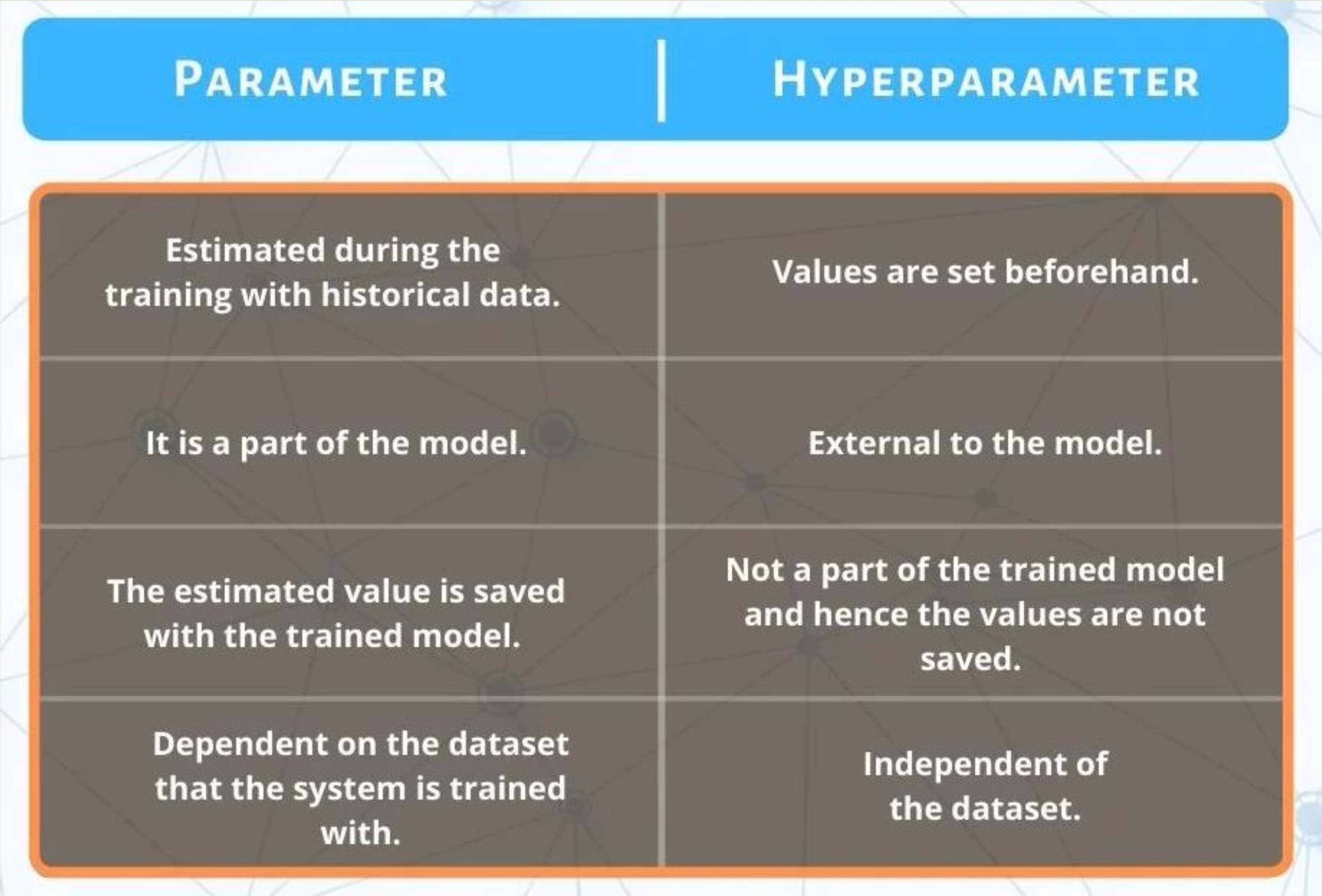
DATA SCIENCE FOR ENGINEERS

Week 7

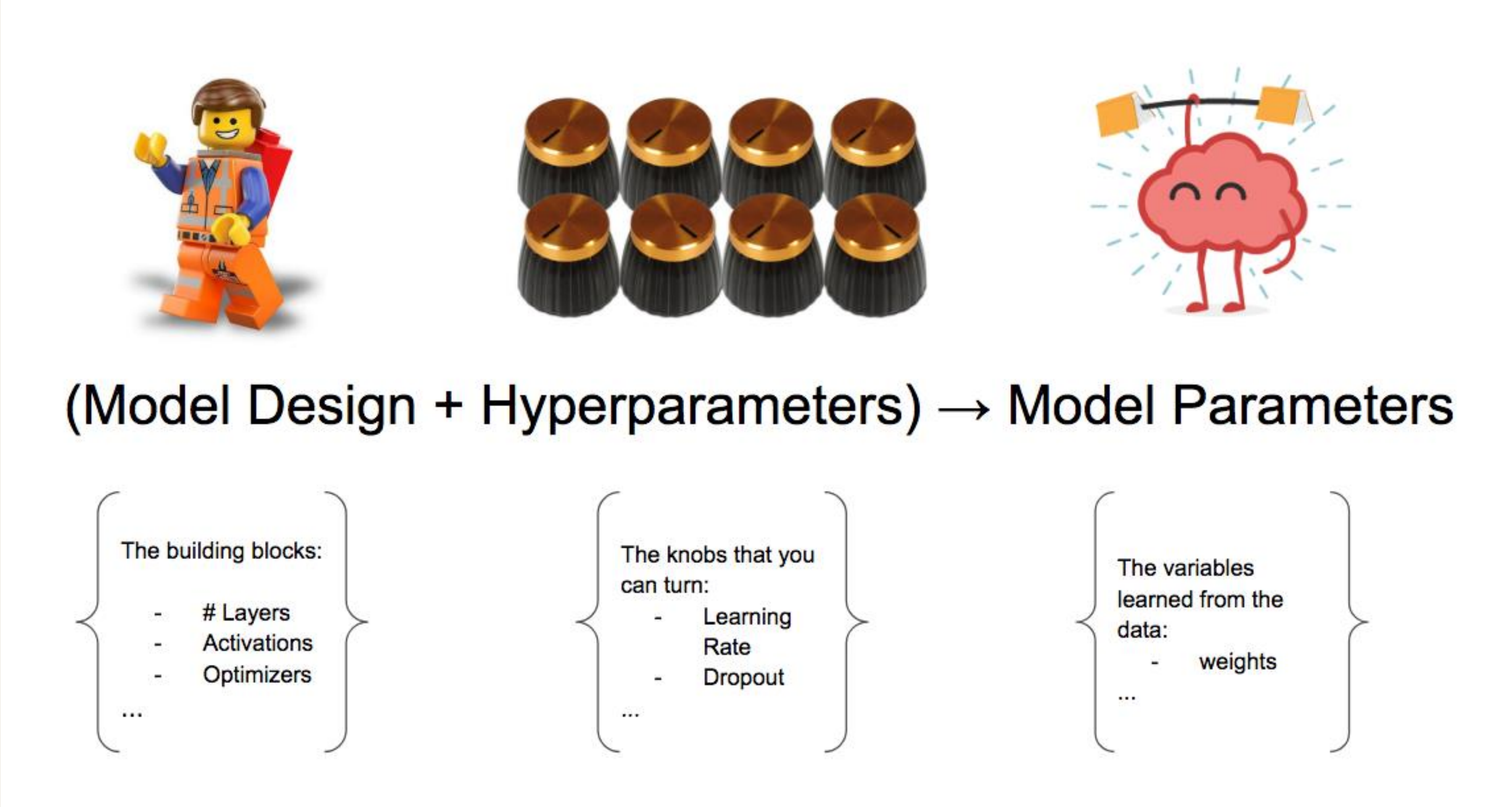
Session Co-Ordinator : Abhijit Bhakte

Parameter and Hyper parameter



Parameter	Hyperparameter
Estimated during the training with historical data.	Values are set beforehand.
It is a part of the model.	External to the model.
The estimated value is saved with the trained model.	Not a part of the trained model and hence the values are not saved.
Dependent on the dataset that the system is trained with.	Independent of the dataset.

Example: Neural Network Model



(Model Design + Hyperparameters) → Model Parameters

The building blocks:

- # Layers
- Activations
- Optimizers

...

The knobs that you can turn:

- Learning Rate
- Dropout

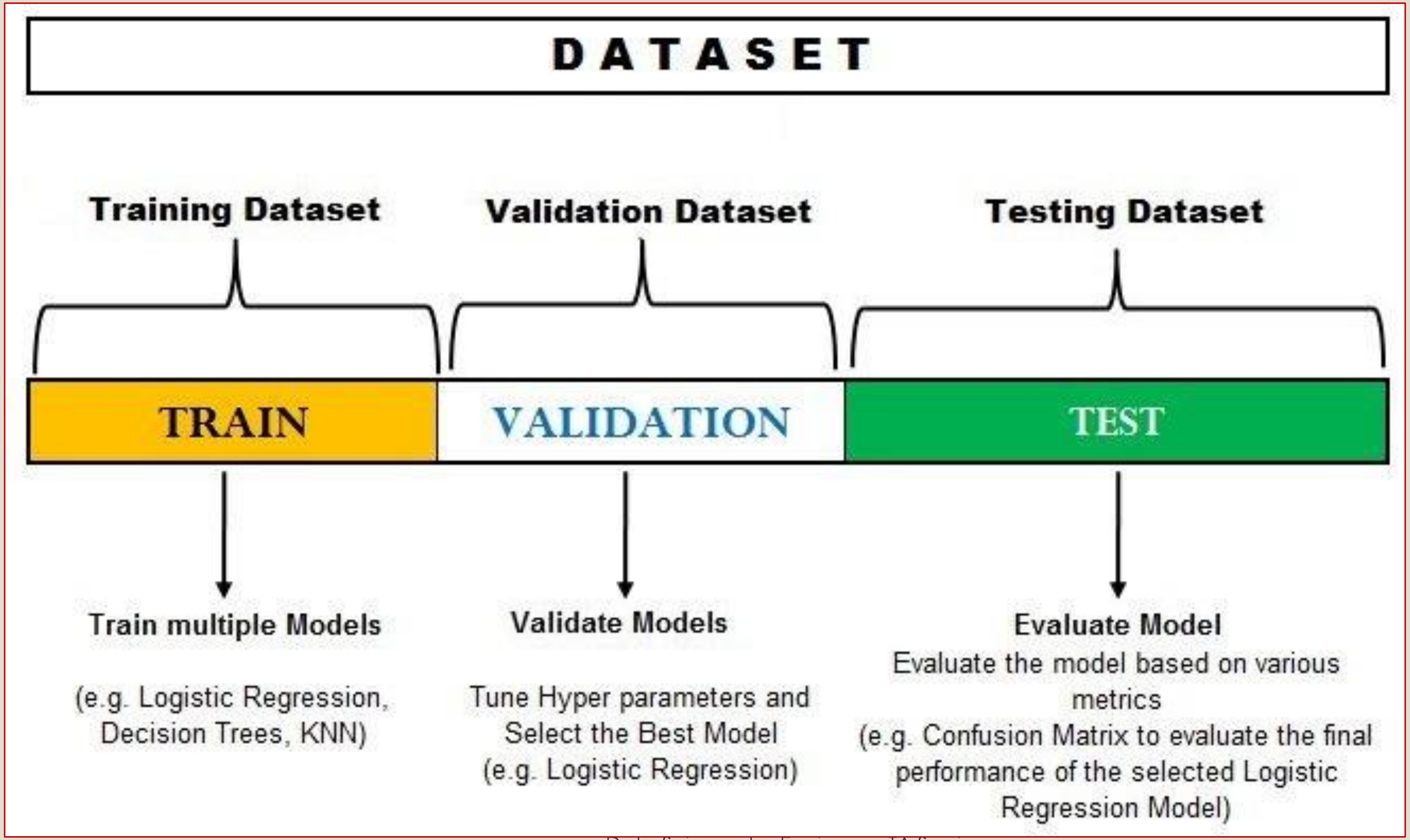
...

The variables learned from the data:

- weights

...

How to select hyparameters?



DATASET

Training Dataset

Validation Dataset

Testing Dataset

TRAIN

VALIDATION

TEST

Train multiple Models

(e.g. Logistic Regression, Decision Trees, KNN)

Validate Models

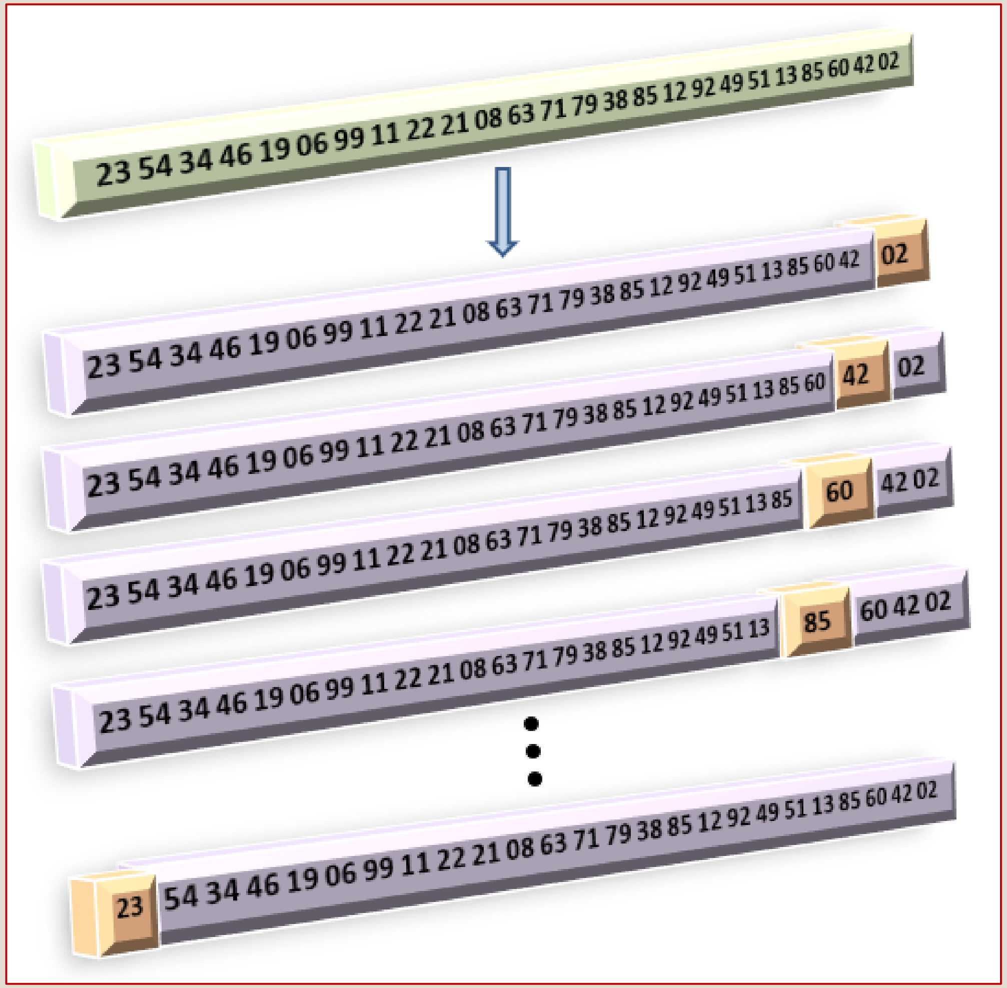
Tune Hyper parameters and Select the Best Model (e.g. Logistic Regression)

Evaluate Model

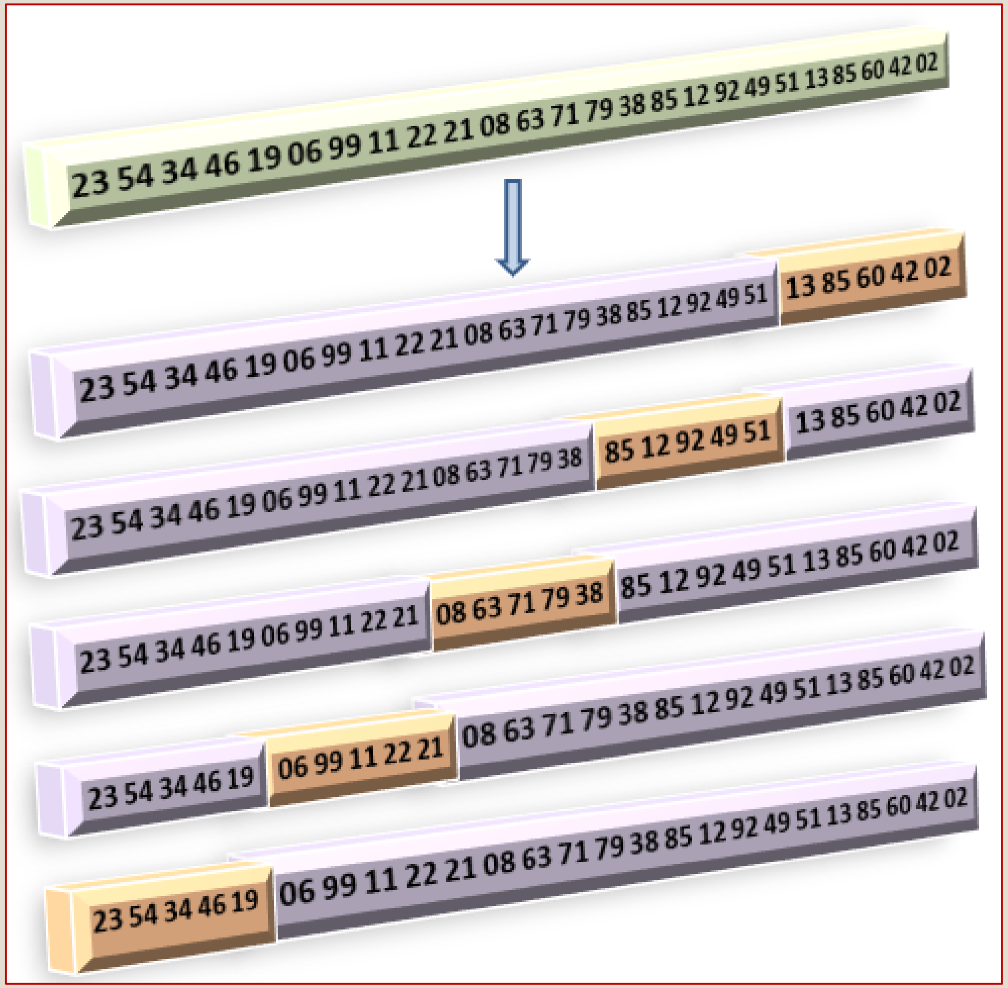
Evaluate the model based on various metrics (e.g. Confusion Matrix to evaluate the final performance of the selected Logistic Regression Model)

Methods

- **Leave one out cross validation (LOOCV)**



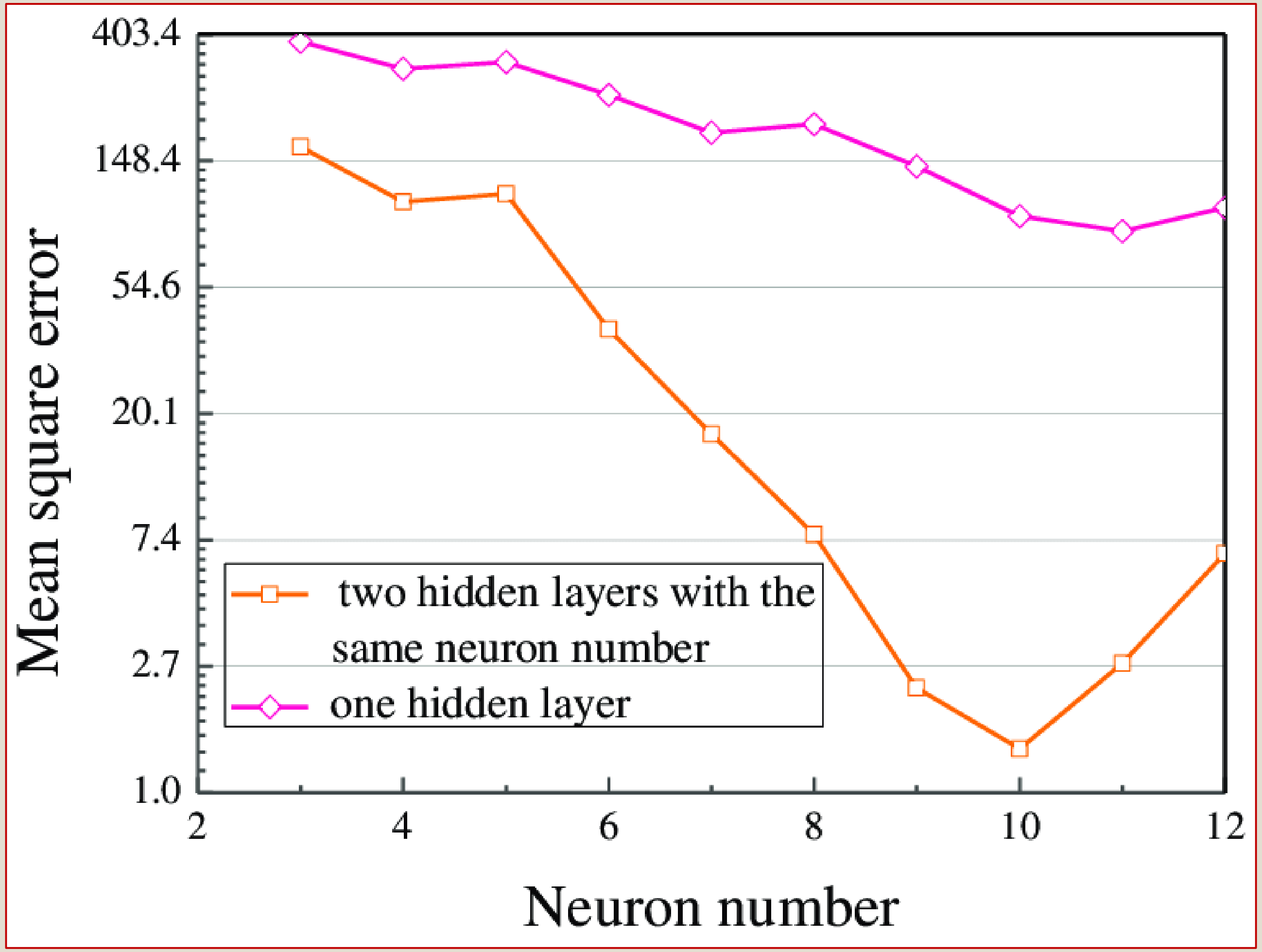
- **K-fold cross validation**



Selection of hyperparameter

- Given graph is for the Neural network model that predicts the square of a number

- When cross validation is performed the best hyper parameters are chosen as **2 hidden layers with 10 neuron**



Q) What is the primary purpose of cross-validation in machine learning?
A) To train multiple models simultaneously for faster convergence.
B) To evaluate a model's performance and assess its generalization to unseen data.
C) To increase the complexity of a model for better accuracy.
D) To reduce overfitting by adding more training data.

Solution

It helps in estimating how well a model will perform on new, unseen data by simulating the process of training and testing on different subsets of the data.

Q) Which of the following is not a commonly used cross-validation technique?

A) K-Fold Cross-Validation
B) Leave-One-Out Cross-Validation (LOOCV)
C) Stratified Cross-Validation
D) Train-Test Split

Solution

It involves splitting the dataset into two parts: a training set and a test set, where the model is trained on the training set and evaluated on the test set. While it is a common method for evaluation, it is not a cross-validation technique.

Q) Which statement best describes Leave-One-Out Cross-Validation (LOOCV)?

A) It is computationally efficient for large datasets.
B) It divides the data into K subsets and uses K-folds for training and 1 fold for testing.
C) It uses a single data point for testing and the remaining data for training.
D) It is primarily used for time series data.

Q) Which of the following is an example of a hyperparameter?

A) The weights of a neural network's hidden layers.
B) The learning rate of an optimization algorithm.
C) The input features of a dataset.
D) The training data used for model training.

Solution

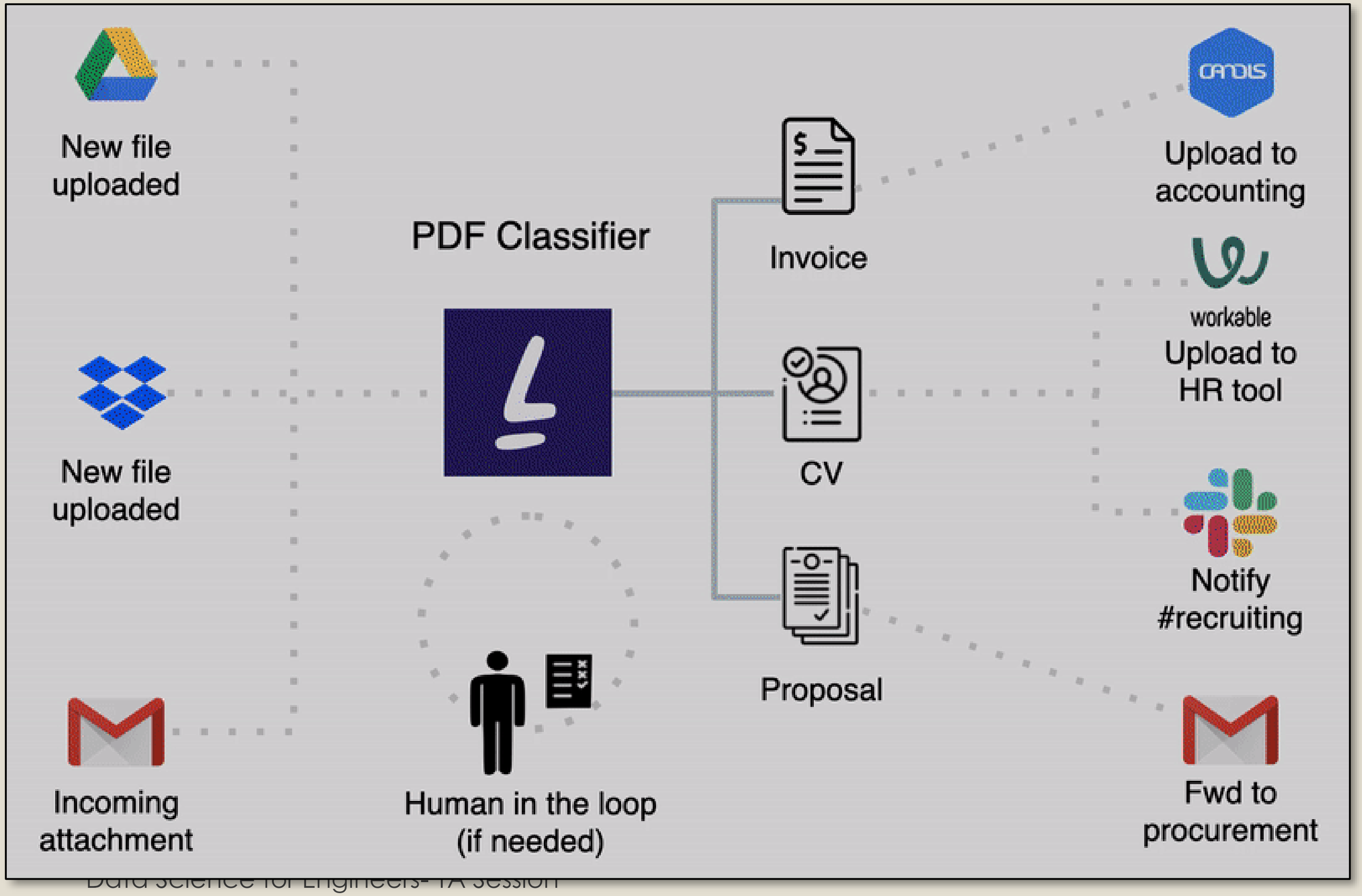
A hyperparameter is a configuration setting that is set before the model training begins

Classification

- Classification is a supervised learning that **assign class to the data points**

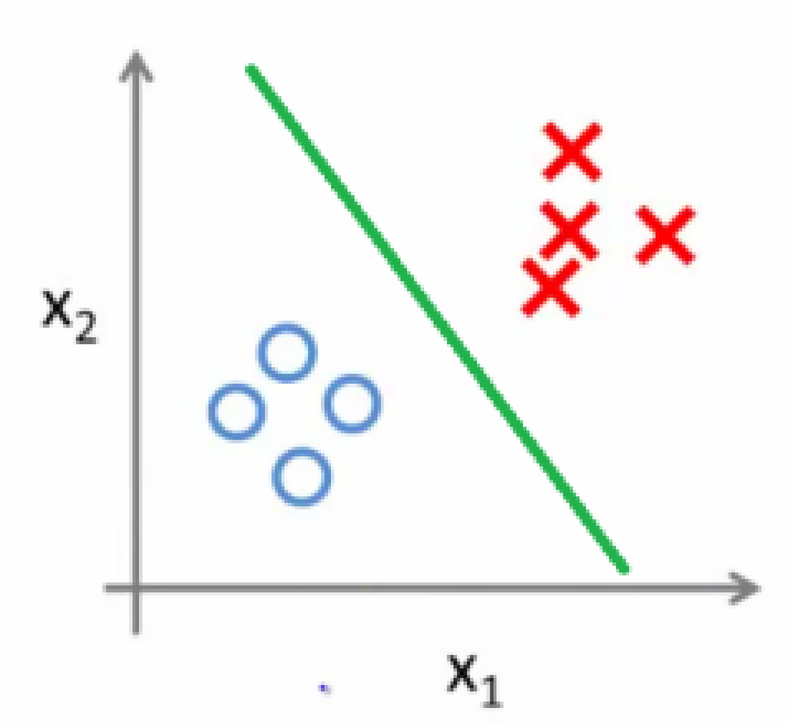
- Examples:

- Email Spam Detection
- Sentiment Analysis
- Image Classification
- Medical Diagnosis
- Fraud Detection
- Language Identification

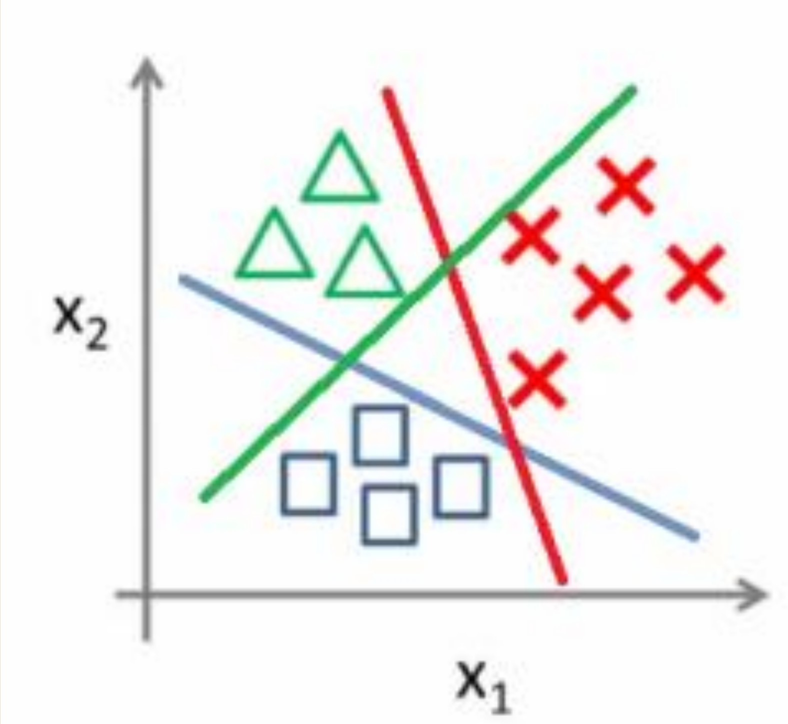


Types of Classification

Binary classification
Includes two Classes

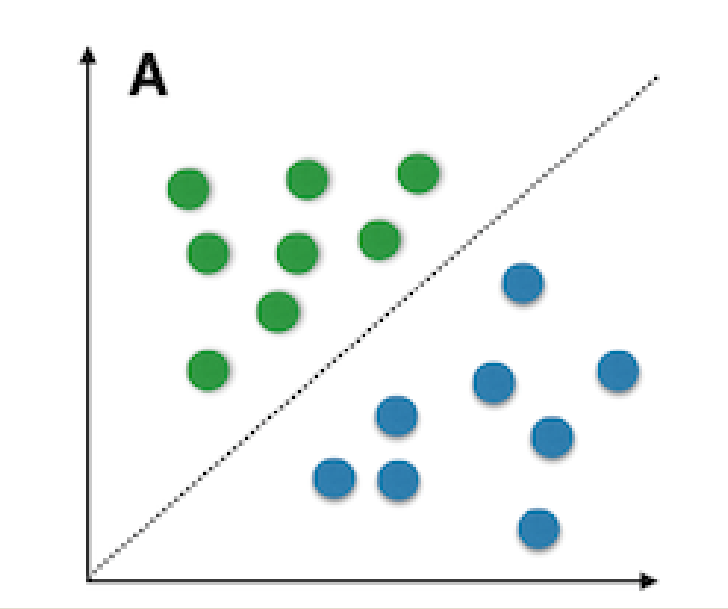


Multiclass classification
Includes more than two Classes

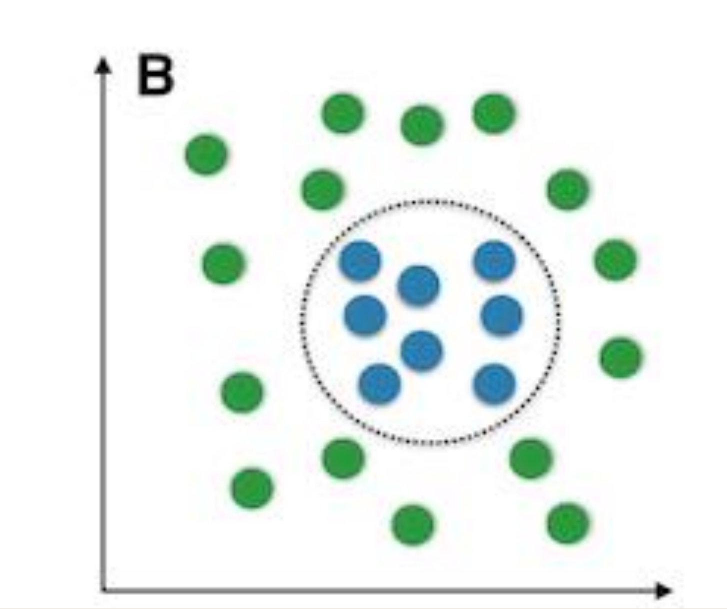


Types of Classification

Linearly Separable Problem
Linear Boundary

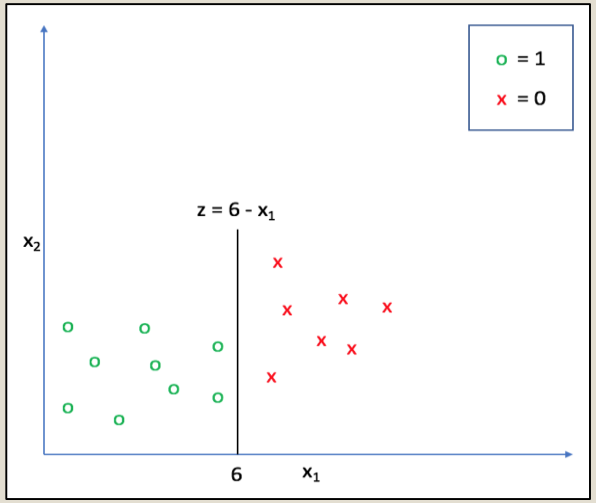


Non-linearly Separable Problem
Nonlinear Boundary



Logistic Regression

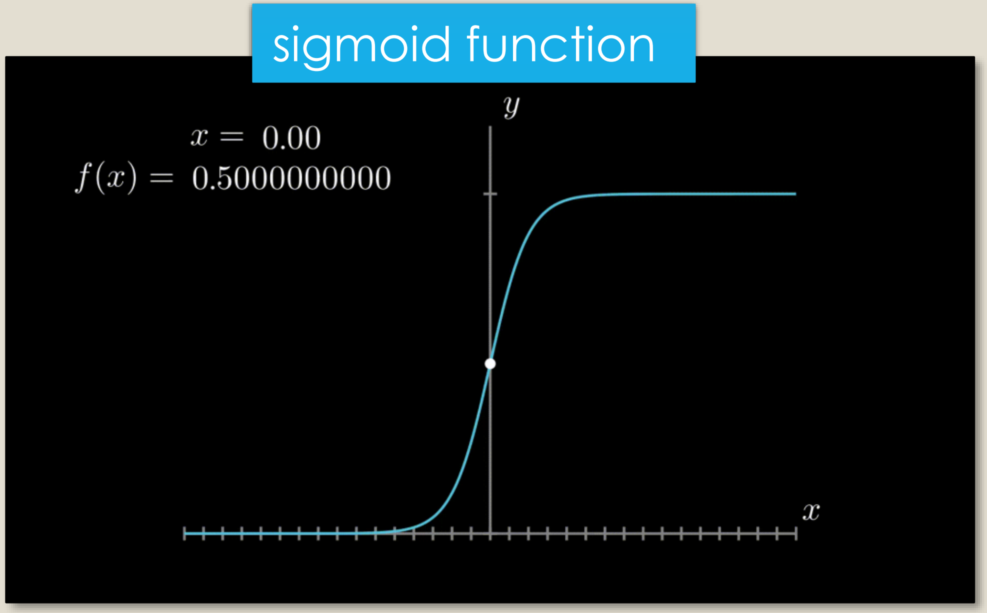
- ➢ Classification is the technique which draw linear boundary
- ➢ Goal: Given new sample data predict the class from which the data point likely to have originated
- ➢ Simply guess of the class is not the good way to classify the sample hence probability is introduces to provide better understanding



How to model probability

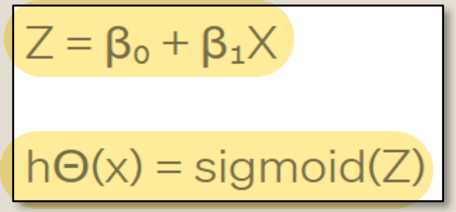
- To classify the two classes we use decision boundary (linear equation)
- The value of equation may vary from (-inf to +inf)
- To bring this value in (0 to 1) range we use sigmoid function

$$f(x) = \frac{1}{1 + \mathrm{e}^{-(x)}}$$



Logistic Regression

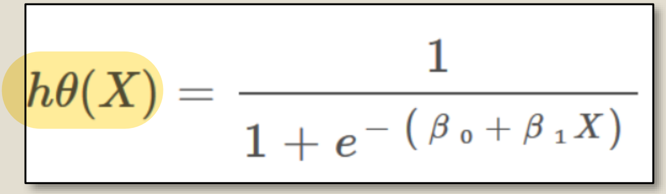
- We have linear boundary equation



Z = β₀ + β₁X

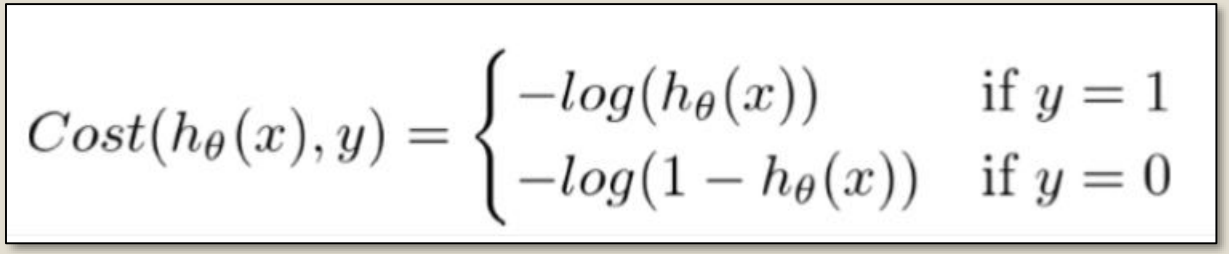
hΘ(x) = sigmoid(Z)

- Applying Sigmoid function



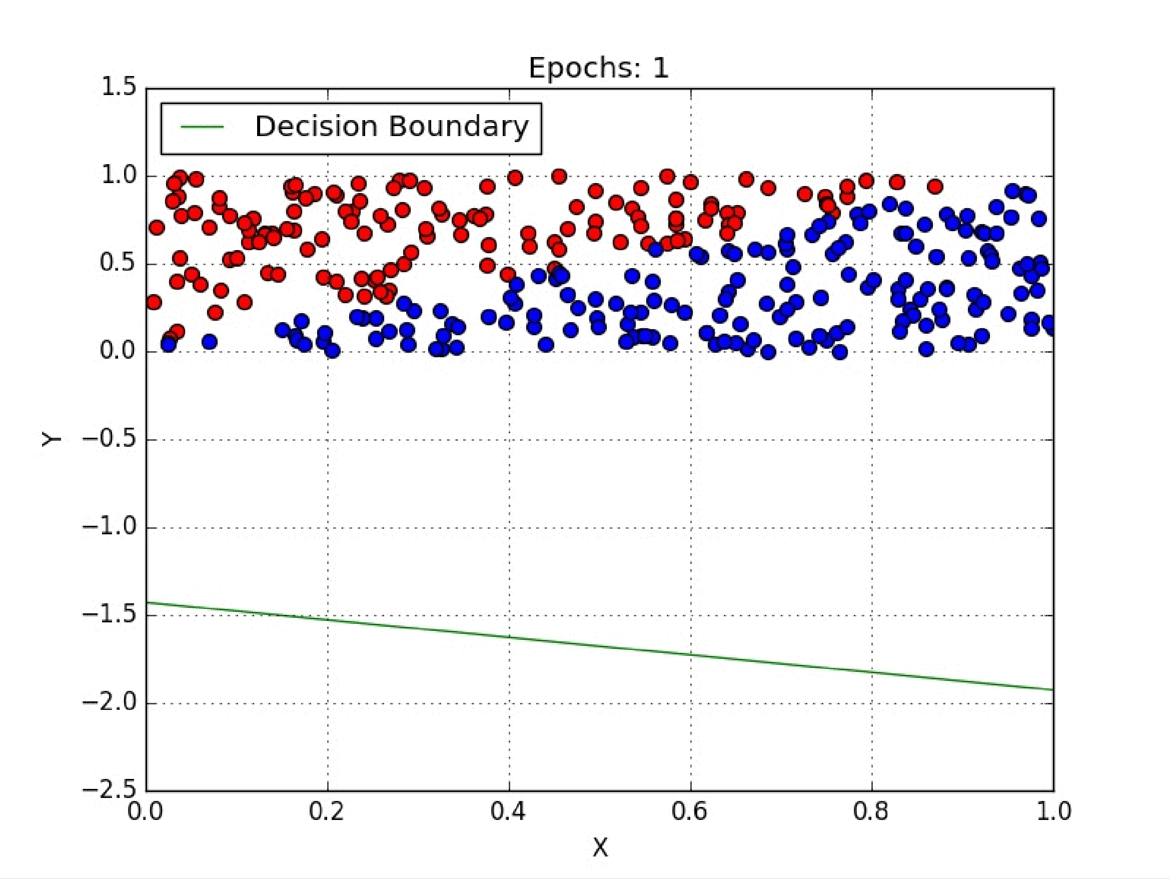
$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- To learn the model parameter we use loss function



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

How the model learns during training



Q) What type of machine learning problem is logistic regression primarily used for?

A) Regression
B) Classification
C) Clustering
D) Dimensionality Reduction

Solution

Logistic regression is a classification algorithm used to model the probability of a binary outcome.

Q) In logistic regression, what is the output range of the logistic function (sigmoid function)?

A) [-1, 1]
B) [0, 1]
C) [0, ∞)
D) (-∞, ∞)

Solution

The logistic function outputs probabilities between 0 and 1, making it suitable for binary classification.

Q) What is the purpose of the sigmoid function in logistic regression?

A) To convert odds to probability.
B) To model the linear relationship between features and target.
C) To normalize the feature values.
D) To calculate the mean squared error.

Solution

The sigmoid function is used to map the log-odds (logit) to a probability value between 0 and 1.

Q) In logistic regression, what is the cost function that is minimized during training?

A) Mean Absolute Error (MAE)
B) Mean Squared Error (MSE)
C) Cross-Entropy Loss (Log Loss)
D) Root Mean Square Error (RMSE)

Solution

The cross-entropy loss, also known as log loss, is used as the cost function in logistic regression.

Q) In logistic regression, how are model coefficients (weights) typically determined during training?

A) Randomly initialized
B) Calculated using gradient descent
C) Set to 1 for all features
D) Assigned based on feature importance

Solution

Gradient descent is commonly used to iteratively update model coefficients to minimize the cost function.

Q) Which evaluation metric is commonly used to assess the performance of a logistic regression model?

A) R-squared (R^2)
B) Mean Absolute Error (MAE)
C) Accuracy, Precision, Recall, F1-Score
D) Root Mean Square Error (RMSE)

Solution

Logistic regression models are often evaluated using classification metrics such as accuracy, precision, recall, and F1-Score, depending on the problem and requirements.

R Studio

- **Wheat Dataset**
- **Input variables:** Perimeter, Area, Compactness, length and width of kernel …(# of features = 7)
- **Output Labels:** Seed Types (Types of seed = 3)

Q) What is one-hot encoding used for in classification?

A) Reducing the dimensionality of data.
B) Encoding categorical variables as binary vectors.
C) Scaling numerical features.
D) Visualizing data in scatter plots.

Solution

One-hot encoding is a technique used to represent categorical variables as binary vectors to make them compatible with machine learning algorithms.

Q) In binary classification, what does precision measure?

A) The ability to correctly identify positive instances.
B) The ability to correctly identify negative instances.
C) The ratio of true positives to all positive predictions.
D) The ratio of true negatives to all negative predictions.

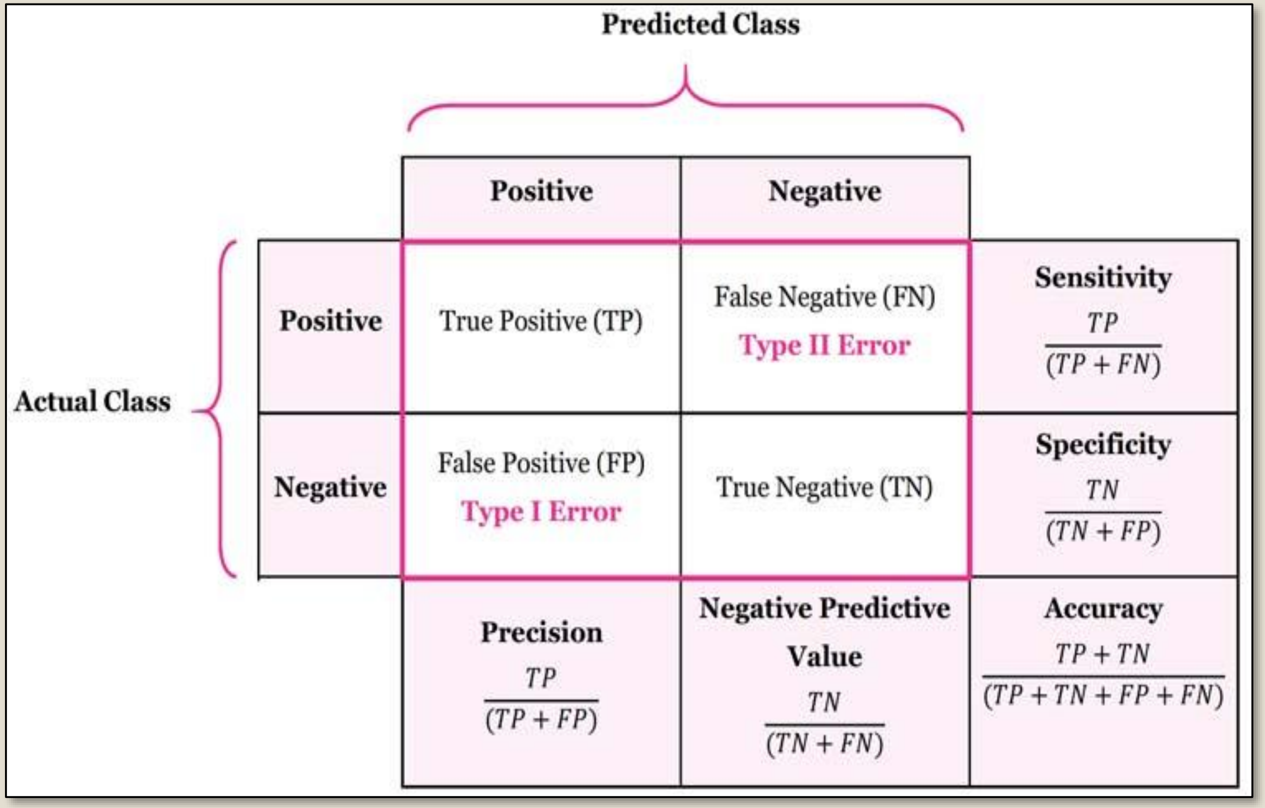
Solution

Precision measures how many of the positive predictions made by a model are actually correct, indicating the model's ability to identify positive instances accurately.

Performance metric

Model

Patient



		Predicted Class	
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Actual	Covid-Test	Type
Positive	Positive	TP
Positive	Negative	FN
Negative	Positive	FP
Negative	Negative	TN

Q) What is the purpose of a confusion matrix in classification?

A) To visualize data in 3D space.
B) To measure the accuracy of a regression model.
C) To assess the performance of a classification model.
D) To calculate the mean squared error.

Solution

A confusion matrix is used to evaluate the performance of a classification model by providing information about true positives, true negatives, false positives, and false negatives.

Q) In binary classification, what does precision measure?

A) The ability to correctly identify positive instances.
B) The ability to correctly identify negative instances.
C) The ratio of true positives to all positive predictions.
D) The ratio of true negatives to all negative predictions.

Solution

Precision measures how many of the positive predictions made by a model are actually correct, indicating the model's ability to identify positive instances accurately.

Q) In a binary classification problem, if a model makes 80 true positive predictions, 10 false positive predictions, 5 false negative predictions, and 105 true negative predictions, what is the accuracy of the model?

A) 0.44
B) 0.92
C) 0.96
D) 0.90

Solution

Acc = (TP+TN) / (Total Predictions)
Acc = (80 + 105) / (80 + 10 + 5 + 105)
Acc = 185 / 200 = 0.925

Q) A classification model predicts 120 instances as positive, out of which 100 are actually positive. What is the precision of the model?

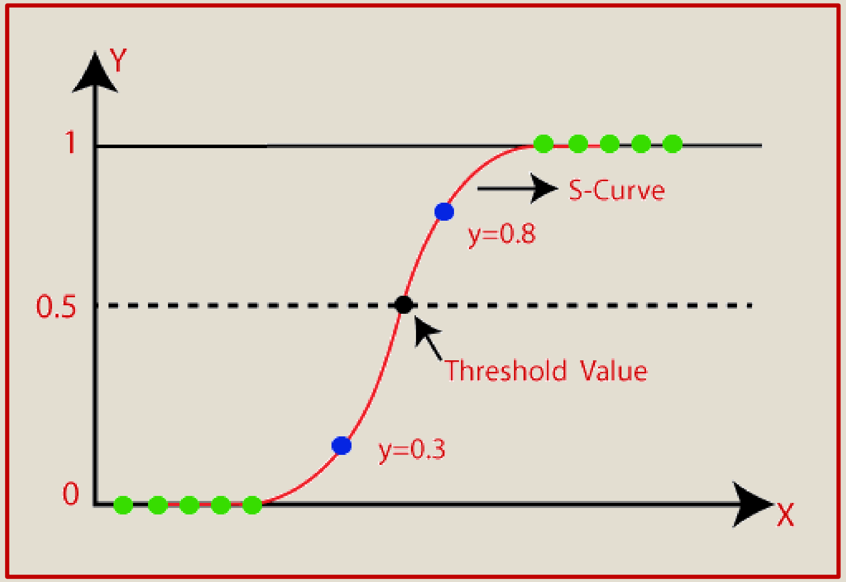
A) 0.95
B) 0.90
C) 0.83
D) 0.75

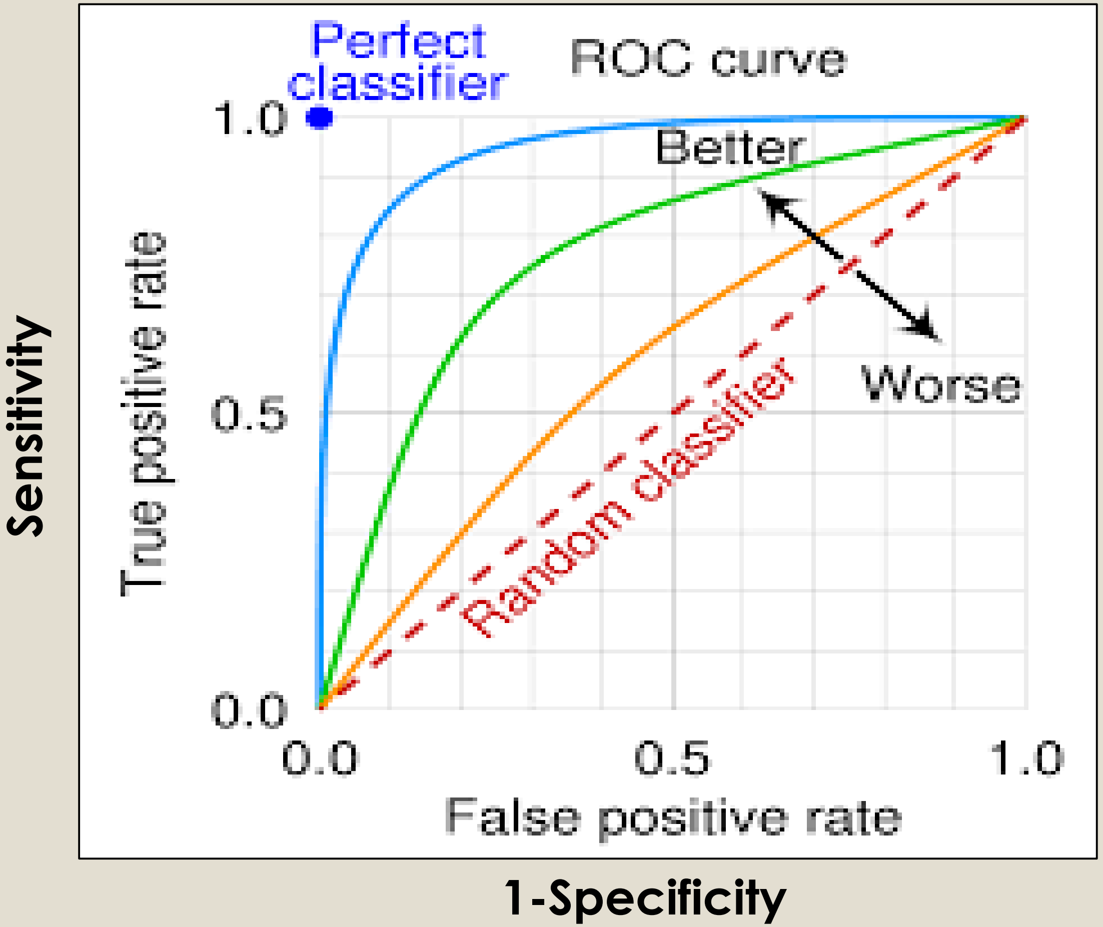
Solution

Precision = (TP) / (TP+ FP)
Precision = 100 / (100 + 20) = 0.83

ROC-(receiver operating characteristic) curve

- ROC Curve is a graphical representation of a model's ability to distinguish between two classes.
- Used in classification problems, particularly in machine learning and medical diagnosis





Q) If a classifier has a true positive rate of 0.90 and a false positive rate of 0.15, what is the specificity of the classifier?

A) 0.15
B) 0.10
C) 0.85
D) 0.90

Solution

Specificity = 1 - False Positive Rate
Specificity = 1 - 0.15 = 0.85

Q) A classification model has 120 true negatives and 30 false positives. What is the false positive rate of the model?

A) 0.20
B) 0.25
C) 0.80
D) 0.10

Solution

FPR = (FP) / (FP+TN)
FPR = 30 / (30 + 120) = 0.2