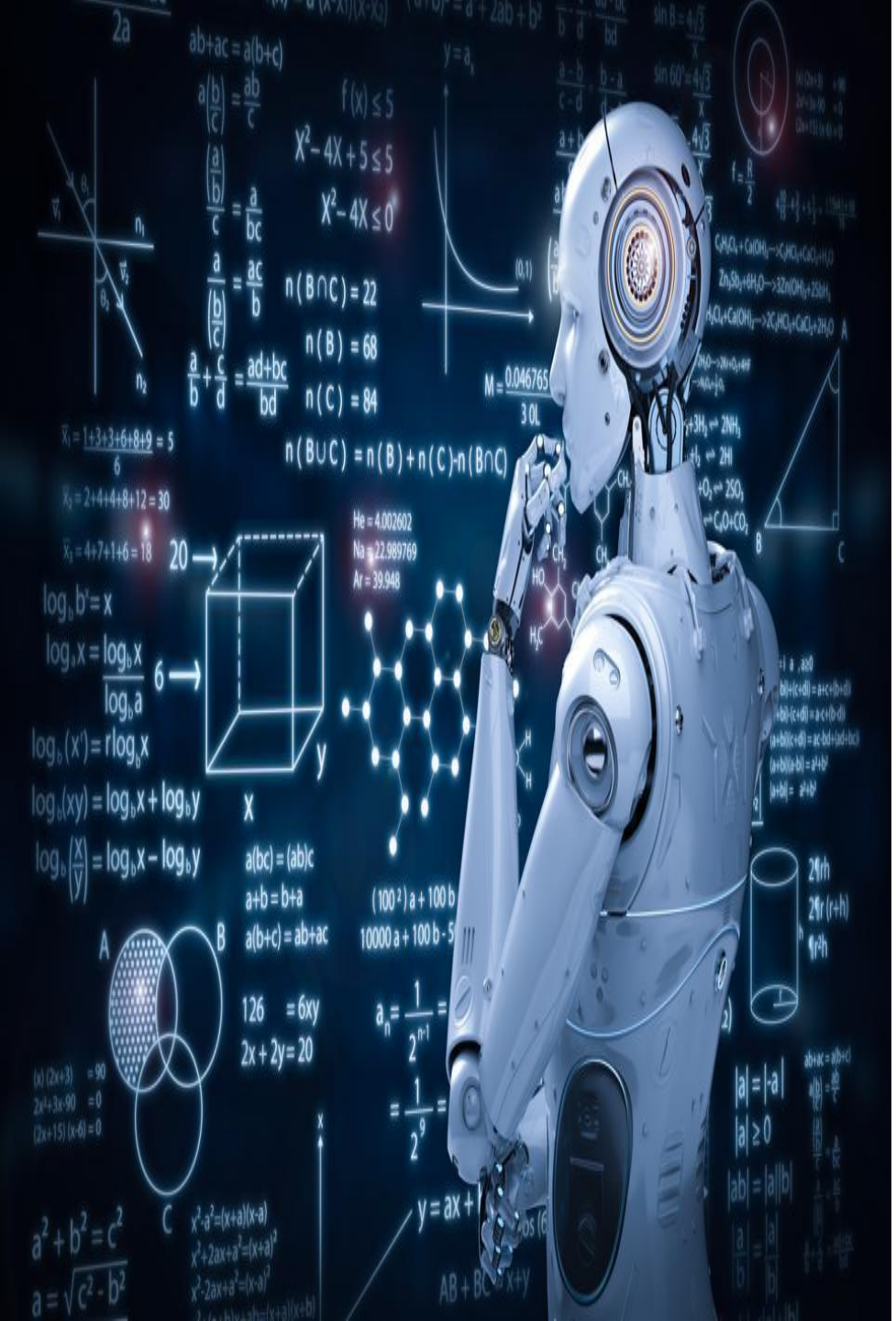


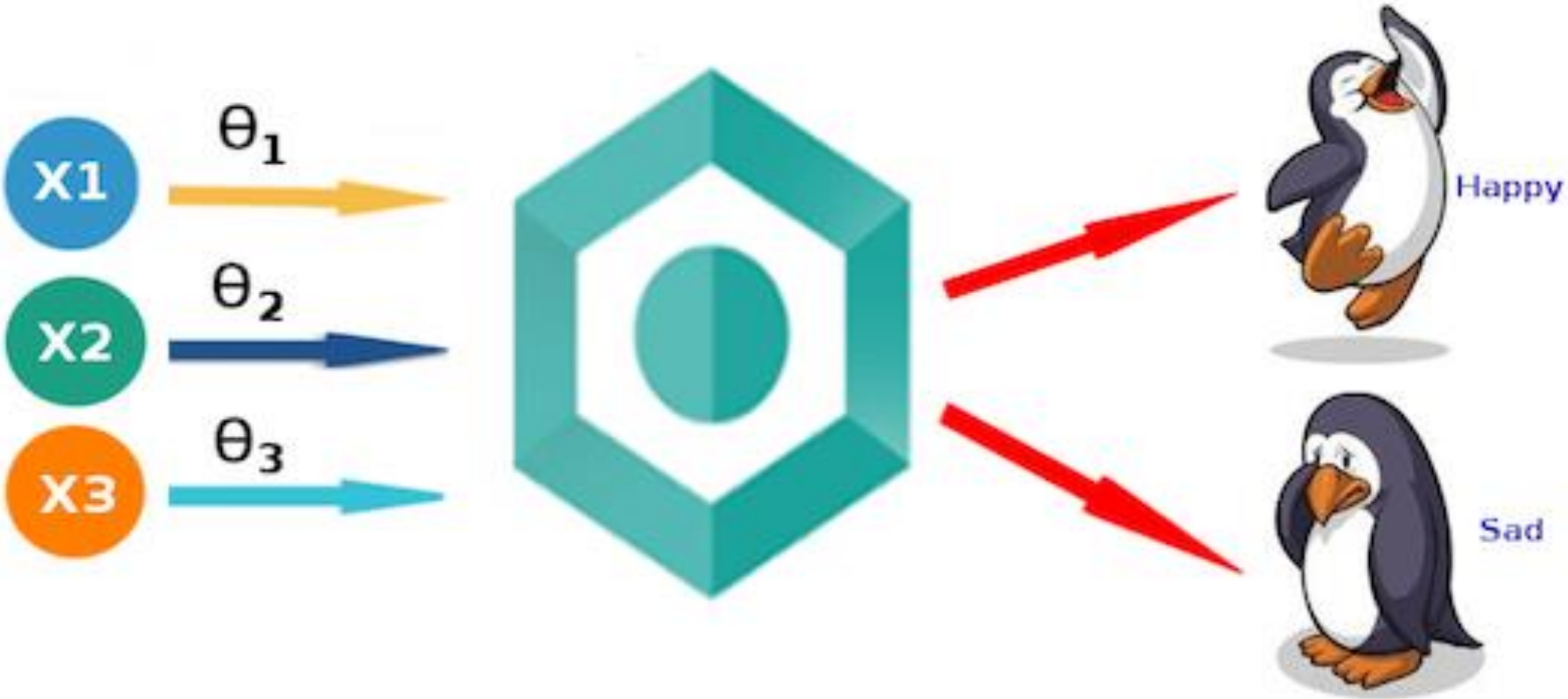
Practical Machine Learning

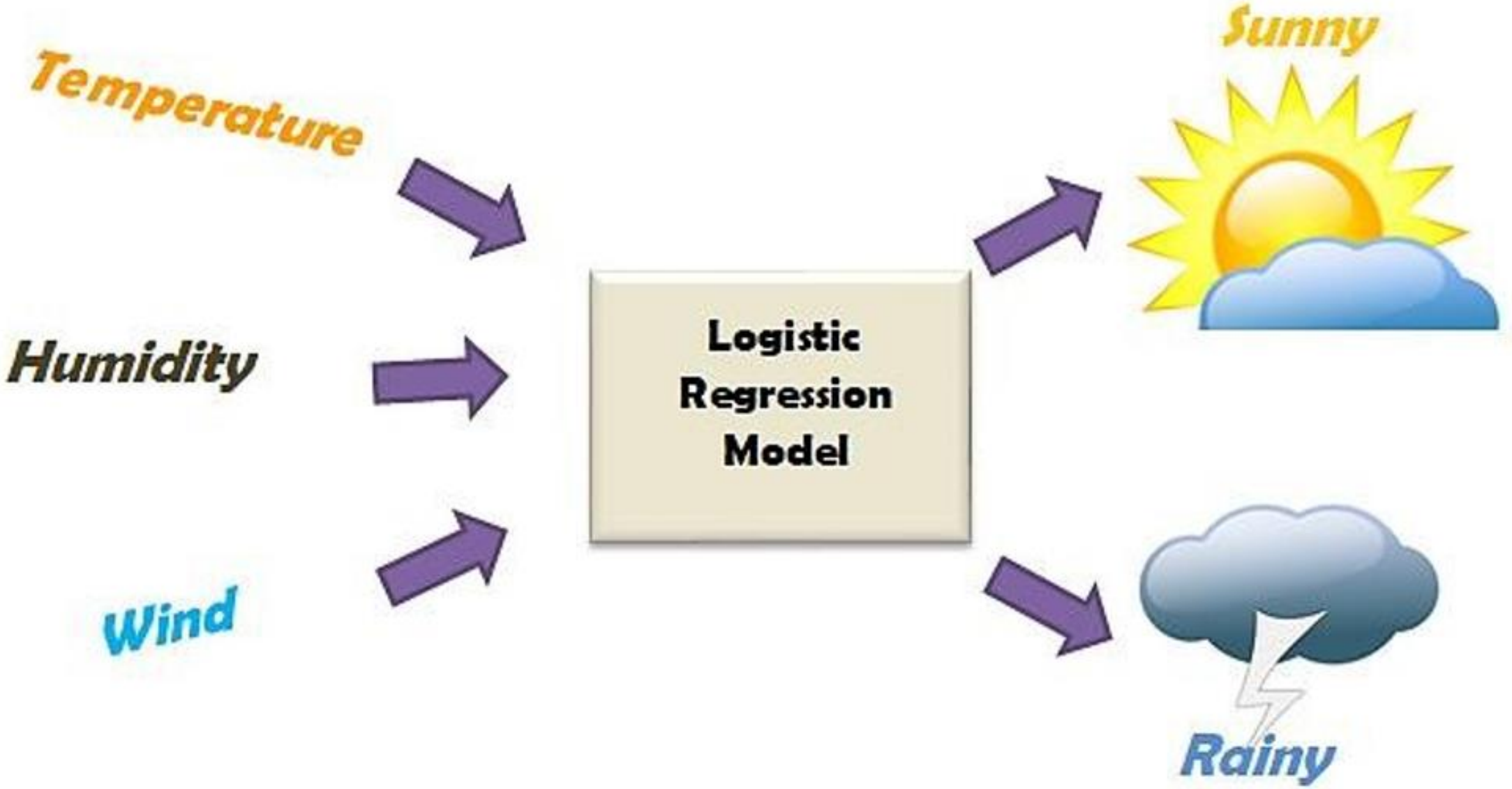
Day 8: Mar23 DBDA

Kiran Waghmare

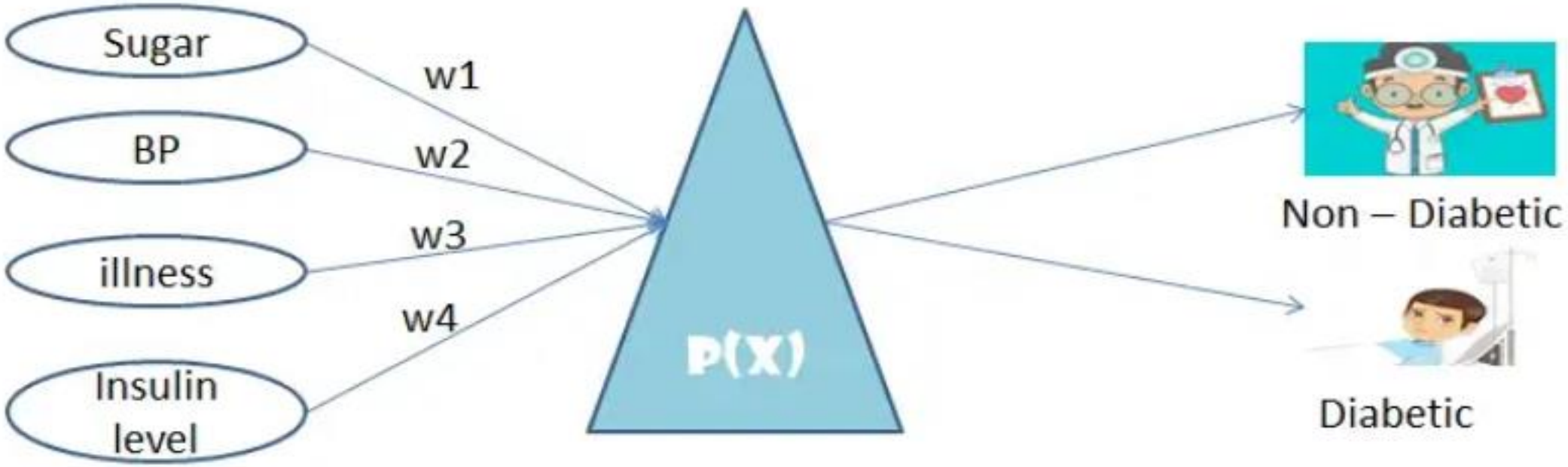
Agenda

- Logistic Regression
- Classification
- Measures for classification
- KNN


X1
X2
X3
θ1
θ2
θ3
Happy
Sad
@dataaspirant.com
Inputs: X1, X2, X3 || Weights: θ1, θ2, θ3 || Outputs: Happy or Sad

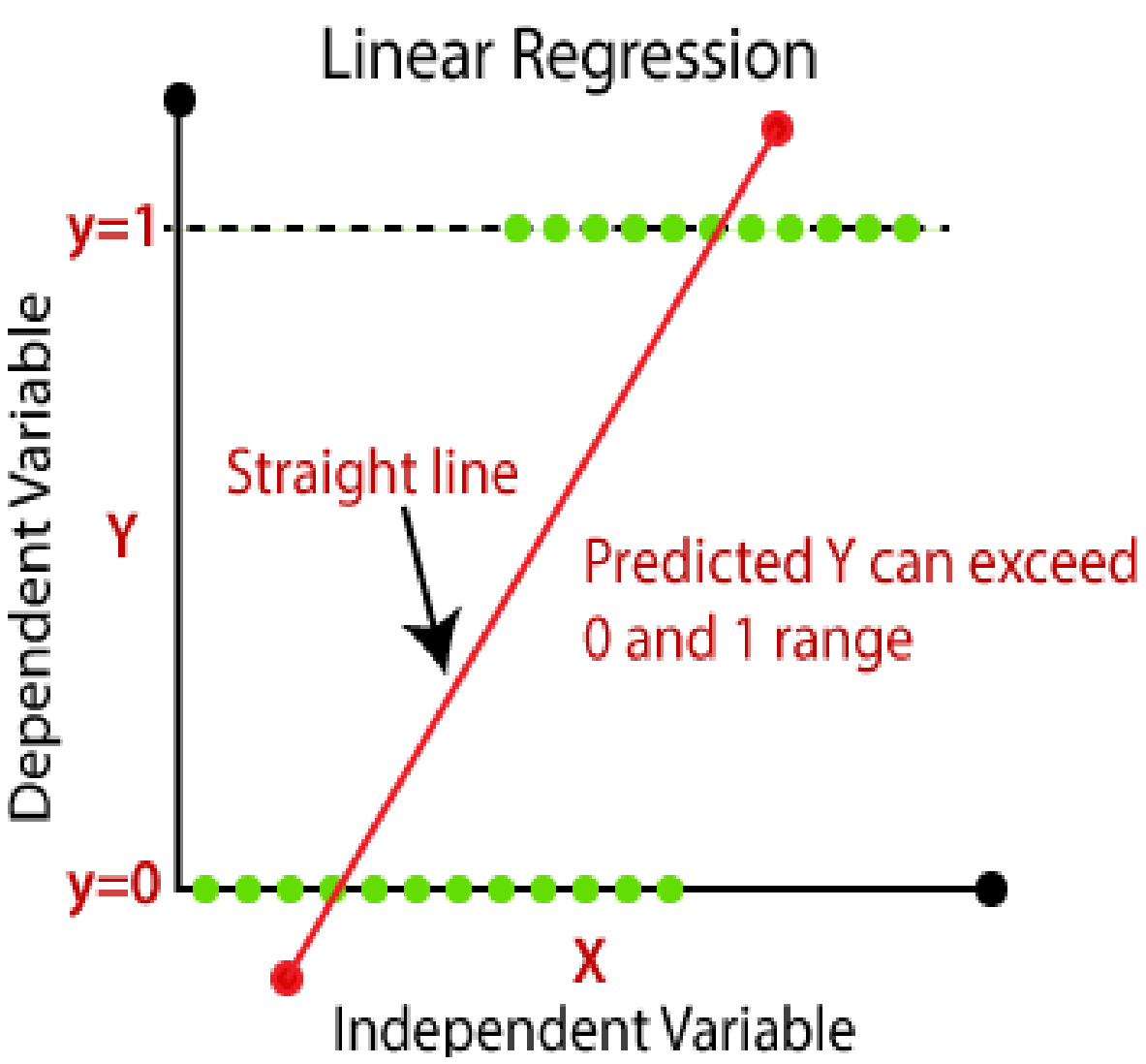

Temperature
Humidity
Wind
Logistic Regression Model
Sunny
Rainy

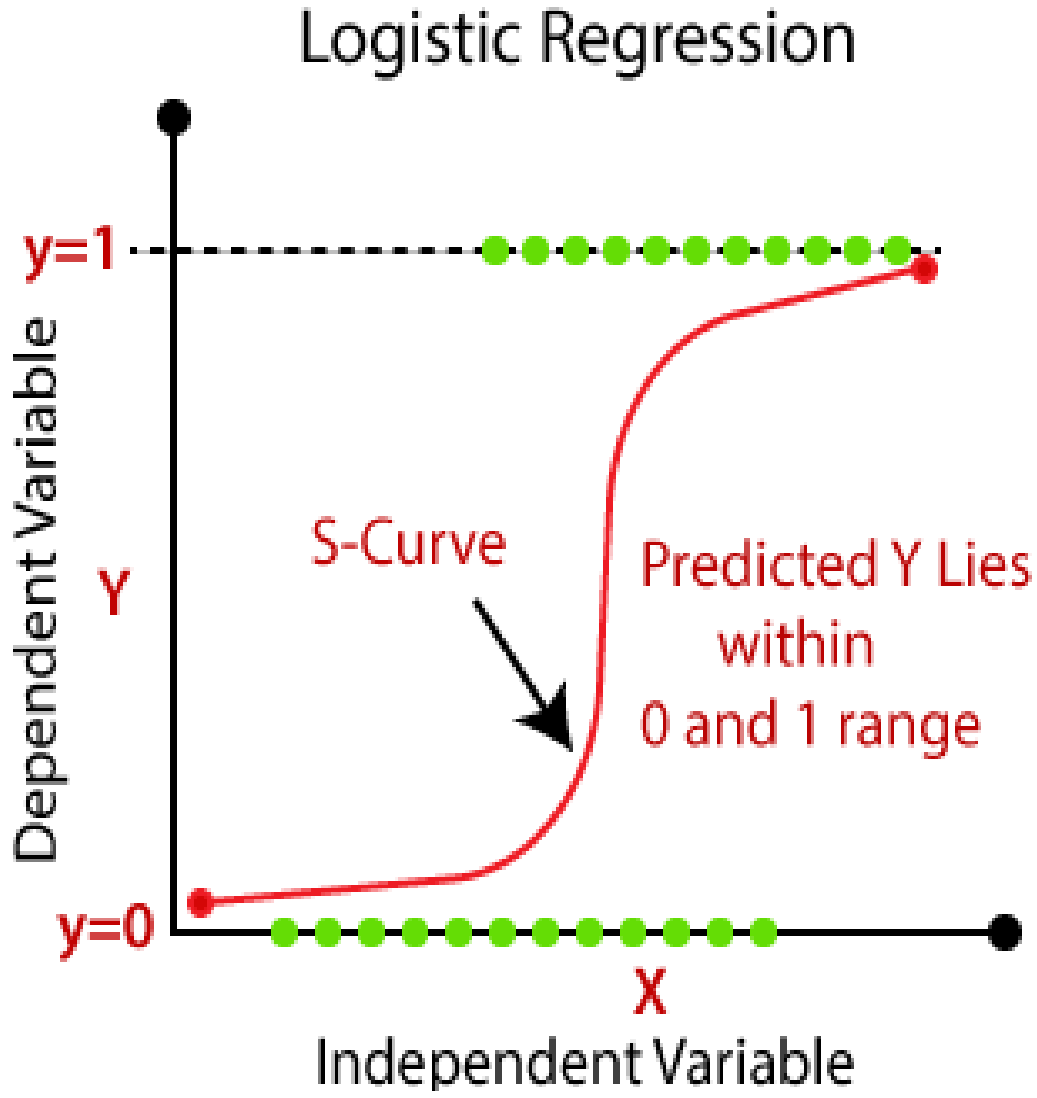
LOGISTIC REGRESSION MODELLING

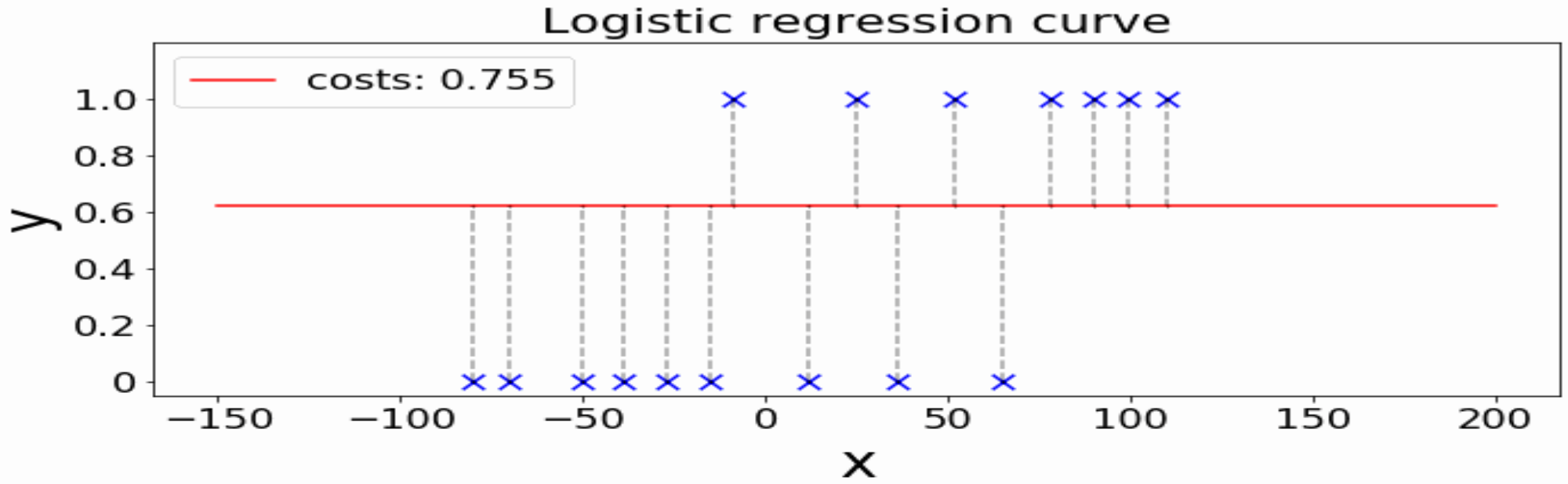


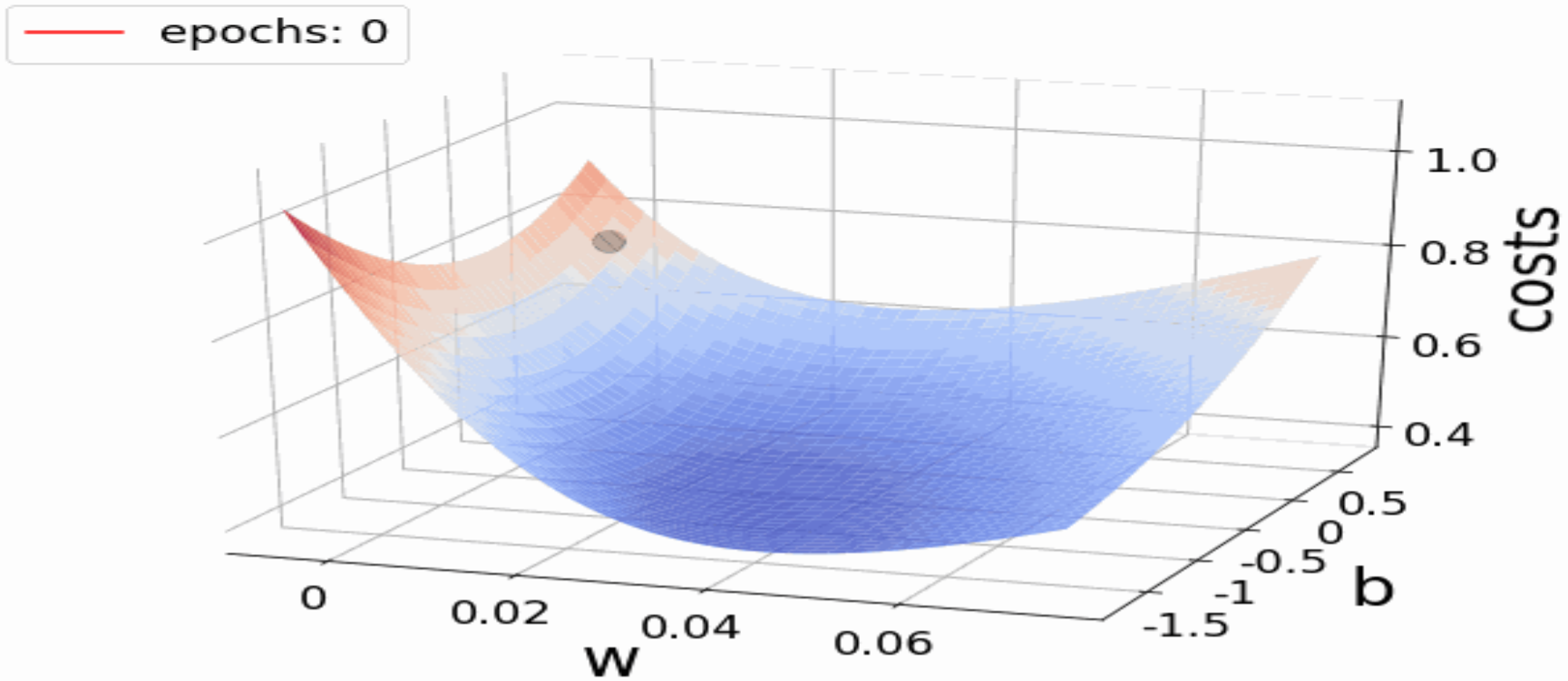
W1,w2,w3,w4 – Amount of each individual medical problem
P(x) – Probability Calculation

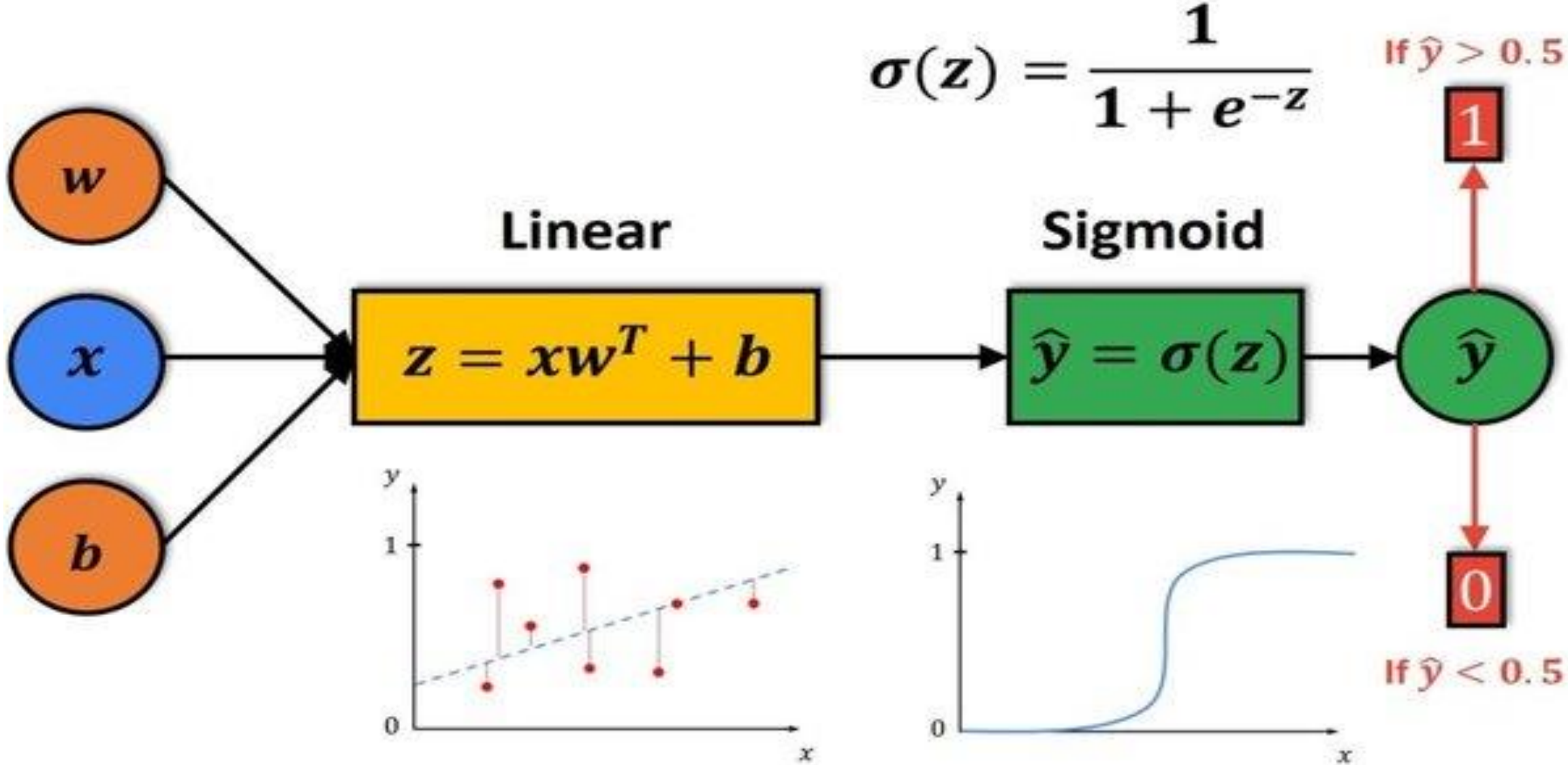
Logistic Regression


Linear Regression
y=1
y=0
Y
Dependent Variable
Straight line
Predicted Y can exceed
0 and 1 range
X
Independent Variable

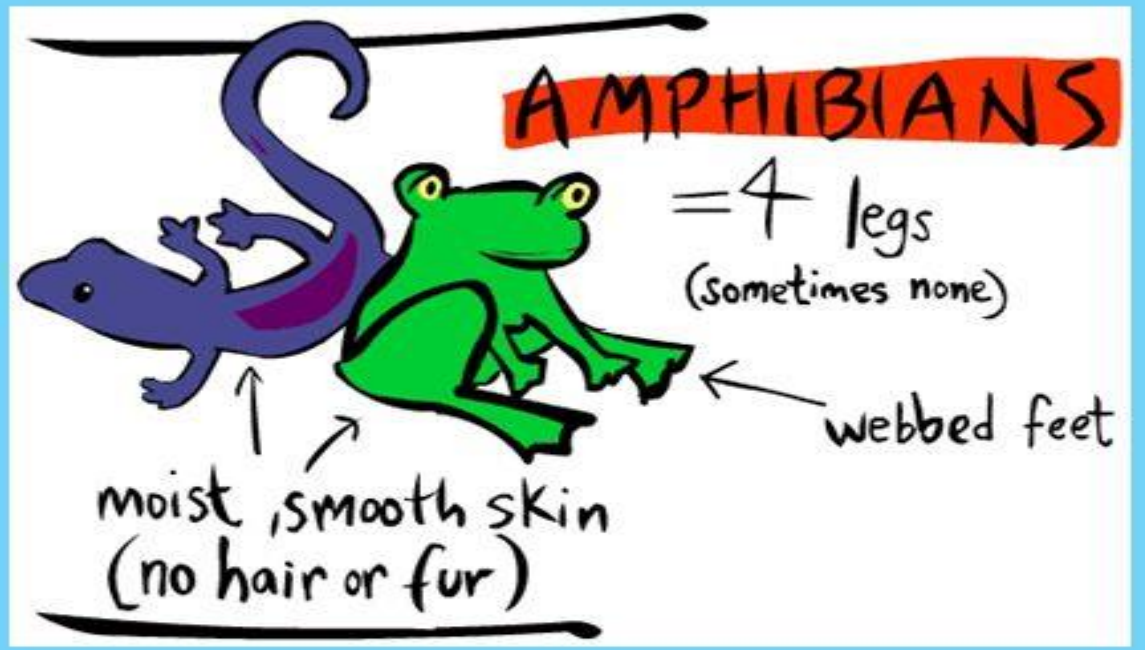

Logistic Regression
y=1
y=0
Y
Dependent Variable
S-Curve
Predicted Y Lies
within
0 and 1 range
X
Independent Variable


Logistic regression curve
costs: 0.755
y
x
1.0
0.8
0.6
0.4
0.2
0
−150
−100
−50
0
50
100
150
200

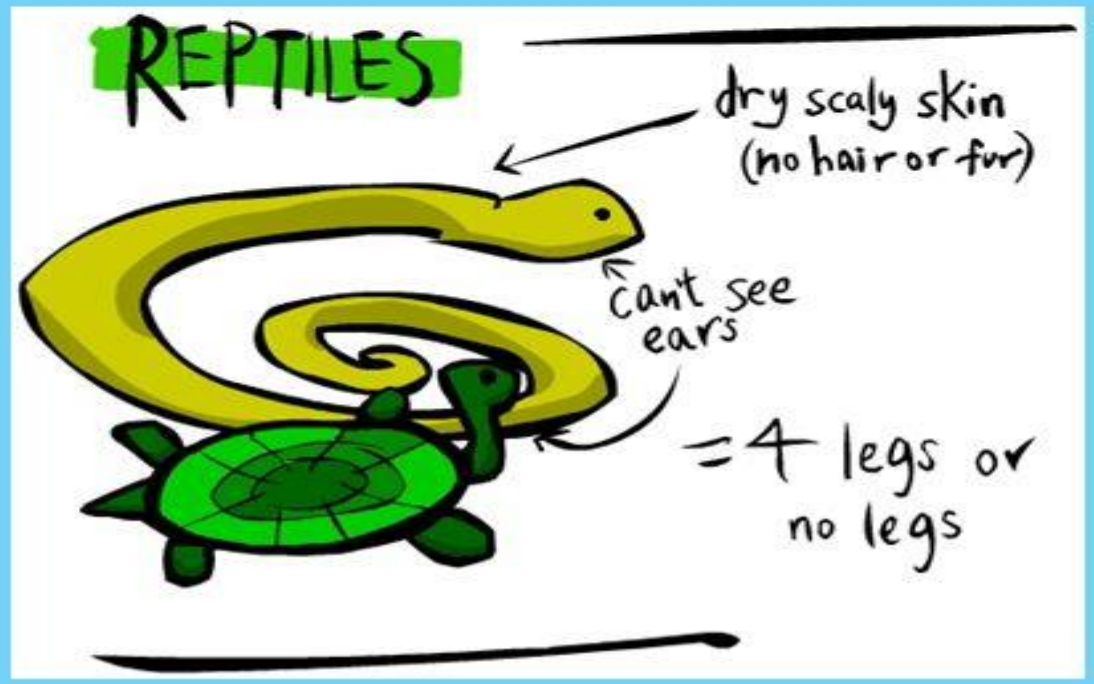

epochs: 0
costs
w
b
0
0.02
0.04
0.06
1.0
0.8
0.6
0.4
0.5
0
-0.5
-1
-1.5


$\sigma(z) = \frac{1}{1+e^{-z}}$
If $\hat{y} > 0.5$
1
Linear
Sigmoid
w
x
b
$z = xw^T + b$
$\hat{y} = \sigma(z)$
$\hat{y}$
y
1
0
x
y
1
0
x
0
If $\hat{y} < 0.5$

Amphibians



Reptiles



General Approach for Building Classification Model

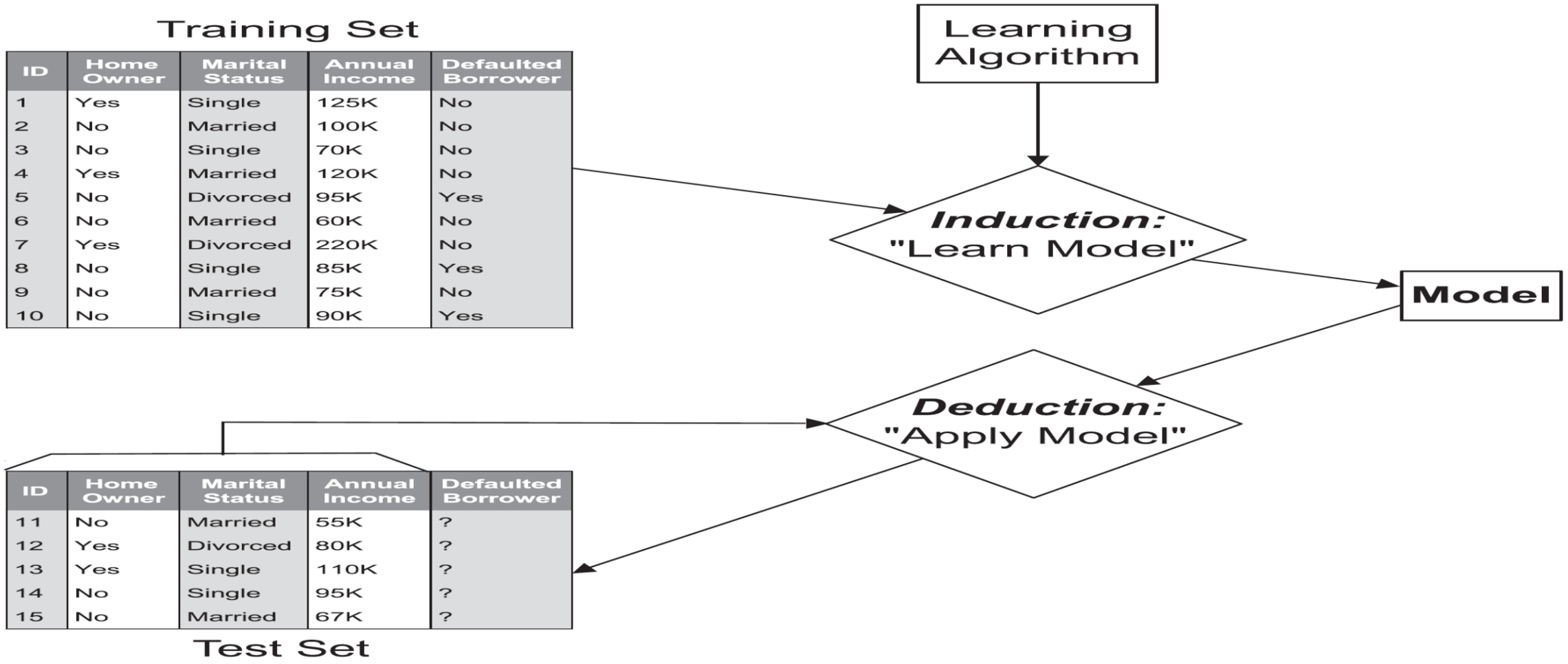


Figure 3.3. General framework for building a classification model.

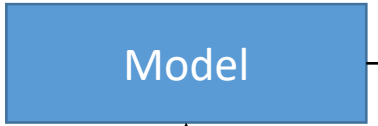
Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

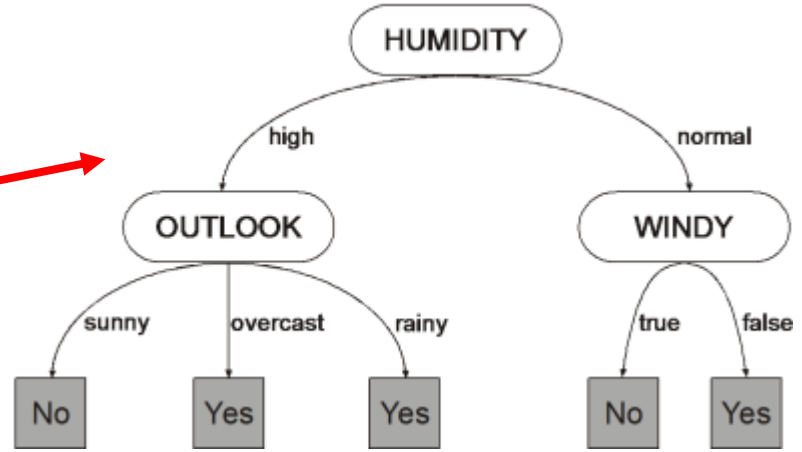
Classification



lion

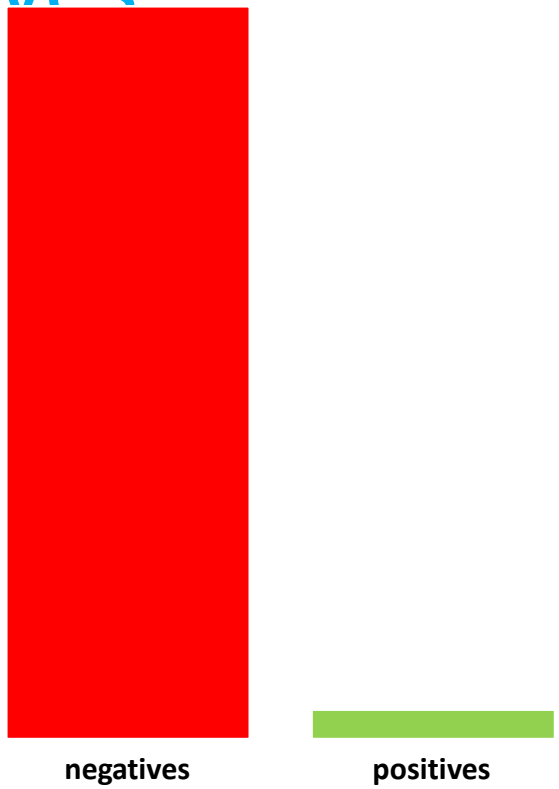
The resulting *model* is also called the *hypothesis*.

Given a model space and an optimality criterion, a *model* satisfying this criterion is sought.



Optimal tree!

After plotting your class distribution you see that you have thousands of negative examples but just a couple of positives



Performance metrics

- Most of the time accuracy will not be enough to assess performance.

- $accuracy = \frac{TP+TN}{P+N}$ Percentage of correctly classified instances.

- $sensitivity = \frac{TP}{P}$ The proportion of positives that are correctly identified as such.

- $precision = \frac{TP}{TP+FP}$ Equivalently, it is the fraction of relevant instances among the selected ones.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$ Matthews correlation coefficient (takes into account imbalance)

Confusion Matrix

- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Accuracy

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

- Most widely-used metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Problem Statement

- Titanic dataset
- **Explore:** How does each feature relate to whether a person survives/alives?
- Do the EDA in more detail than usual and explain the results!
  - Splitting: 80-20, stratify: y, random_state = 0
- **Preprocessing:**
  - * Drop decks
  - * Fill in the missing value using a simple imputer
  - * One hot encoding: sex, alone
  - * Ordinal encoding: class
  - * Binary encoding: embark town
- **Model selection:**
  - * Evaluation metrics used: F1_score
  - Logistic Regression