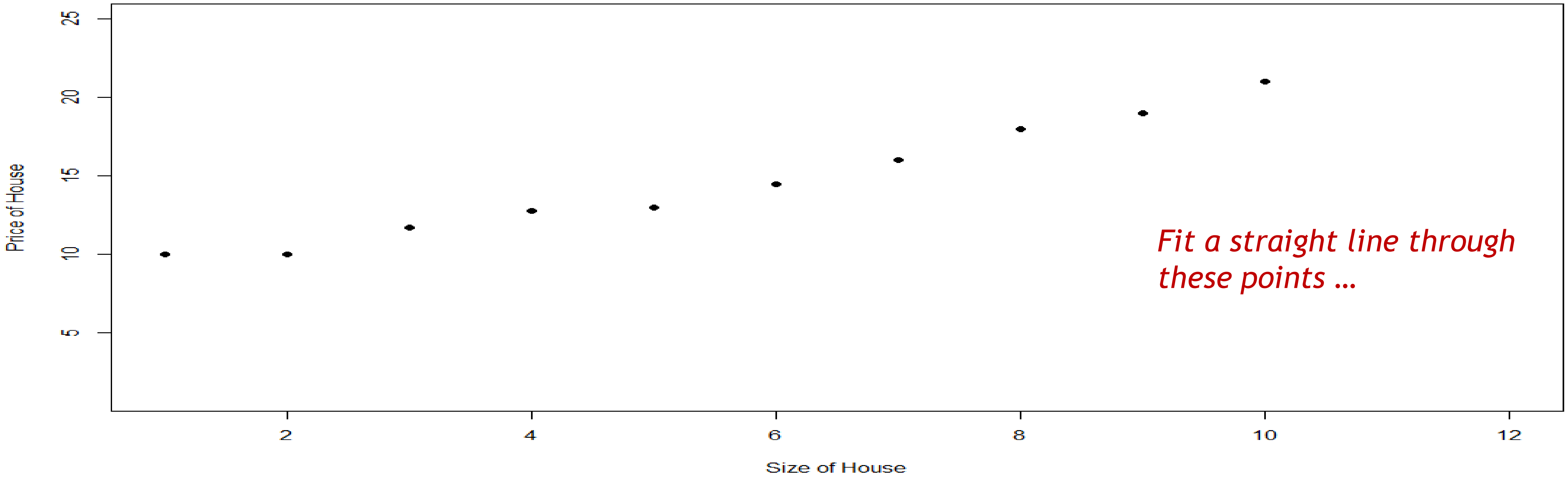
- Supervised Classification algorithms –
  - Logistic Regression
  - Support Vector Machines
  - Decision Trees
  - Random Forests
  - Naïve Bayes

- Unsupervised Algorithms
  - K-means Clustering

Linear and Logistic Regression

Linear Regression

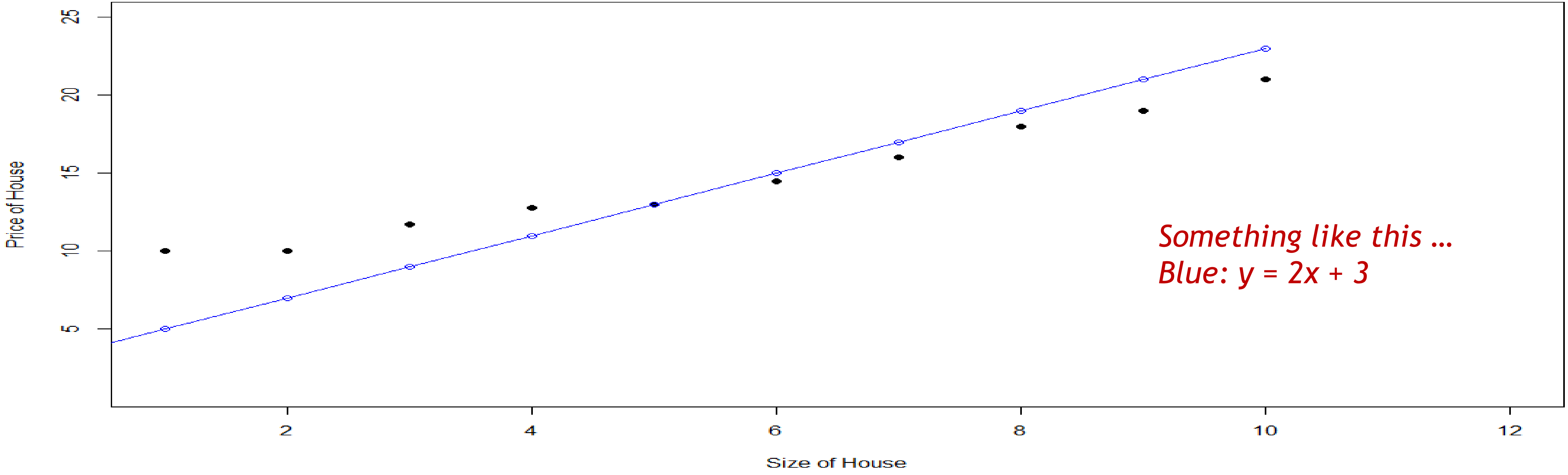
Simple Linear Regression Example – Price of house vs Size of house



Fit a straight line through these points …

Linear Regression

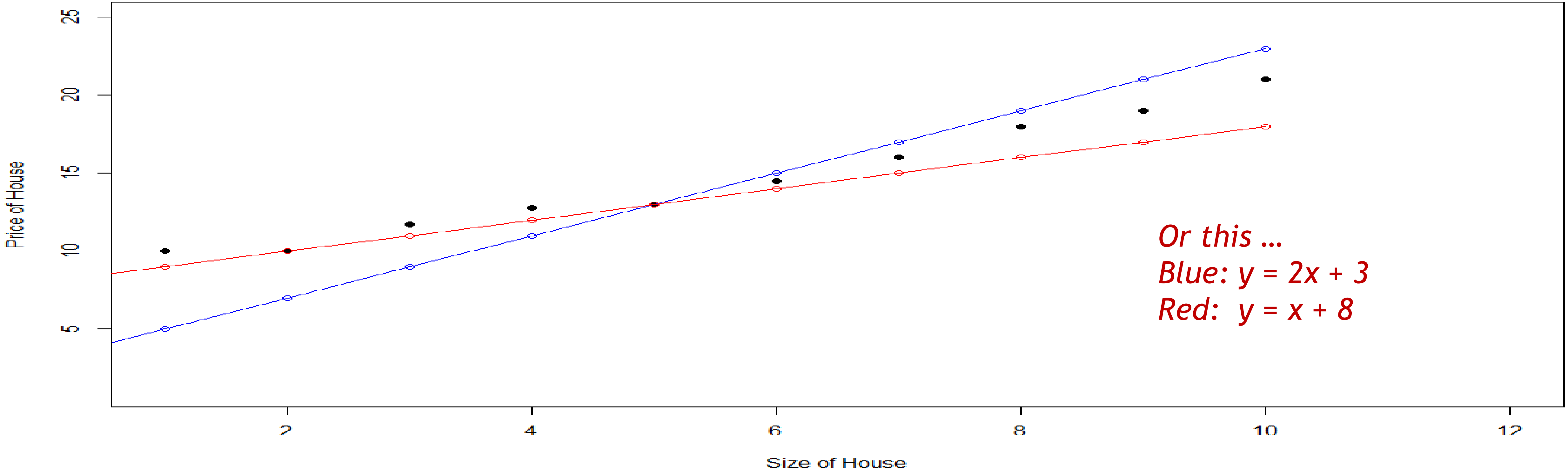
Simple Linear Regression Example – Price of house vs Size of house



Something like this …
Blue: $y = 2x + 3$

Linear Regression

Simple Linear Regression Example – Price of house vs Size of house



Linear Regression

Simple Linear Regression Example – Price of house vs Size of house

Hypothesis Function:

$h_\theta(x) = \theta_0 + \theta_1 x$

Cost Function:

$$J(\theta_0,\theta_1) = \frac{\sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2}{2m}$$

where m is the number of observations

Goal:

Minimize $J(\theta_0,\theta_1)$

Linear Regression

Simple Linear Regression Example – Price of house vs Size of house

Goal:

Minimize $J(\theta_0,\theta_1)$

Iterative approach to find θ_0, θ_1 to minimize Cost Function:

Gradient Descent

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

Where –

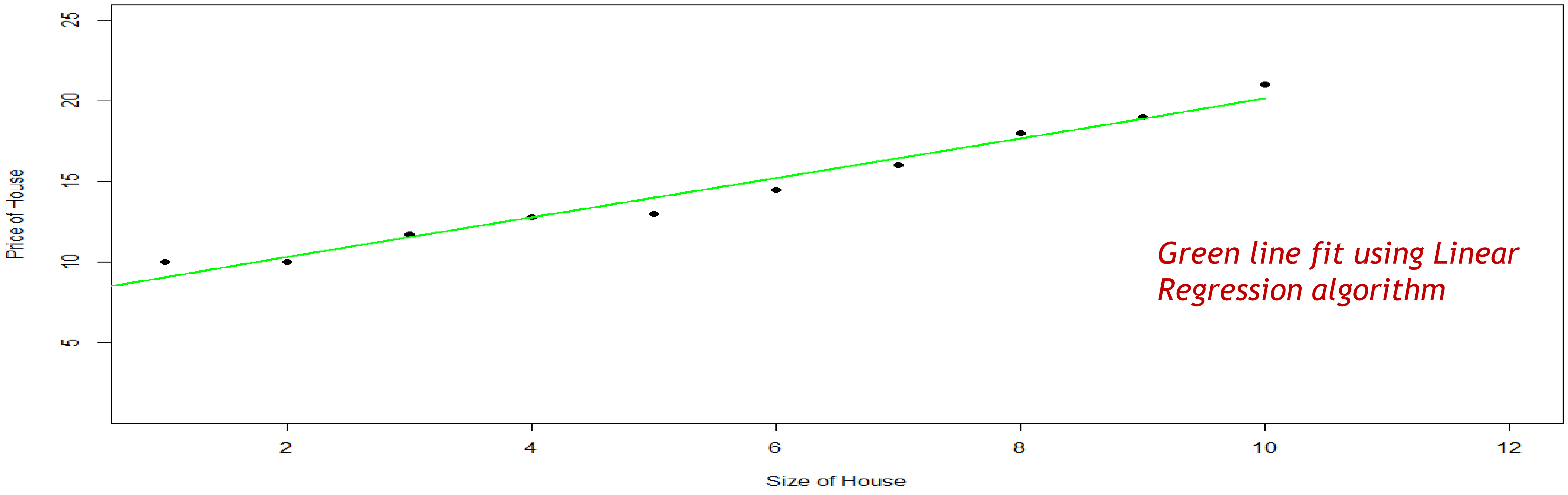
θ_j represents θ_0 and θ_1

α is called the learning rate

j is iteration number taking values 1, 2, …, n

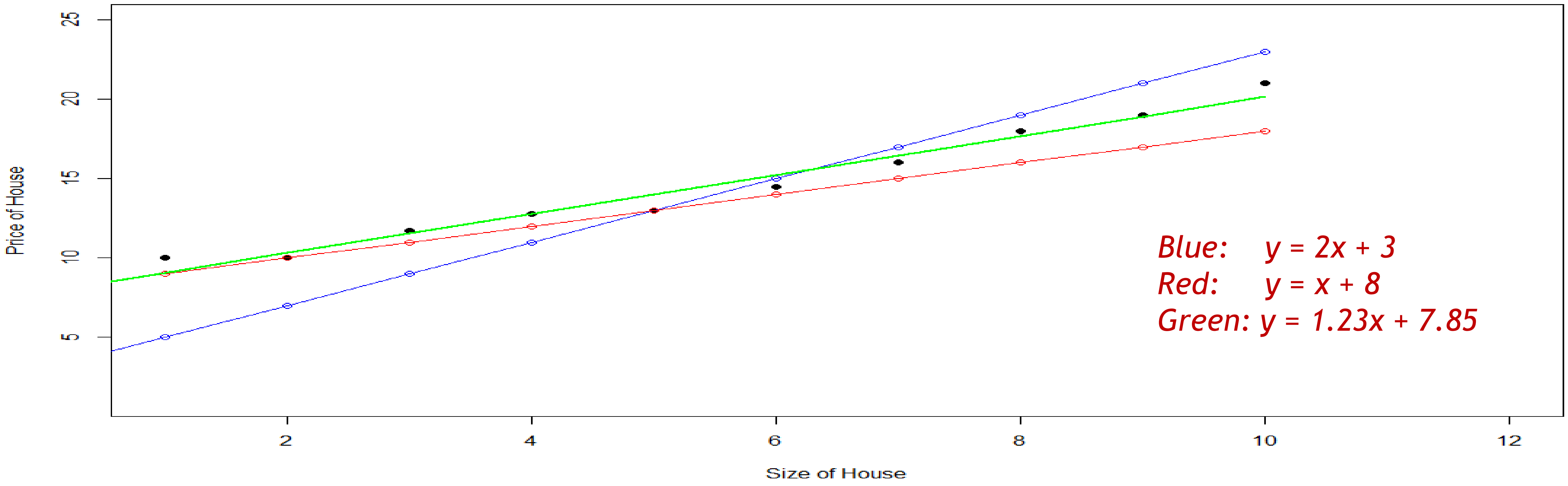
Linear Regression

Simple Linear Regression Example – Price of house vs Size of house



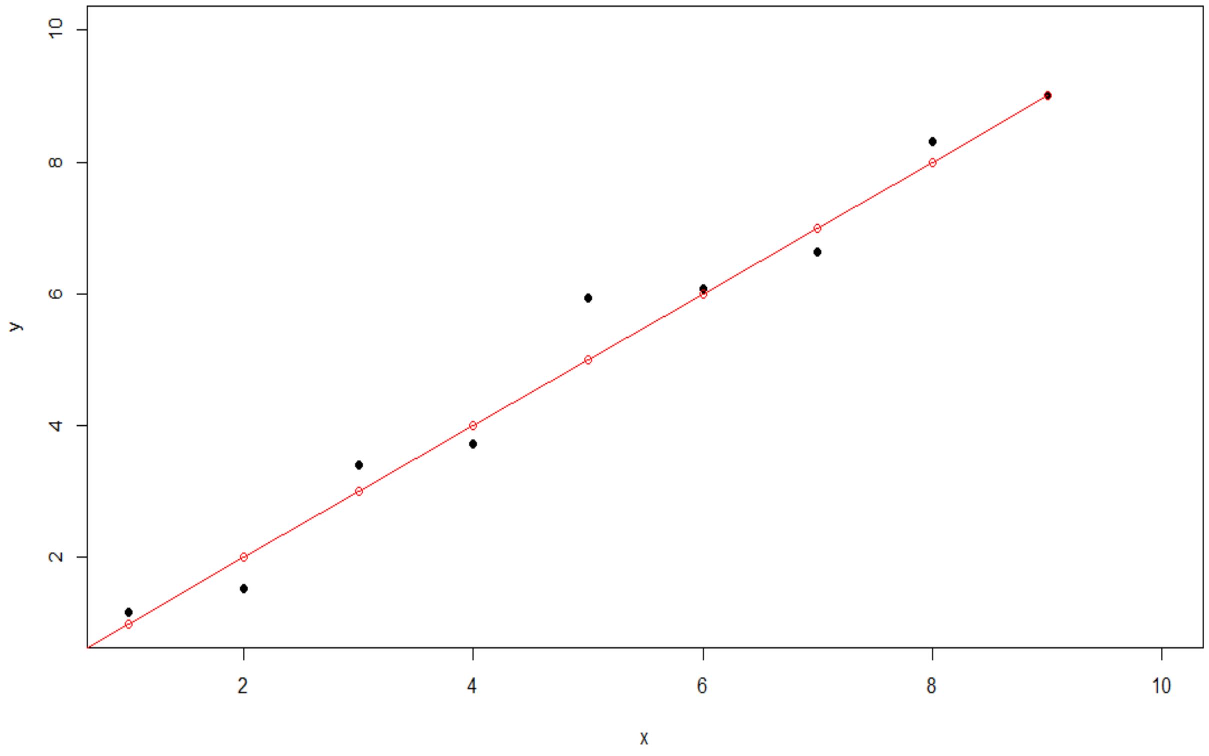
Linear Regression

Simple Linear Regression Example – Price of house vs Size of house



Linear Regression

Why Gradient Descent And what is significance of α



Example with $\theta_0 = 0$

Hypothesis function: $h_\theta(x) = \theta_1 x$

Cost Function:

$$J(\theta_1) = \frac{\sum_{i=1}^{m} (\theta_1 x_i - y_i)^2}{2m}$$

Derivative of $J(\theta_1)$ w.r.t. θ_1 to minimize $J(\theta_1)$:

$J'(\theta_1) = (\sum_{i=1}^{m} (\theta_1 x_i - y_i) * x_i) / m$

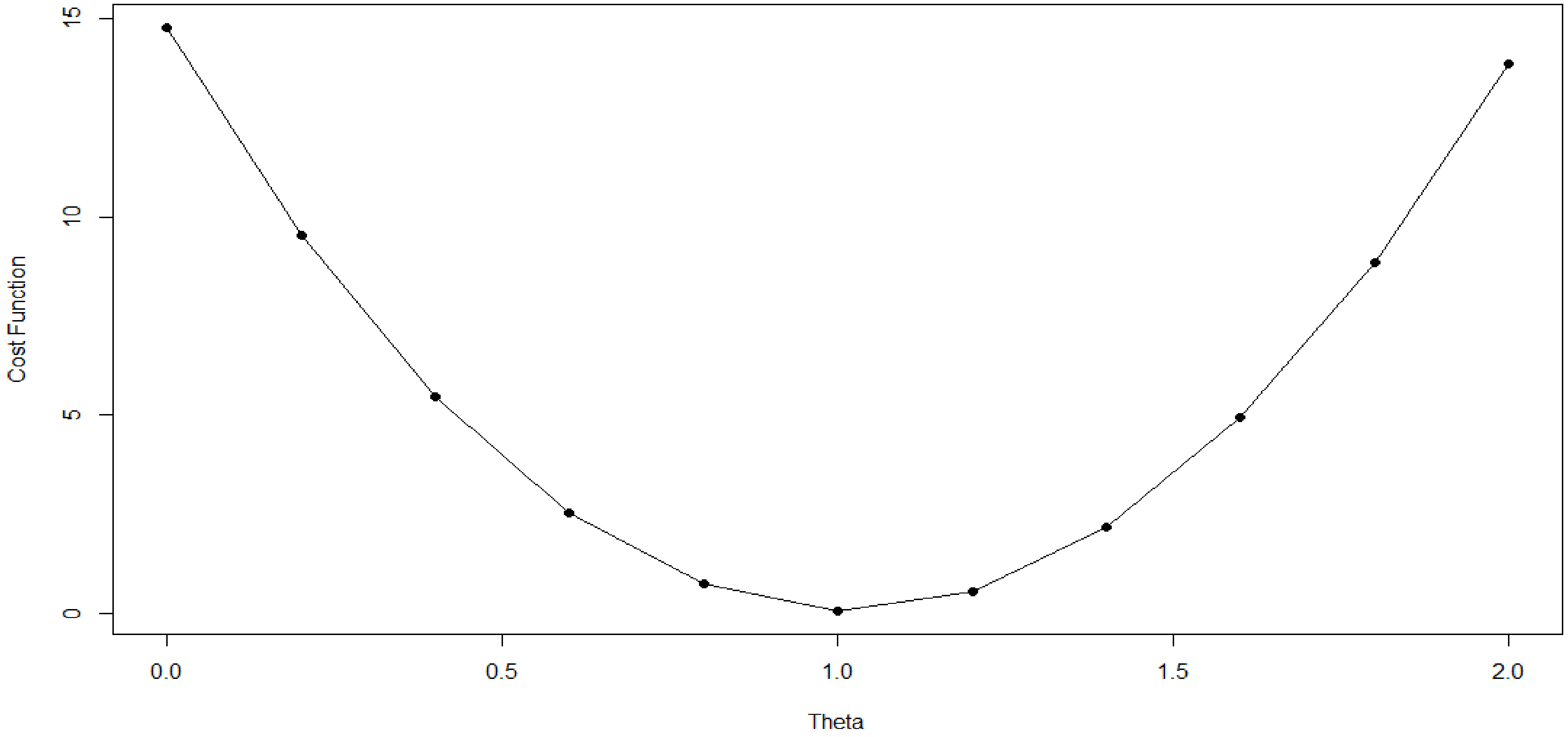
Equation for θ_1:

$\theta_j := \theta_j - (\alpha / m) * \sum_{i=1}^{m} (\theta_1 x_i - y_i) * x_i$

Linear Regression

Significance of α

Plot of $J(\theta_1)$ vs θ_1



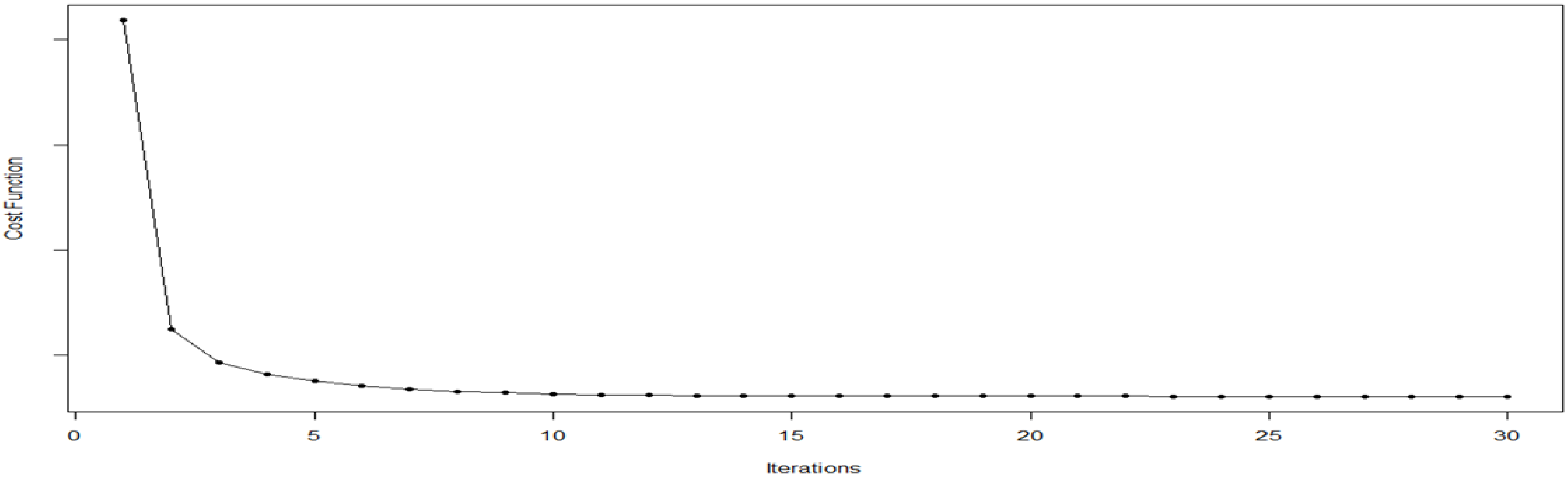
The parameter α will help you control the ***step*** through which θ_1 changes.

In this example, controlling α will prevent θ_1 oscillating between values <1 and >1 without ever reaching the desired value of 1.

Linear Regression

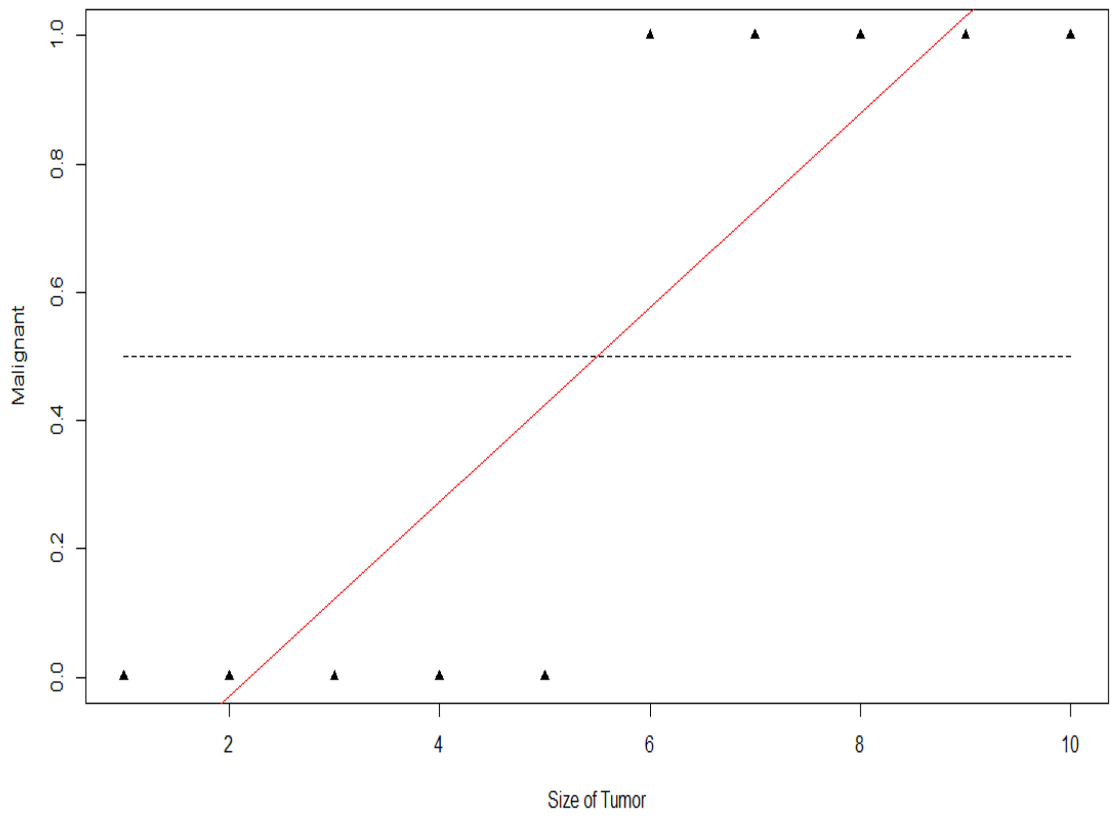
Gradient Descent

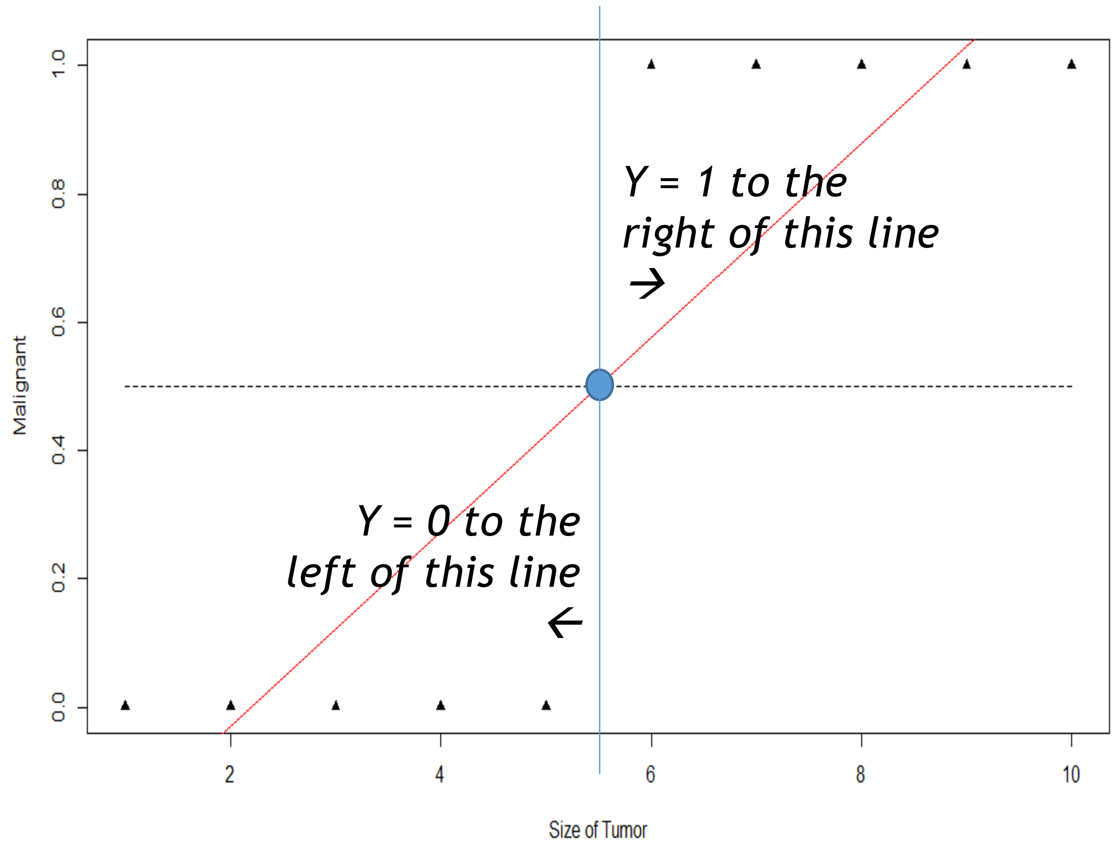
Plot of $J(\theta_1)$ vs n (number of iteration)



Logistic Regression

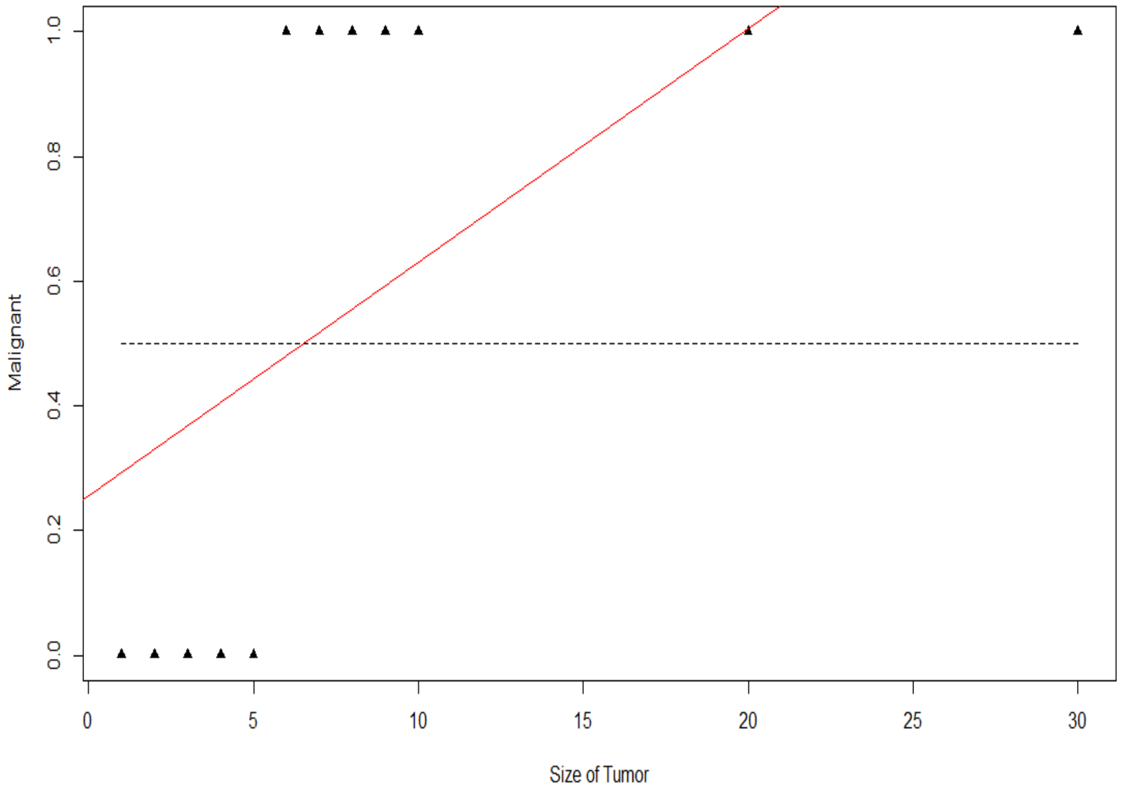
Solving Binary Classification using Linear Regression - Case 1

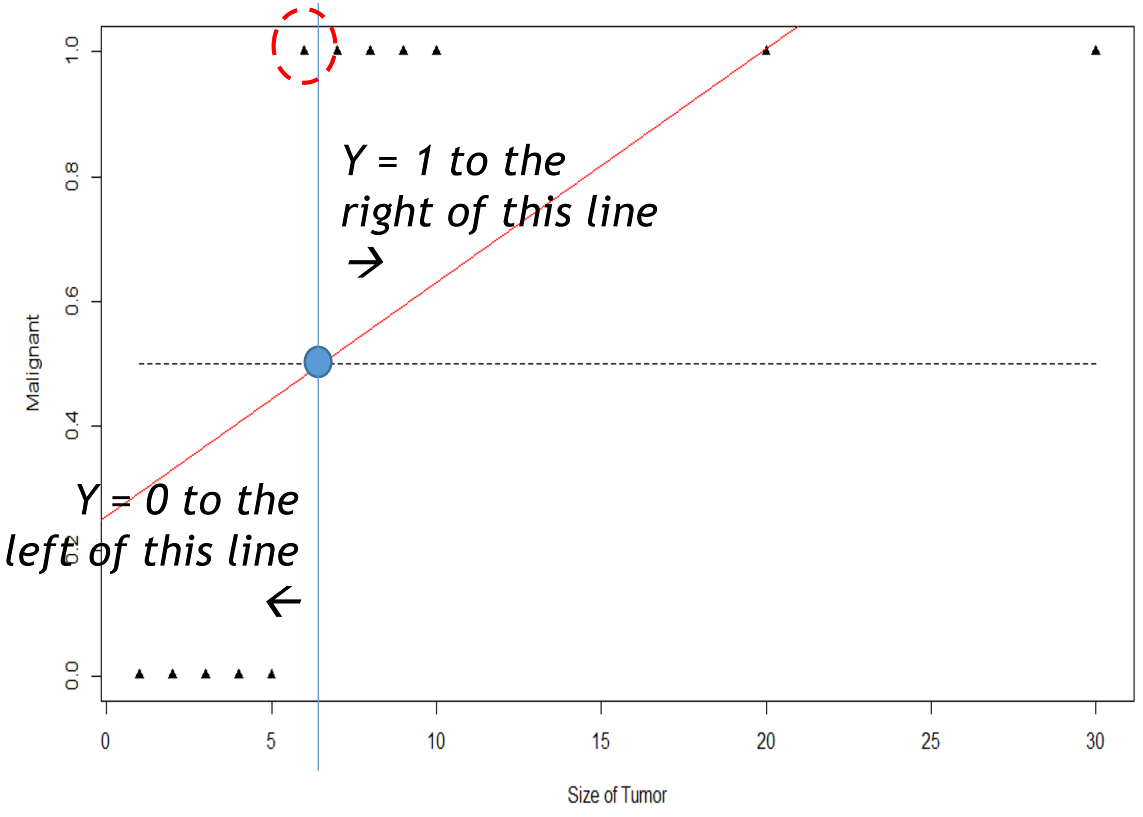




Logistic Regression

Solving Binary Classification using Linear Regression – Case 2





Logistic Regression

Although actual output Y will be either 0 or 1, the Hypothesis Function $h_\theta(x)$ will have values >1 or <0 if we solve the problem using Linear Regression.

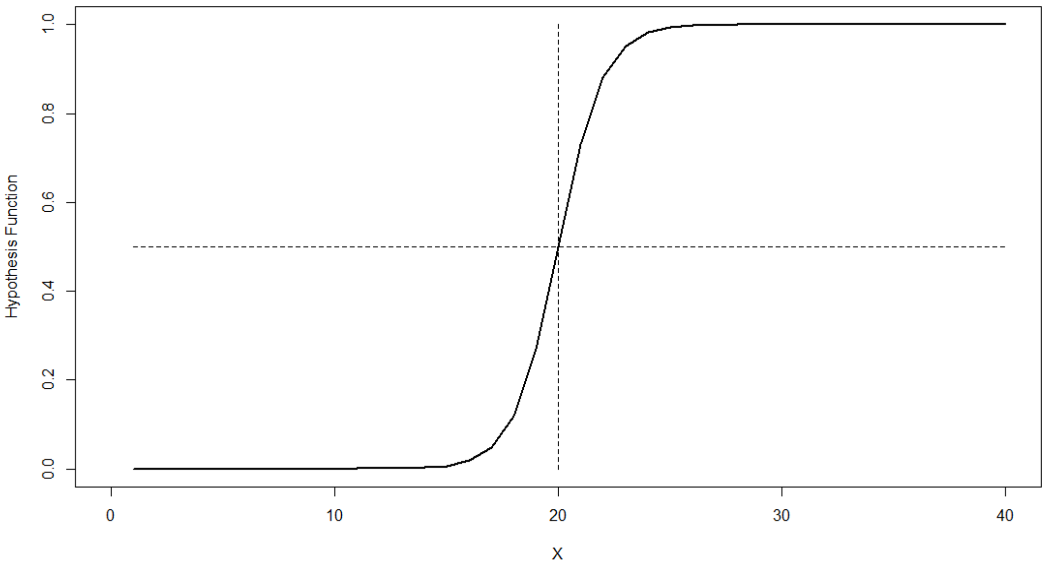
For Classification problems, the Hypothesis Function for Regression (Logistic Regression) is defined such that $0 \le h_\theta(x) \le 1$.

Linear Regression Hypothesis Function in the Matrix form:

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

Logistic Regression Hypothesis Function in the Matrix form (Sigmoid Function):

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



Logistic Regression

Logistic Regression Hypothesis Function in the Matrix form (Sigmoid Function):

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

Logistic Regression Cost Function in the Matrix form (Sigmoid Function):

$$J(\theta) = \frac{-Y^T \log(h_\theta(X)) - (1-Y)^T \log(1 - h_\theta(X))}{m}$$

The final equation for θ_j for both Linear and Logistic Regression –

$$\theta_j := \theta_j - \alpha * \frac{\partial}{\partial \theta_j} J(\theta) = \theta_j - (\alpha/m) * \sum_{i=1}^{m} [(h_\theta(x_i) - y_i) x_{ij}]$$

Intuition about Another Classifier

Spam E-mail Detection

Training Data	
Survey, Win, Tickets	Spam
Buy, Movie, Tickets	Not-Spam
Contest, Win, Prize	Not-Spam
Win, Prize, Survey	Spam

Test Data	
Survey, Buy, Prize	??

Sentiment Analysis of Movie Reviews

Training Data	
Disappointing, Average	Negative
Okay, Bad	Negative
Average, Good, Bad	Neutral
Disappointing, Great, Superb	Positive

Test Data	
Average, Superb	??

What kind of classifier would work the best for such type of problems?

Intuition about Another Classifier

Spam E-mail Detection

Training Data	
Survey, Win, Tickets	Spam
Buy, Movie, Tickets	Not-Spam
Contest, Win, Prize	Not-Spam
Win, Prize, Survey	Spam

Test Data	
Survey, Buy, Prize	??

What can I derive from this data?

- ❑ P(Spam) and P(Not-Spam)

- ❑ P(Survey | Spam)
  P(Survey | Not-Spam)
  P(Survey) = P(Survey | Spam) + P(Survey | Not-Spam)
  And so on for all other words…

What do I need to find out from this data?

- ❑ P(Spam | Survey, Buy, Prize)

What kind of classifier would work the best for such type of problems?

Intuition about Another Classifier

Sentiment Analysis of Movie Reviews

Training Data	
Disappointing, Average	Negative
Okay, Bad	Negative
Average, Good, Bad	Neutral
Disappointing, Great, Superb	Positive

Test Data	
Average, Superb	??

What can I derive from this data?

- ❑ P(Negative), P(Neutral), P(Positive)
- ❑ P(Bad | Negative),
  P(Bad | Neutral)
  P(Bad | Positive)
  P(Bad) = P(Bad | Negative) + P(Bad | Neutral) + P(Bad | Positive)
  And so on for other words…

What do I need to find out from this data?

- ❑ P(Positive | Average, Superb)

What kind of classifier would work the best for such type of problems?

Intuition about Another Classifier

Spam E-mail Detection

Training Data	
Survey, Win, Tickets	Spam
Buy, Movie, Tickets	Not-Spam
Contest, Win, Prize	Not-Spam
Win, Prize, Survey	Spam

Test Data	
Survey, Buy, Prize	??

If we assume that all the individual terms such as Survey, Win, Tickets, Buy and so on occur independently of each other (it's quite a Naïve assumption!), then –

We can write P(Survey, Buy, Prize | Spam) = P(Survey | Spam) * P(Buy | Spam) * P(Prize| Spam)
Since the RHS values can be derived, we can say we derive LHS.

What kind of classifier would work the best for such type of problems?

Intuition about Another Classifier

Spam E-mail Detection

Training Data	
Survey, Win, Tickets	Spam
Buy, Movie, Tickets	Not-Spam
Contest, Win, Prize	Not-Spam
Win, Prize, Survey	Spam

Test Data	
Survey, Buy, Prize	??

Let -
{Survey, Buy, Prize} be X and
{Spam, Not-Spam} be Y

This means -

- ✓ I know P(X | y1),
- ✓ I know P(y1) and
- ✓ I can get P(X) as P(X | y1) + P(X | y2)

➢ **I need to find out P(y1 | X)**

What kind of classifier would work the best for such type of problems?

Intuition about Another Classifier

Spam E-mail Detection

Training Data	
Survey, Win, Tickets	Spam
Buy, Movie, Tickets	Not-Spam
Contest, Win, Prize	Not-Spam
Win, Prize, Survey	Spam

Test Data	
Survey, Buy, Prize	??

- ✓ I know P(X | y1),
- ✓ I know P(y1) and
- ✓ I can get P(X) as P(X | y1) + P(X | y2)

- ➢ **I need to find out P(y1 | X)**

$$P\,(y1 \mid X) = \frac{P\,(X \mid y1) * P\,(y1)}{P\,(X)}$$

NOW WE KNOW !!! -- It's Bayes Theorem!

Intuition about Another Classifier

Bayes Theorem

$$P(Y \cap X) = P(X \cap Y) = P(Y|X) * P(X) = P(X|Y) * P(Y)$$

$$\boxed{P(Y|X) = P(X|Y) * P(Y) \;/\; P(X)}$$

X = {Survey, Win, Tickets, Buy, Movie, Contest, Prize} = $\{x_1, x_2, x_3, \ldots, x_7\}$
Y = {Spam, Not-Spam} = $\{y_1, y_2\}$

Since there are two output levels –
$P(X) = P(x_1, x_2, x_3, \ldots, x_7) = P(x_1, x_2, x_3, \ldots, x_7 \mid y_1) * P(y_1) + P(x_1, x_2, x_3, \ldots, x_7 \mid y_2) * P(y_2)$

$$\boxed{P(y_1|X) = \frac{P(x_1, x_2, x_3, \ldots, x_7 \mid y_1) * P(y_1)}{P(x_1, x_2, x_3, \ldots, x_7 \mid y_1) * P(y_1) + P(x_1, x_2, x_3, \ldots, x_7 \mid y_2) * P(y_2)}}$$

Intuition about Another Classifier

$$P(y_1|X) = \frac{P(x_1, x_2, x_3, \ldots, x_7 \mid y_1) * P(y_1)}{P(x_1, x_2, x_3, \ldots, x_7 \mid y_1) * P(y_1) + P(x_1, x_2, x_3, \ldots, x_7 \mid y_2) * P(y_2)}$$

The biggest challenge in this equation is that the term $P(x_1, x_2, x_3, \ldots, x_7 \mid y_i)$ is difficult to solve. To simplify the term, we use Naïve Bayes algorithm.

The Naïve Bayes Algorithm assumes that all the attributes in the set X are independent.

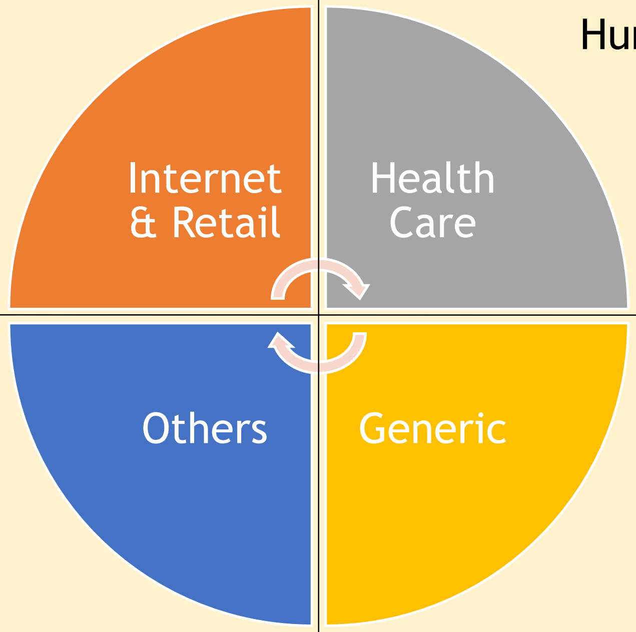
Therefore, $P(x_1, x_2, x_3, \ldots, x_7 \mid y_i) = P(x_1 \mid y_i) * P(x_2 \mid y_i) * \ldots * P(x_7 \mid y_i)$

$$P(y_1|X) = \frac{[\, P(x_1 \mid y_1) * P(x_2 \mid y_1) * \ldots * P(x_7 \mid y_1) \,] * P(y_1)}{[\, P(x_1 \mid y_1) * P(x_2 \mid y_1) * \ldots * P(x_7 \mid y_1) \,] * P(y_1) + [\, P(x_1 \mid y_2) * P(x_2 \mid y_2) * \ldots * P(x_7 \mid y_2) \,] * P(y_2)}$$

Appendix

Exciting Data Science Applications

Internet Search
Targeted Advertising
Recommender Systems
Delivery Logistics



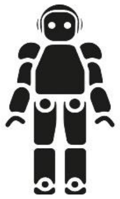
Human Genome Research
Cancer Research
Precision Medicine
Preventive Care

Airline Route Planning
Churn Prediction
Fraud detection
Spam e-mail detection

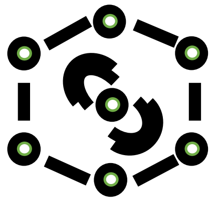
Image Recognition
Speech Recognition
Sentiment Analysis
Gaming

Self Driving Cars

Robot



Internet of Things



Thank You

Please provide your valuable feedback at - hrishikesh.bhatkhande@gmail.com

My webpages -

- ✓ LinkedIn - https://www.linkedin.com/in/hrishikesh-bhatkhande-pmp-341b8421
- ✓ Kaggle - https://www.kaggle.com/hrishikesh312
- ✓ GitHub - https://github.com/Hrishikesh312/