UNIT - II

SUPERVISED LEARNING

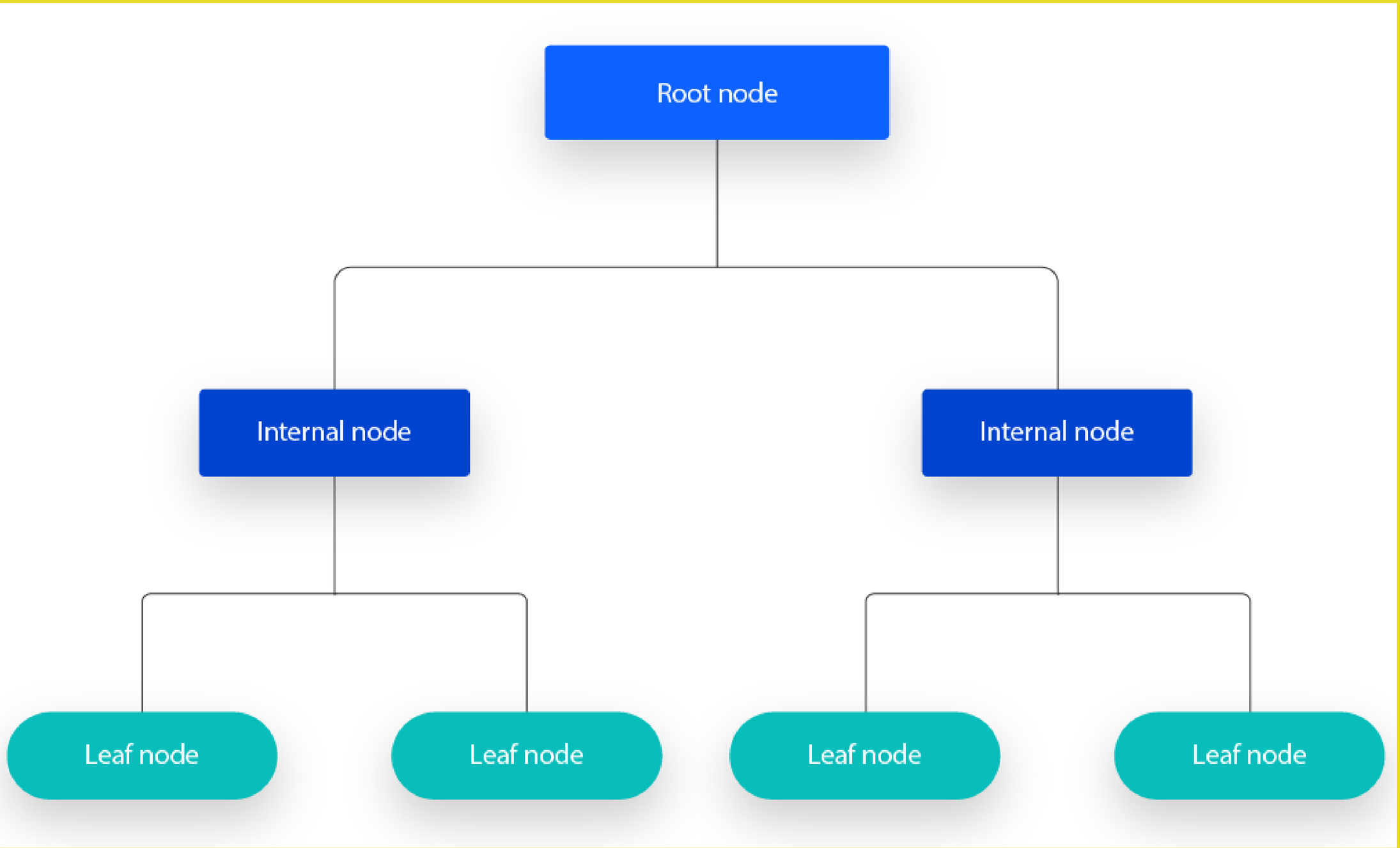
UNIT II SUPERVISED LEARNING 11

Linear Regression Models: Least squares, single & multiple variables, Bayesian linear regression, gradient descent, Linear Classification Models: Discriminant function – Perceptron algorithm, Probabilistic discriminative model - Logistic regression, Probabilistic generative model – Naive Bayes, Maximum margin classifier – Support vector machine, Decision Tree, Random Forests

DECISION TREE

DECISION TREE

- A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks.
- It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.
- A decision tree starts with a root node, which does not have any incoming branches.
- The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes.


Root node
Internal node
Internal node
Leaf node
Leaf node
Leaf node
Leaf node

DECISION TREE

- Based on the available features, both internal node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes.
- The leaf nodes represent all the possible outcomes within the dataset.
- Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree.
- This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels

How to choose the best attribute at each node

- To select the best attribute at each node, two methods, act as popular splitting criterion for decision tree models.

1. Information gain

2. Gini impurity

Entropy

- Entropy is a concept that stems from information theory, which measures the impurity of the sample values.
- Entropy is the measure of the degree of randomness or uncertainty in the dataset.

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

  - S represents the data set that entropy is calculated
  - c represents the classes in set, S
  - p(c) represents the proportion of data points that belong to class c to the number of total data points in set, S

Entropy

- Entropy values can fall between 0 and 1.
- If all samples in data set, S, belong to one class, then entropy will equal zero.
- If half of the samples are classified as one class and the other half are in another class, entropy will be at its highest at 1.
- To select the best feature to split on and find the optimal decision tree, the attribute with the smallest amount of entropy should be used.

Information Gain

- Information gain represents the difference in entropy before and after a split on a given attribute.
- The attribute with the highest information gain will produce the best split as it’s doing the best job at classifying the training data according to its target classification.

Information Gain

$$\text{Information Gain}(\text{H, A}) = H - \sum \frac{|H_V|}{|H|} H_v$$

|HV| is the number of instances in the subset S that have the value v for attribute A

|H| is the entropy of dataset sample S

A is the specific attribute or class label

Types of Decision Trees

1. ID3

2. C 4.5

3. CART

Decision Tree Terminologies

- **Root Node: The initial node at the beginning of a decision tree, where the entire population or dataset starts dividing based on various features or conditions.**
- **Decision Nodes: Nodes resulting from the splitting of root nodes are known as decision nodes. These nodes represent intermediate decisions or conditions within the tree.**

Decision Tree Terminologies

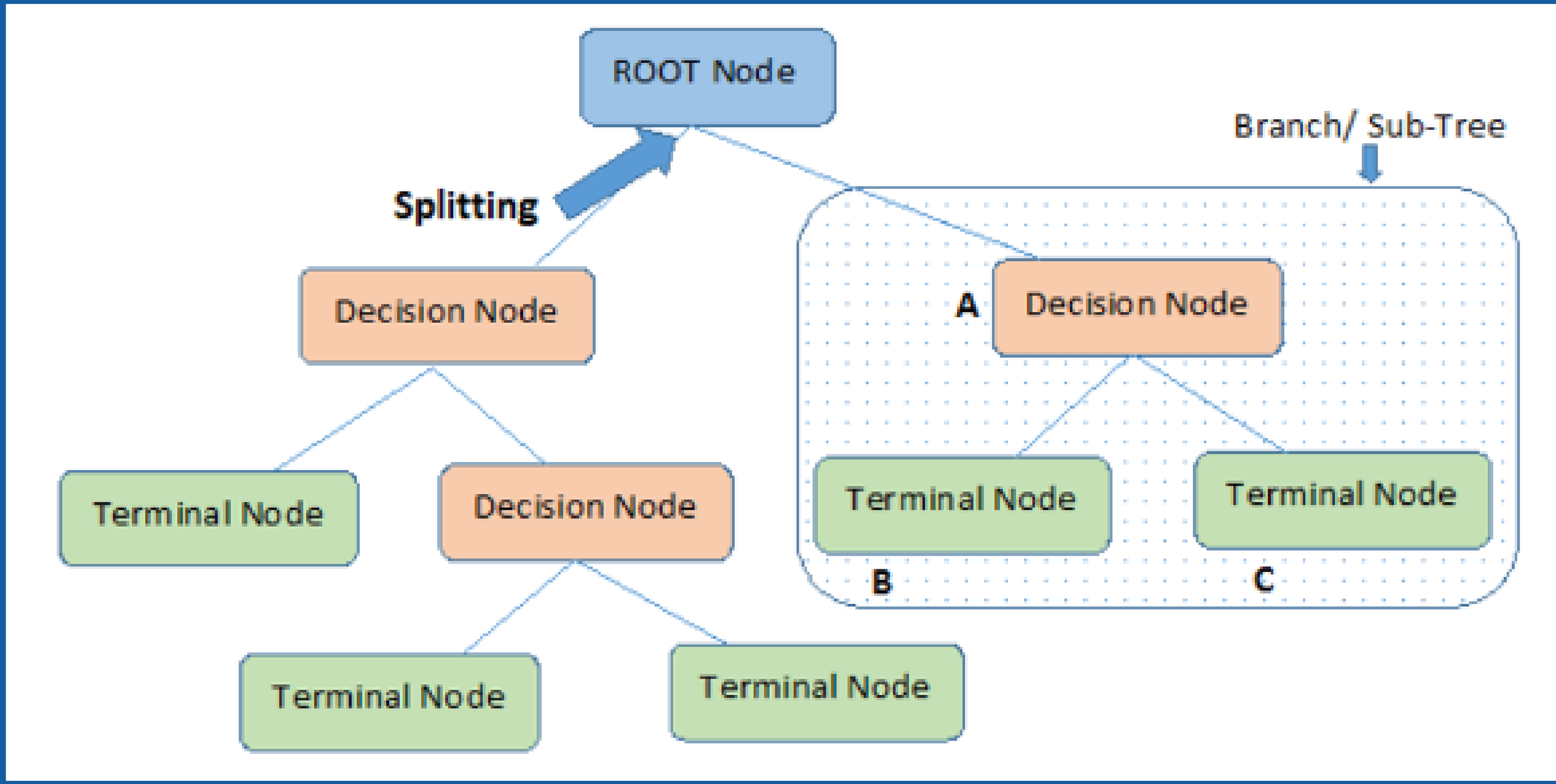
- **Leaf Nodes: Nodes where further splitting is not possible, often indicating the final classification or outcome. Leaf nodes are also referred to as terminal nodes.**
- **Sub-Tree: Similar to a subsection of a graph being called a sub-graph, a sub-section of a decision tree is referred to as a sub-tree. It represents a specific portion of the decision tree.**

Decision Tree Terminologies

- Pruning: The process of removing or cutting down specific nodes in a decision tree to prevent overfitting and simplify the model.
- Branch / Sub-Tree: A subsection of the entire decision tree is referred to as a branch or sub-tree. It represents a specific path of decisions and outcomes within the tree.

Decision Tree Terminologies

- Parent and Child Node: In a decision tree, a node that is divided into sub-nodes is known as a parent node, and the sub-nodes emerging from it are referred to as child nodes.
- The parent node represents a decision or condition, while the child nodes represent the potential outcomes or further decisions based on that condition.


ROOT Node
Branch/ Sub-Tree
Splitting
Decision Node
A
Decision Node
Terminal Node
Decision Node
Terminal Node
Terminal Node
B
C
Terminal Node
Terminal Node

Gini Index

- The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one.
- The Gini Index varies between 0 and 1, where 0 represents purity of the classification and 1 denotes random distribution of elements among various classes.

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Decision Tree using Gini Index – Solved Example

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Compute the Gini Index for the overall collection of training examples.
- There are four possible output variables Cinema, Tennis, Stay In and Shopping.
- The data has 6 instances of Cinema, 2 instances of Tennis, 1 instance of Stay In and 1 of shopping.
- $Gini(S) = 1 - [\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2] = 0.58$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Computation of **Gini Index for Money** Attribute
- It has **two possible values of Rich (7 examples)** and **Poor (3 examples).**
- For **Money = Poor,** there are **3 examples with "Cinema".**
- $Gini(S) = 1 - [\left(\frac{3}{3}\right)^2] = 0$ ✓
- For **Money = Rich,** there are **2 examples with "Tennis", 3 examples with "Cinema"** and **1 example with "Stay in", "Shopping" each**
- $Gini(S) = 1 - [\left(\frac{2}{7}\right)^2 + \left(\frac{3}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2] = 0.694$
- $\boldsymbol{Weighted\ Average(Money)}$

$$= \mathbf{0} * \left(\frac{\mathbf{3}}{\mathbf{10}}\right) + \mathbf{0.694} * \left(\frac{\mathbf{7}}{\mathbf{10}}\right) = \mathbf{0.486}$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Computation of **Gini Index for Parents** Attribute
- It has two possible values of **Yes (5 examples)** and **No (5 examples).**
- For **Parents = Yes,** there are **5 examples**, all with "Cinema".
- $Gini(S) = 1 - [\left(\frac{5}{5}\right)^2] = 0$
- For **Parents = No,** there are **2 examples with "Tennis", 1 example with "Stay in", "Shopping" and "Cinema" each**
- $Gini(S) = 1 - [\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2] = 0.72$
- $\boldsymbol{Weighted\ Average(Parents)}$

$$= \mathbf{0} * \left(\frac{\mathbf{5}}{\mathbf{10}}\right) + [\mathbf{0.72} * \left(\frac{\mathbf{5}}{\mathbf{10}}\right) = \mathbf{0.36}$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Computation of **Gini Index for Weather** Attribute
- It has three possible values of **Sunny (3 examples)**, **Rainy (3 examples)** and **Windy (4 examples).**
- For **Weather = Sunny**, there are **2 examples with "Cinema"** and **1 with "Tennis".**
- $Gini(Sunny) = 1 - [\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2] = 0.444$
- For **Weather = Rainy**, there are **2 examples with "Cinema"** and **1 example with "Stay in"**
- $Gini(Rainy) = 1 - [\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2] = 0.444$
- For **Weather = Windy**, there are **3 examples with "Cinema"** and **1 example with "Shopping"**
- $Gini(Windy) = 1 - [\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2] = 0.375$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$Weighted\ Average(Weather)$$

$$= 0.444 * \left(\frac{3}{10}\right) + 0.444 * \left(\frac{3}{10}\right) + 0.375 * \left(\frac{4}{10}\right)$$

$$= 0.416$$

For Weather - Gini Index: 0.416

For Parents - Gini Index: 0.36

For Money - Gini Index: 0.486

Parents is selected as it has smallest Gini index.

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Weather Attribute

- **Sunny (2 examples)**
- For Parent= No | Weather = Sunny, there are 2 example with "Tennis.
- $Gini(S) = 1 - [\left(\frac{2}{2}\right)^2] = 0$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Weather Attribute

- **Rainy (1 example).**
- For Parents = No | Weather = Rainy, there is 1 example with "Stay In".
- $Gini(S) = 1 - [\left(\frac{1}{1}\right)^2] = 0$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Weather Attribute

- **Windy (2 example)**
- For Parents = No | Weather = Windy, there is 1 example with "Cinema" and 1 example with "Shopping".
- $Gini(S) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$

$$Weighted\ Average(Parents = No\,|\,Weather) = 0 * \left(\frac{2}{5}\right) + 0 * \left(\frac{1}{5}\right) + 0.5 * \left(\frac{2}{5}\right) = 0.2$$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Money Attribute

- Rich (4 examples)
- For Parents = No | Money = Rich, there is 1 example with "stay in" and "Shopping" each and 2 examples of "Tennis".
- $Gini(S) = 1 - [\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{2}{4}\right)^2] = 0.625$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Money Attribute

- Poor (1 example)
- For Parents = No | Money = Poor, there is 1 example with "Cinema".
- $Gini(S) = 1 - [\left(\frac{1}{1}\right)^2] = 0$
- $Weighted\ Average\ (Parents = No|\ Money) = 0.625 * (4/5) + 0 * (1/5) = 0.5$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

For Parents = No | Weather - Gini Index: 0.2

For Parents = No | Money - Gini Index: 0.5

Weather is selected as it has smallest Gini index.

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Now, for Parent=No & Weather=Sunny, we have all instances as Tennis.

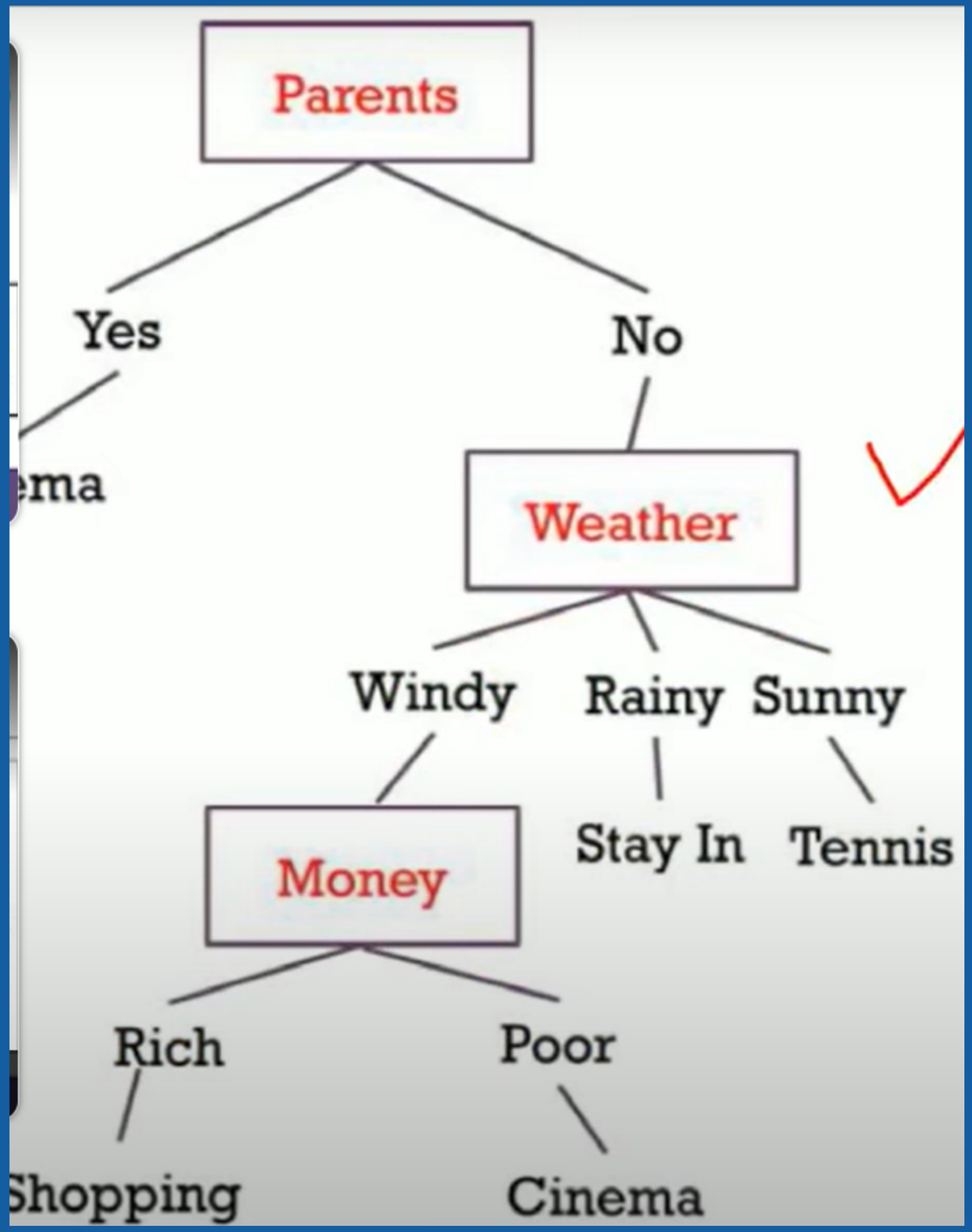
Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

Now, for Parents=No & Weather=Rainy, we have all instances as Stay In.

Weekend	Weather	Parents	Money	Decision
W5	Rainy	No	Rich	Stay In

Now, for Parent=No & Weather=Windy, we need to split.

Weekend	Weather	Parents	Money	Decision
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping


Parents
Yes
No
ema
Weather
Windy
Rainy
Sunny
Money
Stay In
Tennis
Rich
Poor
Shopping
Cinema

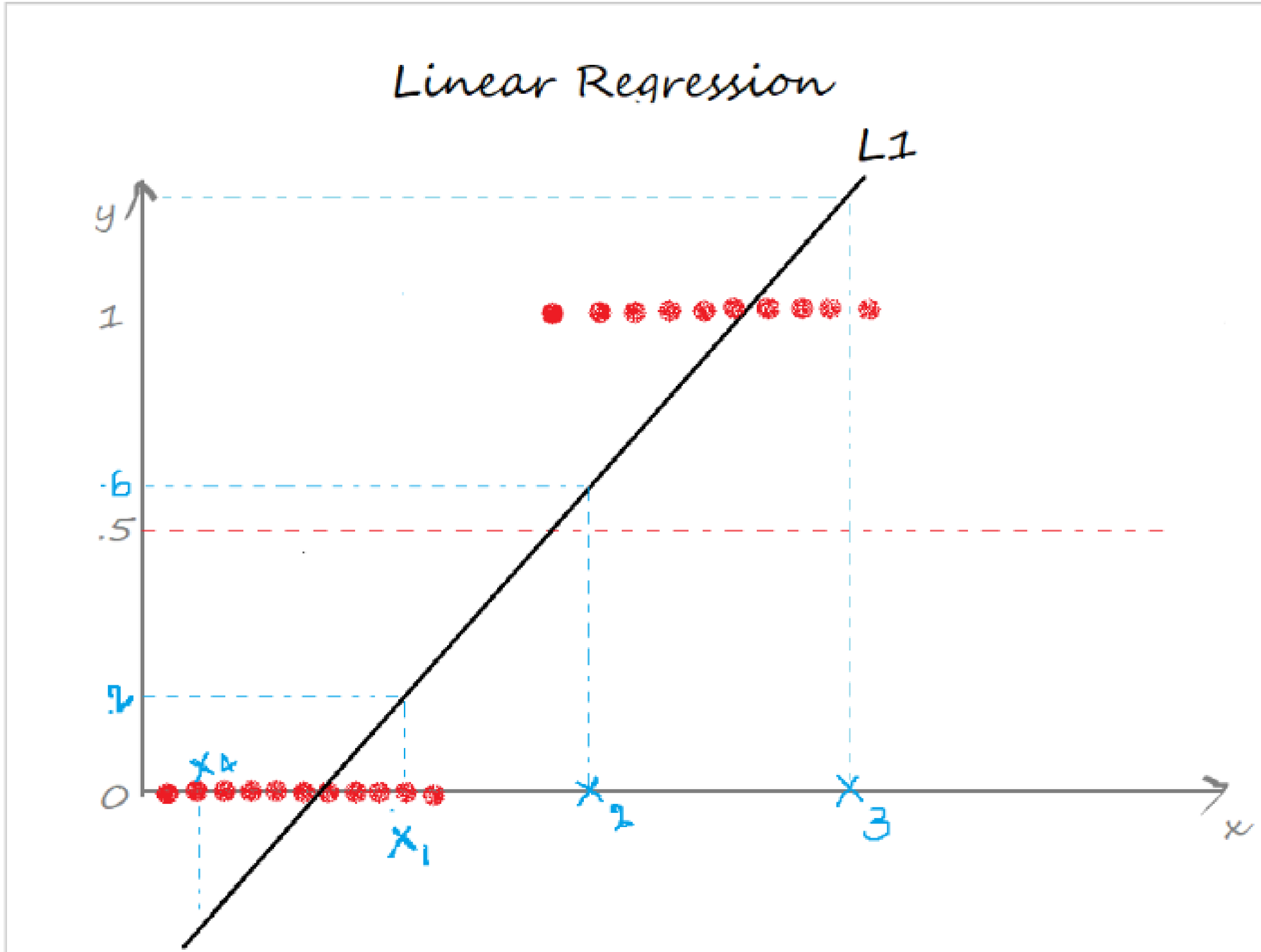
Probabilistic Discriminative Model Logistic Regression

We cannot solve a classification problem using linear regression, Why?

- The equation of the line L1 is y=mx+c, where m is the slope and c is the y-intercept.
- We define a threshold T = 0.5, above which the output belongs to class 1 and class 0 otherwise.

$$y = mx + c, \text{ Threshold } T = 0.5$$

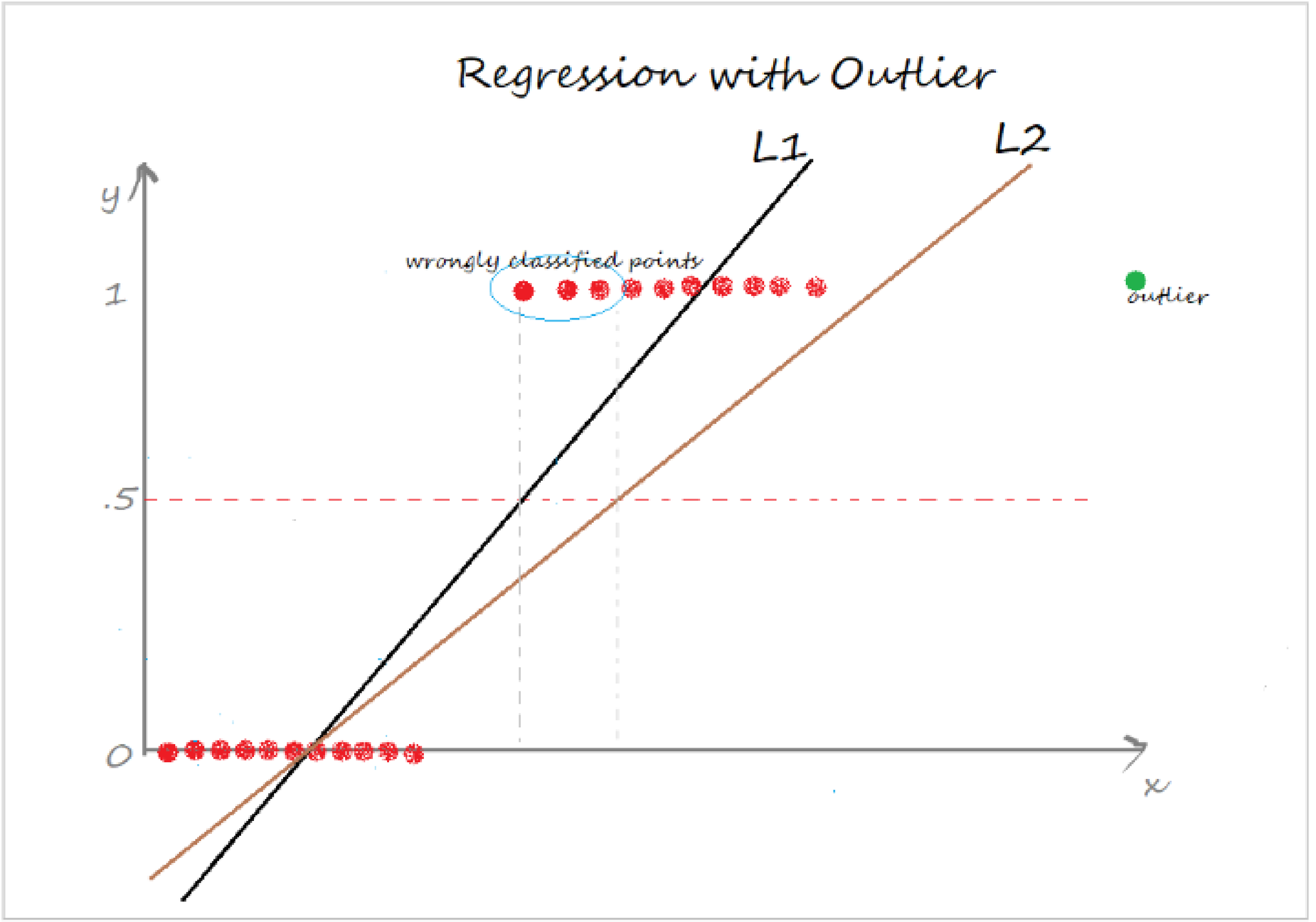
$$y = \begin{cases} 1, & mx + c \geq 0.5 \\ 0, & mx + c < 0.5 \end{cases}$$


Linear Regression
L1
y
1
.6
.5
.2
0
x
x4
x1
x2
x3

- **Case 1:** the predicted value for x1 is ≈0.2 which is less than the threshold, so x1 belongs to class 0.
- **Case 2:** the predicted value for the point x2 is ≈0.6 which is greater than the threshold, so x2 belongs to class 1.
- **Case 3:** the predicted value for the point x3 is beyond 1.
- **Case 4:** the predicted value for the point x4 is below 0.

- The predicted values for the points x3, x4 exceed the range (0,1) which doesn't make sense because the probability values always lie between 0 and 1.
- Our output can have only two values either 0 or 1.
- Hence, this is a problem with the linear regression model.

- Now, introduce an outlier and see what happens.
- The regression line gets deviated to keep the distance of all the data points to the line to be minimal.
- L2 is the new best-fit line after the addition of an outlier.


Regression with Outlier
L1
L2
y
1
.5
0
x
wrongly classified points
outlier

Limitations of Linear Regression

- The two limitations of using a linear regression model for classification problems are:
  - The predicted value may exceed the range (0,1)
  - Error rate increases if the data has outliers

What is Logistic Regression?

- Logistic Regression is a statistical and classification model.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function.
- Logistic regression uses the concept of logistic function or sigmoid function.
- The value of logistic function lies between 0 and 1.

How does Logistic Regression Work?

- Consider we have a model with one predictor "x" and one response variable "ŷ" and p is the probability of ŷ=1.
- The linear regression equation can be written as:

$$p = b_0 + b_1 x \text{ --------> eq 1}$$

- The right-hand side of the equation (b_0+b_1x) is a linear equation and can hold values that exceed the range (0,1).
- But probability will always be in the range of (0,1).
- To overcome that, we predict odds instead of probability.
- **Odds:** The ratio of the probability of an event occurring to the probability of an event not occurring.
- Odds = p/(1-p)

- Equation 1 can be rewritten as:
- $p/(1-p) = b_0 + b_1 x$ --------> eq 2
- Odds can only be a positive value, to tackle the negative numbers, we predict the **logarithm of odds**.
- Equation 2 can be rewritten as:

$$\text{Log of odds} = \ln(p/(1-p))$$

- $\ln(p/(1-p)) = b_0 + b_1 x$ --------> eq 3

To recover p from equation 3, we apply exponential on both sides.

$$\exp(\ln(p/(1-p))) = \exp(b_0+b_1x)$$

$$e^{\ln(p/(1-p))} = e^{(b0+b1x)}$$

From the inverse rule of logarithms,

$$p/(1-p) = e^{(b0+b1x)}$$

Simple algebraic manipulations

$$p = (1-p) * e^{(b0+b1x)}$$

$$p = e^{(b0+b1x)} - p * e^{(b0+b1x)}$$

Taking p as common on the right-hand side

$$p = p * ((e^{(b0+b1x)})/p - e^{(b0+b1x)})$$
$$p = e^{(b0+b1x)} / (1 + e^{(b0+b1x)})$$

Dividing numerator and denominator by $e^{(b0+b1x)}$ on the right-hand side

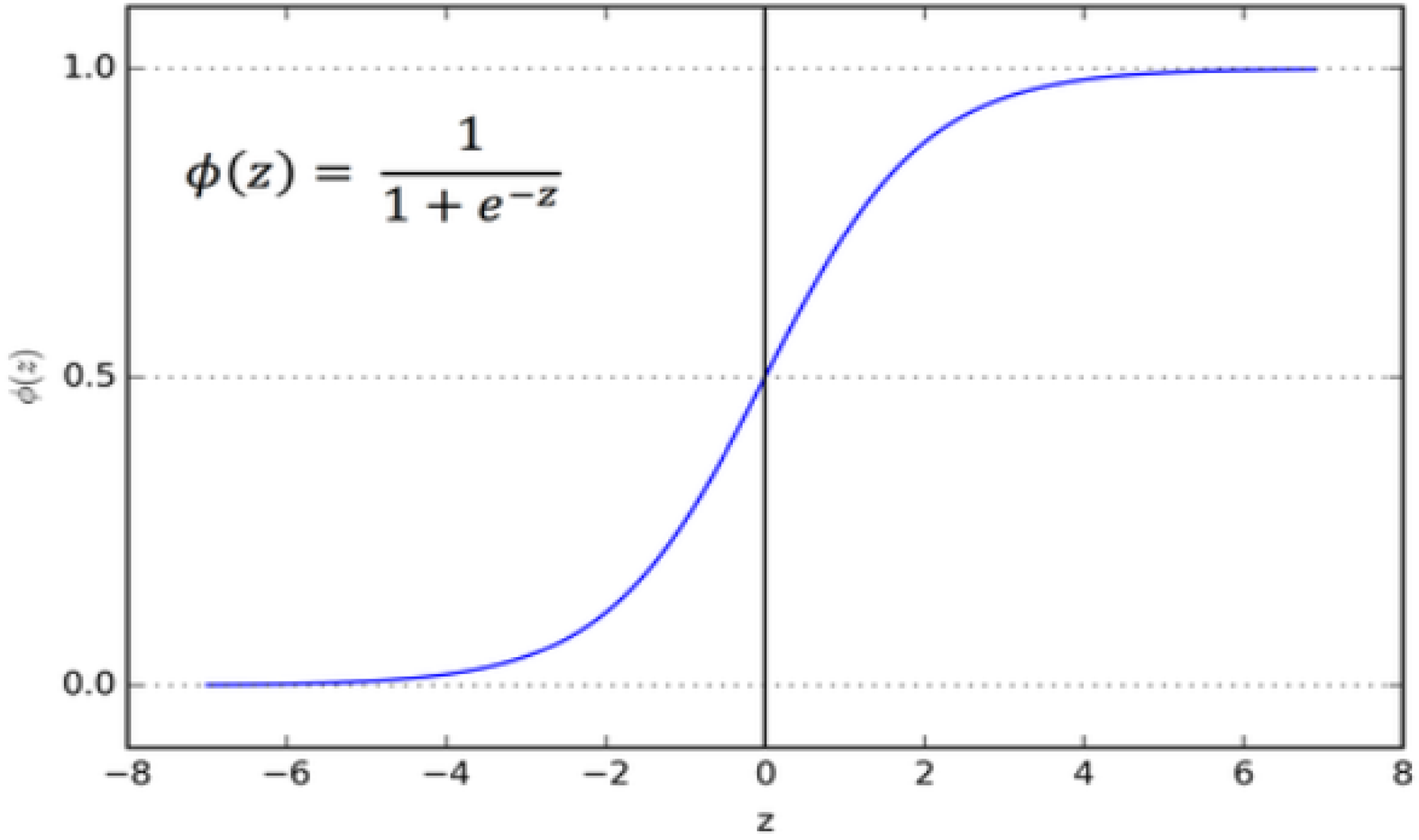
$$p = 1 / (1 + e^{-(b0+b1x)})$$

Similarly, the equation for a logistic model with 'n' predictors is as below:

$$p = 1/ (1 + e^{-(b0+b1x1+b2x2+b3x3+----+bnxn)})$$

- We arrive at the sigmoid function, which helps to squeeze the output to be in the range between 0 and 1.

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



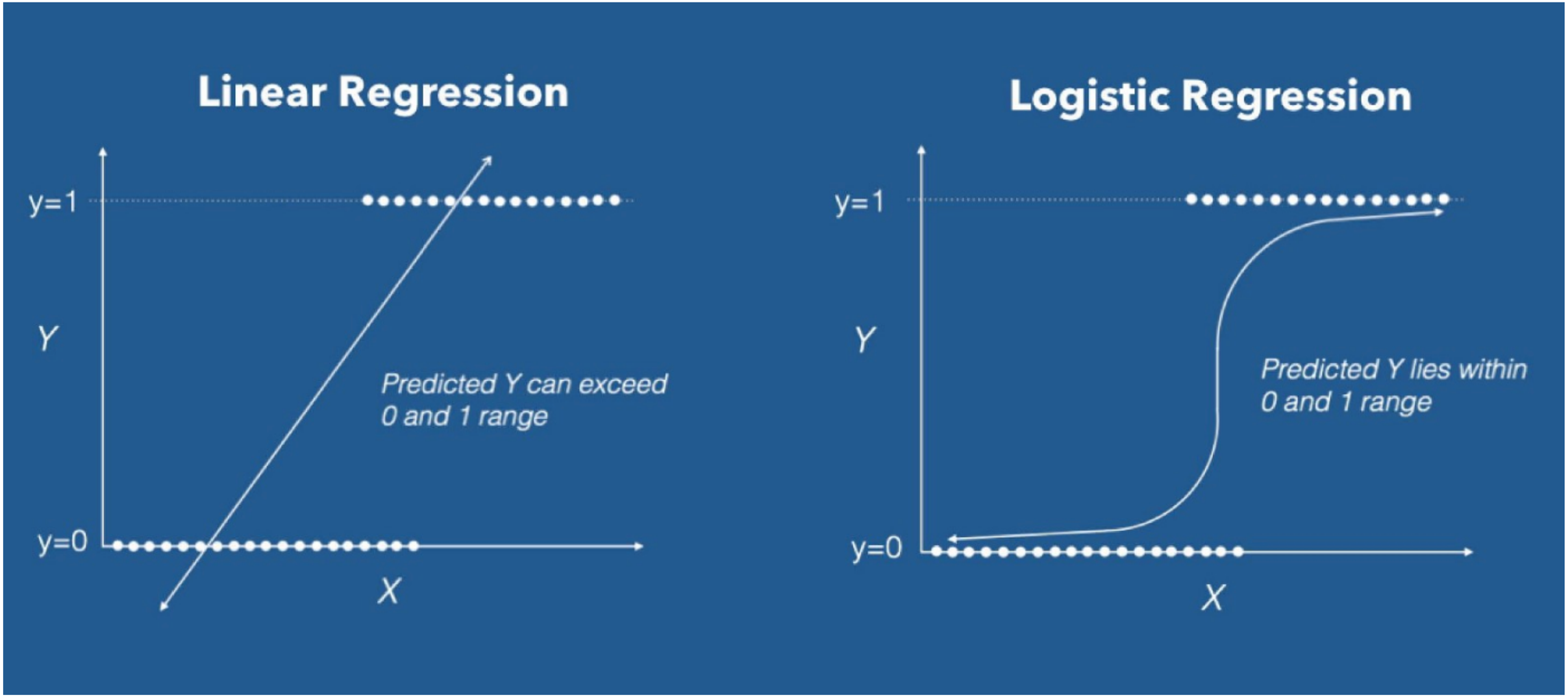
- **Linear model:**

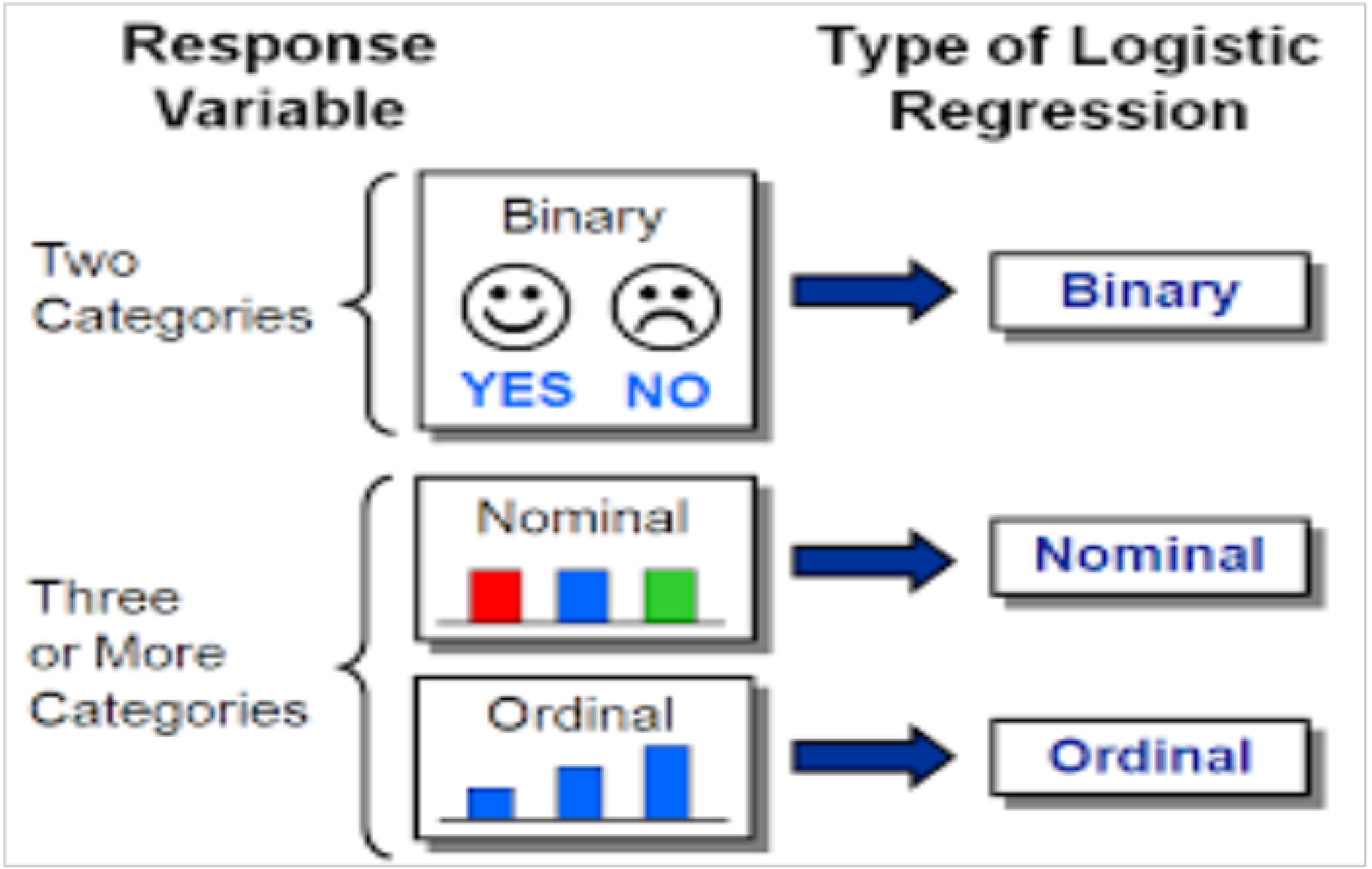
$$\hat{y} = b_0 + b_1 x$$

Sigmoid function: $\sigma(z) = 1/(1+e^{-z})$

Logistic regression model:

$$\hat{y} = \sigma(b_0 + b_1 x) = 1/(1+e^{-(b0+b1x)})$$


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed
0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within
0 and 1 range


Response Variable
Type of Logistic Regression
Two Categories
Binary
YES NO
Binary
Three or More Categories
Nominal
Nominal
Ordinal
Ordinal