SUPERVISED LEARNING

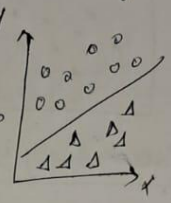
CLASSIFICATION

It is a supervised learning alg that is used to identify the category of new observations on the basis of training data.

→ In the program classifies new observation into a number of class or groups. like yes or no / 0 or 1 / spam or Not spam

→ Classes can be called as targets / labels or categories.

eg: spam detector.



The alg. that implements the Classification on a dataset is known as classifier.

2 types:

Binary Classifier: If the outcome has only 2 possible outcomes
eg: Y/N, Male or F

Multi-class classifier: If it has more than 2 outcomes
eg: types of crops, types of music

Types of learners:

Lazy learners – first stores training dataset & wait until it receives the test dataset. The classi is done based on the most related data stored in training.
less time in training & more time for prediction

eg: kNN, case based reasoning

2) Eager learners: learners develop a classification model (5) based on training data set before receiving a test data. More time in learning and less time in prediction

eg: Decision trees, Naive Bayes, ANN.

Types of classification Alg:

Linear model	Non linear
Logistic Regression	K-NN
SVM	Kernel SVM
	Naive Bayes
	Decision tree
	Random forest.

Evaluating a classification Model:

1) Log loss (or) Cross-entropy loss:

→ Used for evaluating performance of classifier, whose o/p is a probability value b/w 0 and 1

→ For good binary classifier log loss should be near to 0.

→ lower log loss = higher accuracy of the model.

2) Confusion Matrix: / error matrix

→ provides matrix / table as o/p & describe performance.

	Actual true	Actual -ve
Predicted true	True positive	false true
predicted -ve	False -ve	True negative

Multivariate Classification:

To predict the label (or category) of an observation (or sample) based on multiple features or variables.

In this it takes into multiple features simultaneously making it more powerful.

EX: predicting flower species. – Iris dataset.

Dataset Overview:

Features (I/p variables) – sepal & petal length sepal & petal width

Labels (O/p categories) – Iris setosa, Iris Versicolor, etc.

Process:

1) data collection
2) preprocessing
3) Model selection – decision tree, Random forest etc.
4) Training
5) Prediction
6) Evaluation

Benefits:

Improved accuracy

Better feature interaction Understanding

Applicability to complex problems.