Fundamentals of Machine Learning

Sourav Karmakar

souravkarmakar29@gmail.com

OUTLINE

- Introduction to Machine Learning
- Supervised Learning
- Unsupervised Learning
- Machine Learning Process
- Classification and Classifier
- Data and Features
- Training and Validation
- Evaluation of Binary Classifier

INTRO TO MACHINE LEARNING

What is Machine Learning?

- *"The field of study that gives computers the ability to learn without being explicitly programmed"* – Arthur Samuel

- *"A computer program is said to learn from experience 'E' with respect to some class of tasks 'T' and performance measure 'P', if its performance at tasks in T, as measured by P, improves with experience E."* – Tom M. Mitchell

Let's try to understand the definition given by Tom Mitchell with few examples…

INTRO TO MACHINE LEARNING

Example-1:

Classifying Emails as Spam or Not Spam:

- **Task (T):** Classifying emails as spam or not spam
- **Experience (E):** A dataset of emails, each labelled by human as spam or not spam
- **Performance (P):** The number or fraction of emails correctly classified as Spam or Not Spam

Example-2:

Recognizing Hand-written Digits :

- **Task (T):** Recognizing Hand written digits
- **Experience (E):** Labelled dataset of hand-written digits of 10 classes (0-9) & identify the underlying pattern
- **Performance (P):** The number or fraction of hand-written digits correctly classified

INTRO TO MACHINE LEARNING

Example-3:

Translating sentences from English to German:

- **Task (T):** Translating sentences from English to German
- **Experience (E):** Text corpora consist of translations from English to German
- **Performance (P):** The number or fraction of sentences correctly translated from English to German

Example-4:

Movie recommendation:

- **Task (T):** Recommending movies to a user that they are likely to enjoy
- **Experience (E):** A history of movies watched by user, their ratings of those movies
- **Performance (P):** The average rating of recommended movies by the user or the user engagement with the platform.

INTRO TO MACHINE LEARNING

When do we use Machine Learning

Machine learning is used in cases where:

- There is an intuition that a certain rule exists
- But, we do not know it or cannot express it explicitly

So, the machine / computer learns the rule from data…

The performance of the learning agent (in this case the machine / computer) should improve based on experience in either of the following ways:

1. The range of the performance is expanded: *The machine can do more.*
2. The accuracy on tasks is improved: *The machine can do things better.*
3. The speed is improved: *The machine can do things faster.*

TYPES OF MACHINE LEARNING

Machine Learning is a *"data oriented"* discipline.

Depending upon how the machine learning models learn the rule inherent in the dataset there are various types of learning methods:

- **Supervised Learning:** In this type of learning the ground truth of the training dataset is known. i.e. the dataset is already labelled.
- **Unsupervised Learning:** In this type of learning the ground truth of the training dataset is not known. i.e. the dataset is not labelled.
- **Semi-supervised Learning:** In this type of learning few of the training dataset are labelled. It learns based on some assumptions on the dataset.
- **Reinforcement Learning:** In this type of learning the agent will be penalised/rewarded based on the incorrect/correct prediction. The task of the agent is to maximize the reward.

SUPERVISED LEARNING

Supervised Learning: Learning under supervision

Teacher teaches child

Child recognizes an apple when she sees it again

The machine learning model is also able to learn from past data and make prediction / classification. This is called **Supervised Learning**.

SUPERVISED LEARNING

Example of Supervised Machine Learning - 1

Sl. No.	Age	Weight (kg)	Systolic Blood Pressure (mmHg)
1	39	67	144
2	47	78	180
3	45	61	138
. . .	. . .	. . .	. . .
240	21	72	148
241	44	63	127
242	63	72	170

- **Task:** Predicting the blood pressure of an adult based on his/her Age and Weight.
- Here **Age** and **Weight** are two predictors / input variables.
- Here **Blood Pressure** is the target / output variable.
- There are total 242 numbers of examples present in the dataset.
- Each training sample is of the form: ⟨(Age, Weight), Blood Pressure)⟩

- In the supervised learning framework we shall use a large fraction of our data to *"teach"* the model, so that the model can learn and predict the rest of the data with high *confidence*.

SUPERVISED LEARNING

Example of Supervised Machine Learning - 2

Sl. No.	Age	Monthly Income (Rs.)	Existing Loan (Lakh)	Loan Sanction (1 = Yes, 0 = No)
1	39	47000	2.5L	0
2	47	32000	0L	0
3	45	65000	3L	1
. . .	. . .	. . .	. . .	. . .
998	21	12000	0L	0
999	44	38000	1L	1
1000	63	28000	0L	0

- Task: Whether to sanction Loan or not to a person based on his/her Age, Monthly Income and Existing Loan amount.
- Here the predictor variables are **Age, Monthly Income** and **Existing Loan** and the target variable is **Loan Sanction**.
- There are total 1000 labelled samples given.
- Each sample is of the form: ⟨(Age, Monthly Income, Existing Loan), Loan Sanction⟩

- Like previous, here also we take a fraction of the total dataset as training samples to train our supervised learning model.

TYPES OF SUPERVISED LEARNING

Regression and Classification Problem

- What is the main difference between the examples of *Supervised Learning* we just saw?

- In the first example of *"Blood pressure prediction"*, the predicted or output variable assumes continuous values within a given range.

  This type of supervised learning is known as ***Prediction*** *or* ***Regression*** *problem*

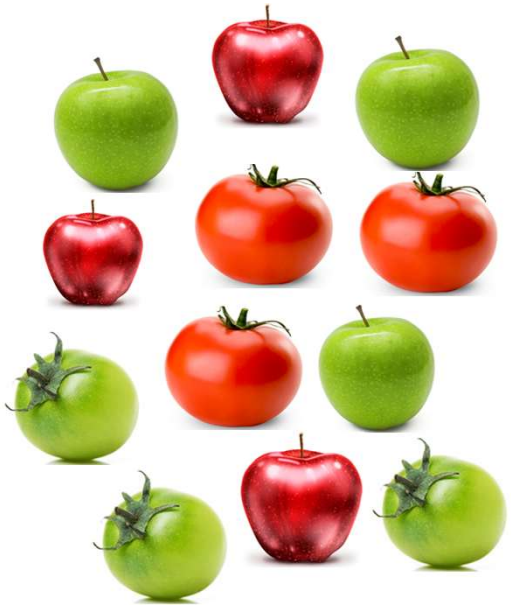
- While in the second example of *"whether to sanction loan or not"*, the predicted or output variable is categorical in nature, as it assumes only finite discrete values.

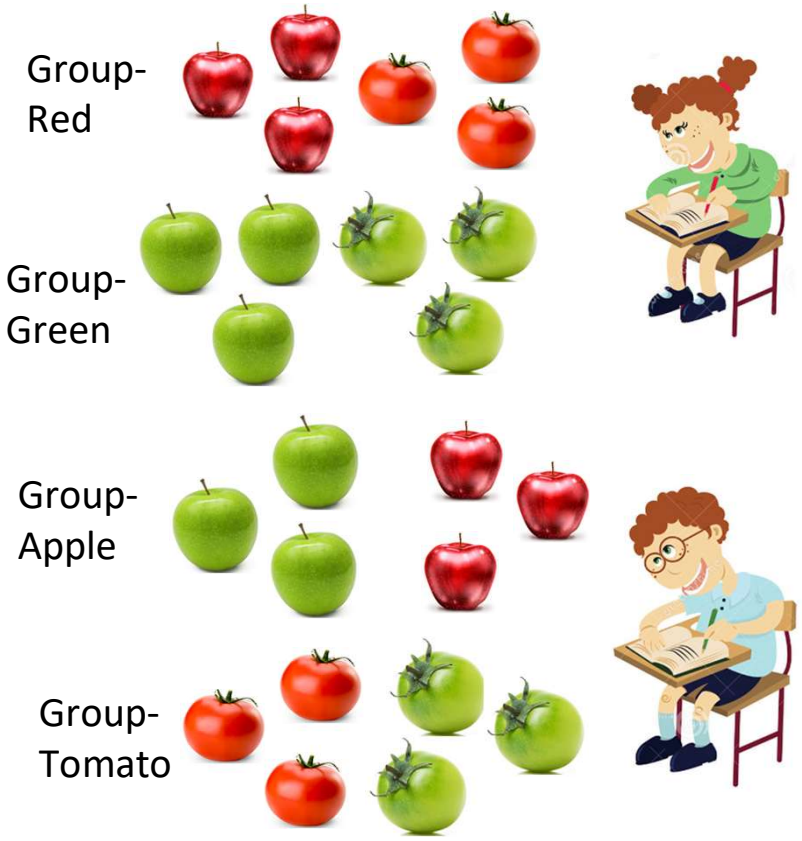
  This type of supervised learning is known as ***Classification*** *problem*

UNSUPERVISED LEARNING

Unsupervised Learning: Learning without teacher (self-organization)

Unsupervised learning is the machine learning task of inferring a function to describe inherent hidden structure from *unlabeled* data.





TYPES OF UNSUPERVISED LEARNING

Clustering: Grouping of objects into two or more groups such that each object fall into exactly one group is called **clustering.**

Example: Customer segmentation

- **Task (T):** Grouping customer data into distinct segments.
- **Experience (E):** A dataset of customer demographics, purchasing history, browse behavior etc. The model has to learn the underlying pattern.
- **Performance (P):** The quality of clusters often measured by internal metrics or by the usefulness of the clusters for business insights.

Other examples of clustering are:

- Organizing a large collection of uncaptioned images into categories.
- Image segmentation
- Document organization and discovery
- Social network analysis

TYPES OF UNSUPERVISED LEARNING

Dimensionality Reduction: Reduce the number of features (variables) in a dataset while retaining as much relevant information as possible. This is useful for visualization, noise reduction, and speeding up other ML algorithms.

Example: Compressing the image / video data

- **Task (T):** Compressing image / video data for efficient storage or transmission.
- **Experience (E):** A dataset of raw image pixel values.
- **Performance (P):** The compression ratio achieved while maintaining an acceptable level of image quality (often measured by metrics like peak signal to noise ratio (PSNR).

Other examples of dimensionality reductions are:

- Visualizing high dimensional biological data.
- Reducing the dimension of input data before applying ML model.

TYPES OF UNSUPERVISED LEARNING

Association rule mining: Discover interesting relationships or dependencies between items in large datasets. Often used in "market basket analysis.

Example:

- **Task (T):** Identifying products that are frequently purchased together in a retail store.
- **Experience (E):** Transactional data (e.g., customer receipts) listing the items bought in each transaction.
- **Performance (P):** The "support" (how frequently items appear together) and "confidence" (the probability that a customer buying one item also buys another) of the discovered association rules. A high confidence rule like "if a customer buys bread, they also buy milk with 70% confidence" is a strong indicator of a useful association.

TYPES OF UNSUPERVISED LEARNING

Anomaly detection (often unsupervised): Identify rare items, events, or observations that deviate significantly from the majority of the data.

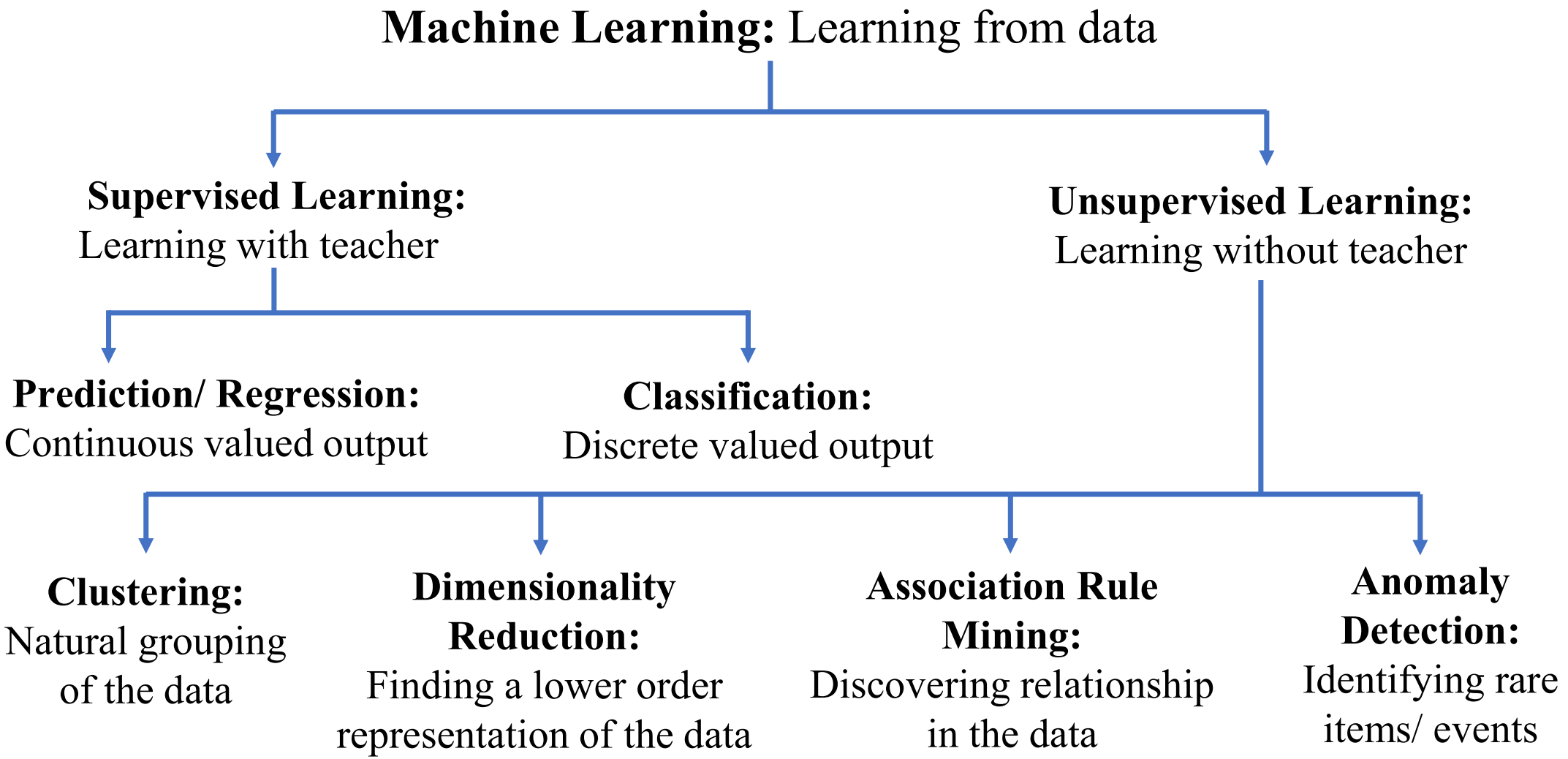
Example:

- **Task (T):** Detecting fraudulent credit card transactions.
- **Experience (E):** A dataset of credit card transactions, most of which are legitimate, but some are fraudulent (unlabeled as such initially). The system learns what "normal" transactions look like.
- **Performance (P):** The number of fraudulent transactions correctly identified (recall) and the number of legitimate transactions incorrectly flagged as fraudulent (false positives), often evaluated against a human-labeled test set after the anomalies are flagged.

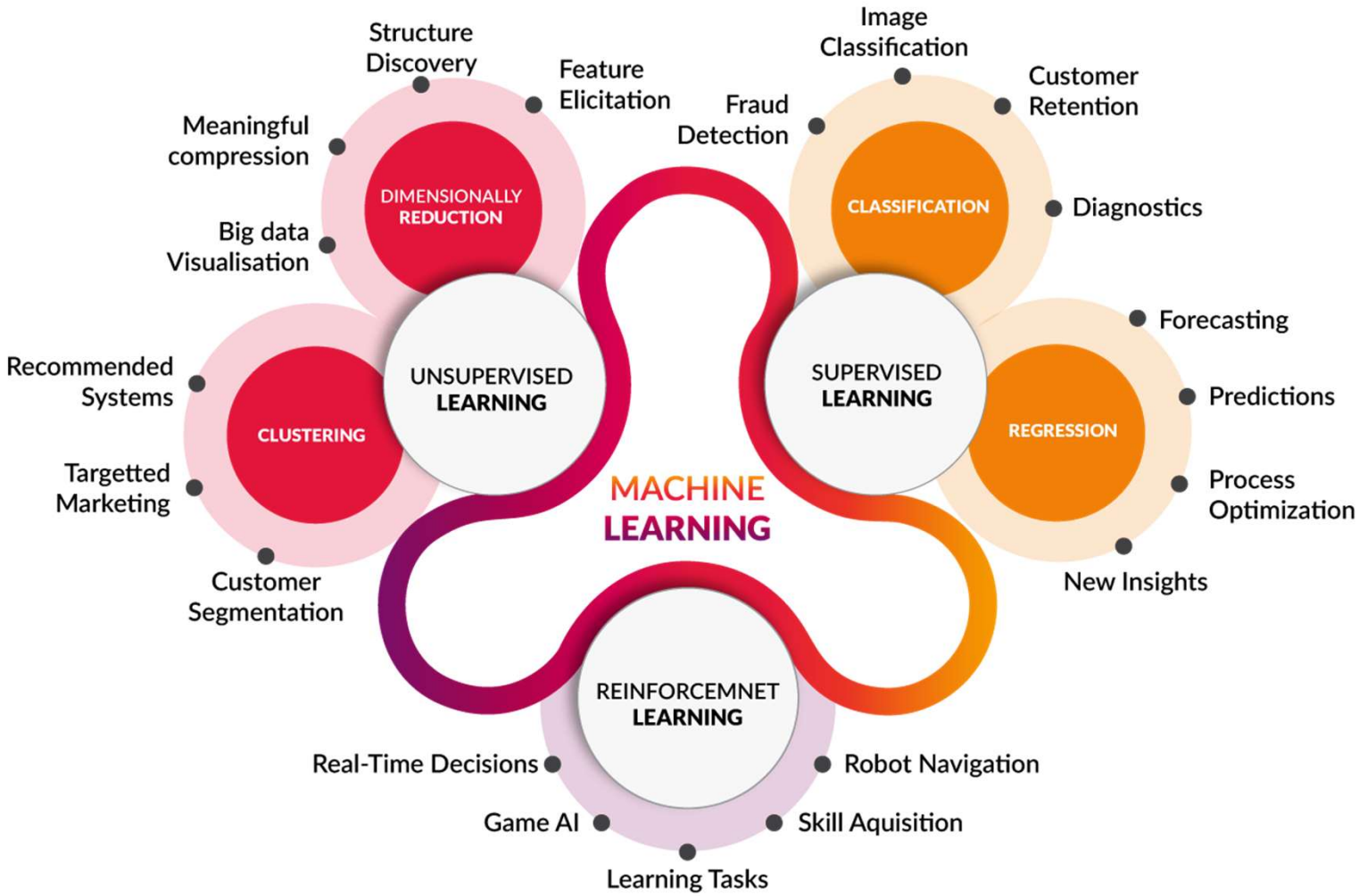
Other examples of anomaly detection are:

- Identifying unusual network activity that could indicate a cyberattack.
- Identifying unusual click pattern in online advertisement that could indicate bots.

TYPES OF MACHINE LEARNING: RECAP



TYPES OF MACHINE LEARNING: RECAP

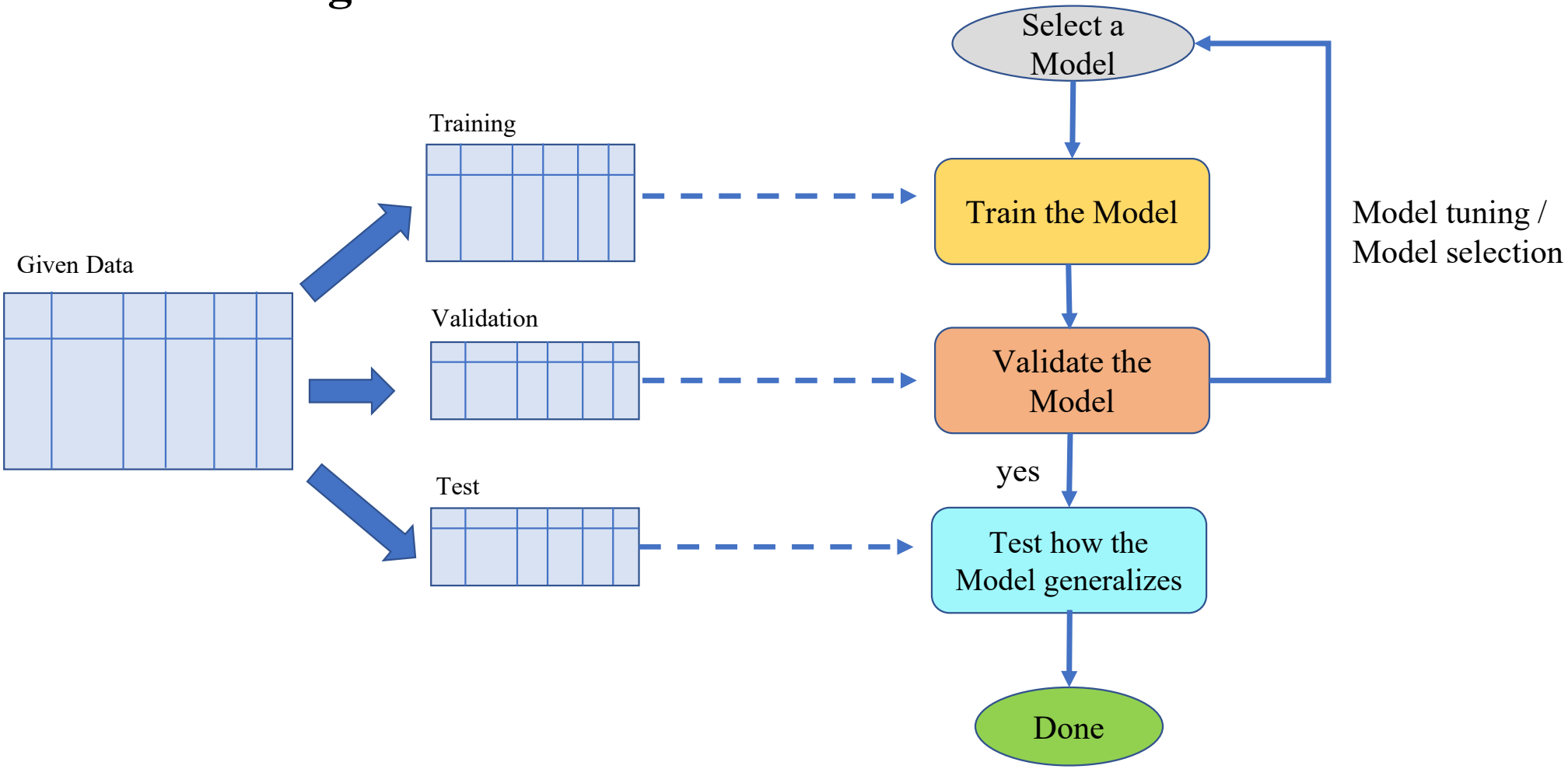


MACHINE LEARNING PROCESS

- Machine Learning is a data oriented discipline. So we need data to start a Machine Learning process. Usually more data or better data leads to better algorithm.
- First identify the task and acquire the required data for the task.
- Then pre-process and clean the data.
- If it's a supervised learning task then, given the pre-processed data first break it into train, validation and test datasets: (Like: 70% training, 20% validation, 10% test). Or we can use techniques like cross-validation.
- Train your model on the training dataset and validate it on the validation dataset.
- Keep changing your model / model parameters until you get desired accuracy on both training and validation sets.
- Finally test your model test dataset (hold out samples) to see how well your model generalizes to unseen dataset.

MACHINE LEARNING PROCESS

Machine Learning Process



CLASSIFICATION

Classification Problem

A **classification problem** is a supervised machine learning task to identify to which of a set of **categories / classes** a new observation belongs, on the basis of a **training set** of data containing observations (or instances) whose category membership or classes are already known.

Classifier

An algorithm trained for classifying a particular set of object into certain number of classes is known as **classifier**. That means a classifier performs classification.

There are different types of classification tasks like: Image Classification, Classification of certain disease to be malignant or benign, classification of a news article to sports / politics / social / finance / weather / entertainment / health etc.

CLASSIFICATION

Doc. Classifier

Social
Political
Sports
Weather
Entertainment

Image Classifier

Monkey
Cat
Dog
frog

Lump Size
Lump Thickness
⋮
Age

Tumour Classifier

Malignant
Benign

Email Classifier

Spam
Non-Spam

- For each type of classification task a separate classifier is needed. A text classifier is unsuitable for image classification and vice versa. Hence, classifiers are task specific.
- If there are only two classes then it is called a **binary classification** problem (and the corresponding classifier is known as **binary classifier**). If there are more than two classes to classify into, then it is called **multi-class classification** problem.

DATA AND FEATURES

Data: Data are nothing but points in the feature space, which denote certain observations / objects.

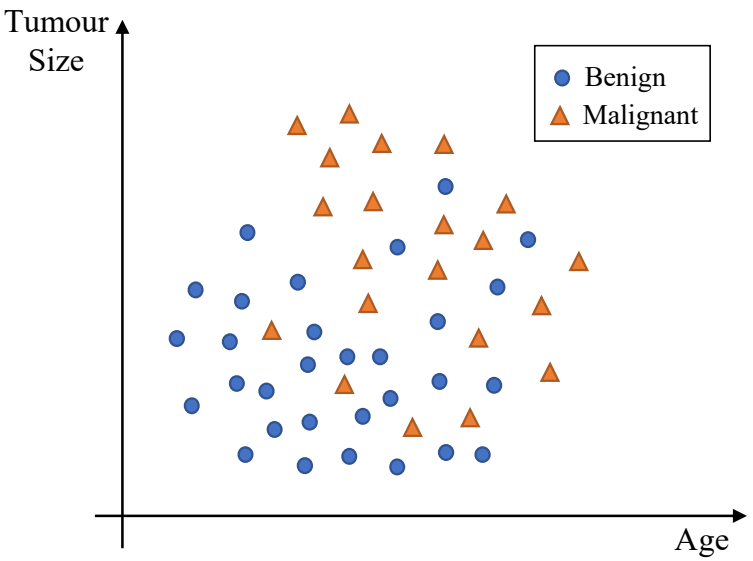
Features / Attributes: Features or Attributes are the properties of the object/pattern which helps us to classify/cluster them.

Examples:

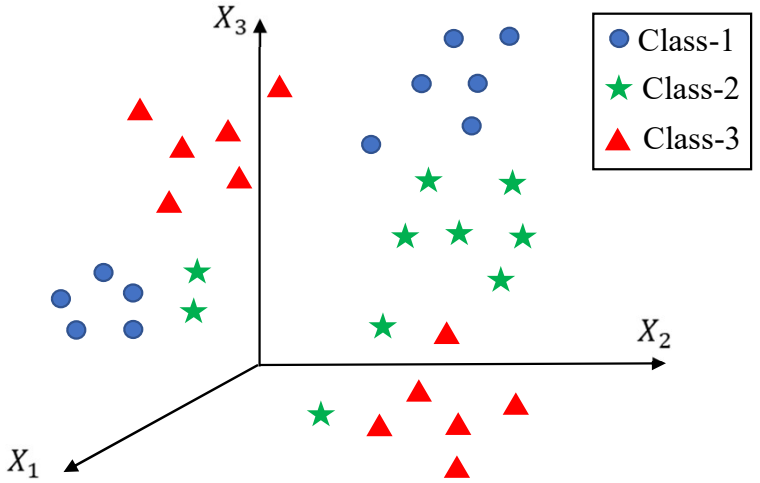
- A student can be represented by a set of features like {Class, Marks, Attendance, Height, Weight, Age, …}. So a particular combination of the values of those features denote a particular student. Similarly some other combinations denote some other students. Collection of several students' features will be called the "Student's dataset".

CLASSIFICATION

Consider the following example of Tumour classification: Here the two features are {Age, Tumour Size}

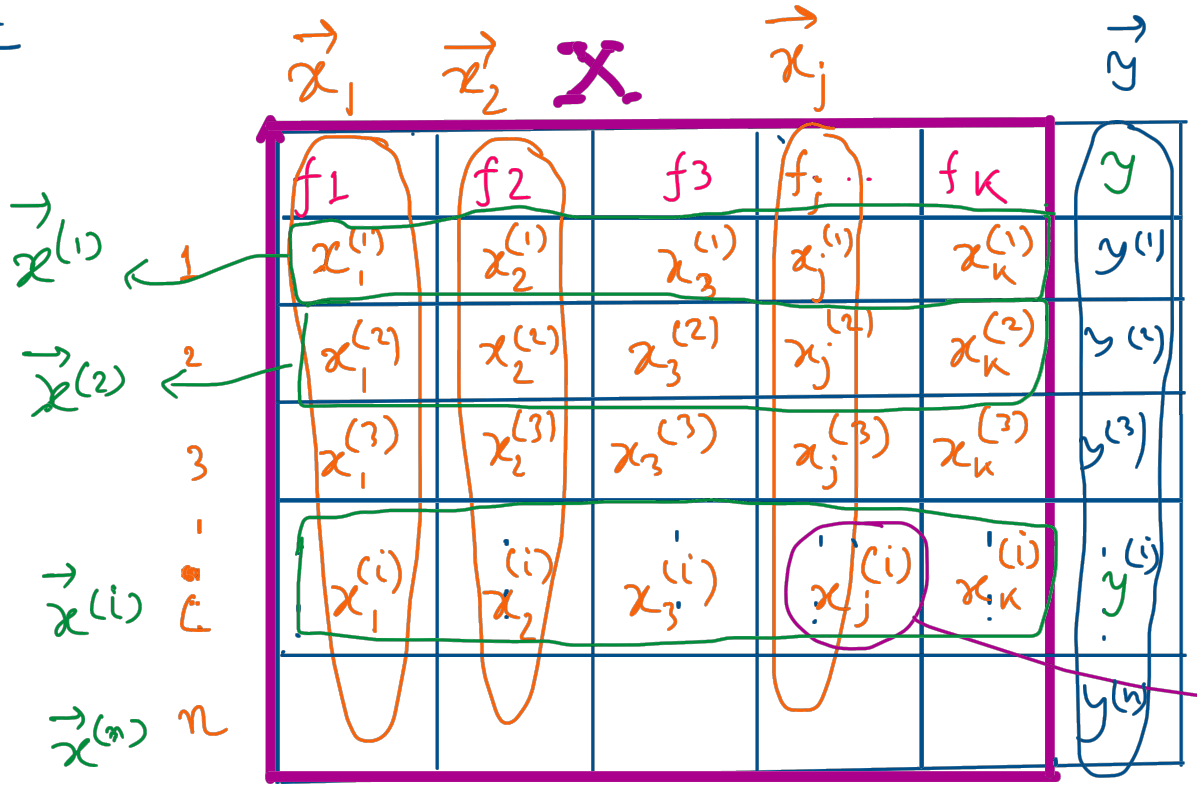


- In general a pattern with k many features/attributes is a vector in k-*dimensional* feature-space. We denote i^{th} observation as $\vec{\boldsymbol{x}}^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots, x_k^{(i)}\right\}$, where $'x_j^{(i)}{}'$ denotes the value of the j^{th} feature/attribute of i^{th} observation.



In classification problem each of the observations in training set are provided with corresponding class label and is usually represented by $\left\{\vec{\boldsymbol{x}}^{(i)}, \boldsymbol{y}^{(i)}\right\}$ where $\boldsymbol{y}^{(i)}$ is the class label of corresponding observation $\vec{\boldsymbol{x}}^{(i)}$.

Data



$X_{n \times K}$ $\vec{y}_{n \times 1}$

attributes / features

observations / records

$$X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_K]$$

$x_j^{(i)}$: The value of j^{th} feature in i^{th} observation

$\langle \vec{x}^{(i)}, y^{(i)} \rangle$: i^{th} observation with label $y^{(i)}$ in the dataset

TRAINING & VALIDATION

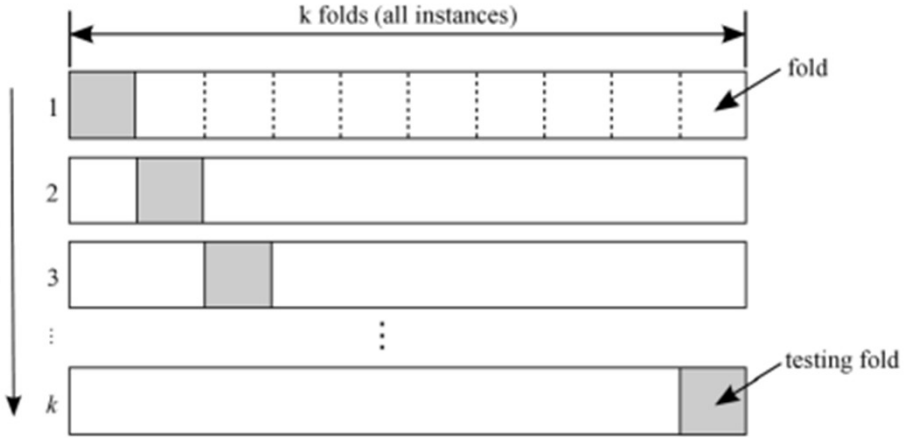
- Usually the given data is subdivided into Training data and Validation data.
  - **Training Data:** Used to train the models and estimate the model parameters.
  - **Validation Data:** Used for evaluation and hyper parameter tuning.
- There are several methods of doing it as discussed following:

(1) Holdout: A fixed percentage of the entire dataset is taken as Training data and rest as Validation data. Usually 70% training and 30% test split is most common. But other splitting ratios are also acceptable.

(2) Bootstrapping: It involves repeatedly sampling with replacement from the original dataset to create multiple bootstrap samples, training the model on each sample, and evaluating its performance on the out-of-bag (OOB) data (data not included in the bootstrap sample).

(3) k-fold Cross Validation:

- Divide into k-parts (folds)
- Train on (k-1) folds, test on the remaining.
- Repeat for different combinations.



(4) Leave-one-out Cross Validation:

- It is a special case of k-fold cross validation
- Here k = No. of samples in the dataset (m)
- All the data except a single observation are used

for training and the model is tested on the single observation.

EVALUATION OF BINARY CLASSIFIER

How to measure the **performance of a classifier**?

Accuracy:

Usually the most common and intuitive measure of classification performance. It is measured as

$$Accuracy = \frac{Number\ of\ Correctly\ classified\ Patterns}{Total\ Number\ of\ Patterns},$$

(Mostly it is represented in perentage)

- We measure both Training and Validation Accuracy of Classifier.
  - **Training Accuracy**: Measure of how well the classifier performs in training set and is evaluated as :

$$Training\ Accuracy = \frac{Number\ of\ Correctly\ classified\ Patterns\ in\ Training\ Set}{Total\ Number\ of\ Patterns\ in\ Training\ Set}$$

  - Validation Accuracy: Measure of how well the classifier performs in validation set and is evaluated as:

$$Validation\ Accuracy = \frac{Number\ of\ Correctly\ classified\ Patterns\ in\ Validation\ Set}{Total\ Number\ of\ Patterns\ in\ Validation\ Set}$$

A **"good"** Classifier should have good Training and Validation accuracy. Why?

EVALUATION OF BINARY CLASSIFIER

Confusion Matrix:

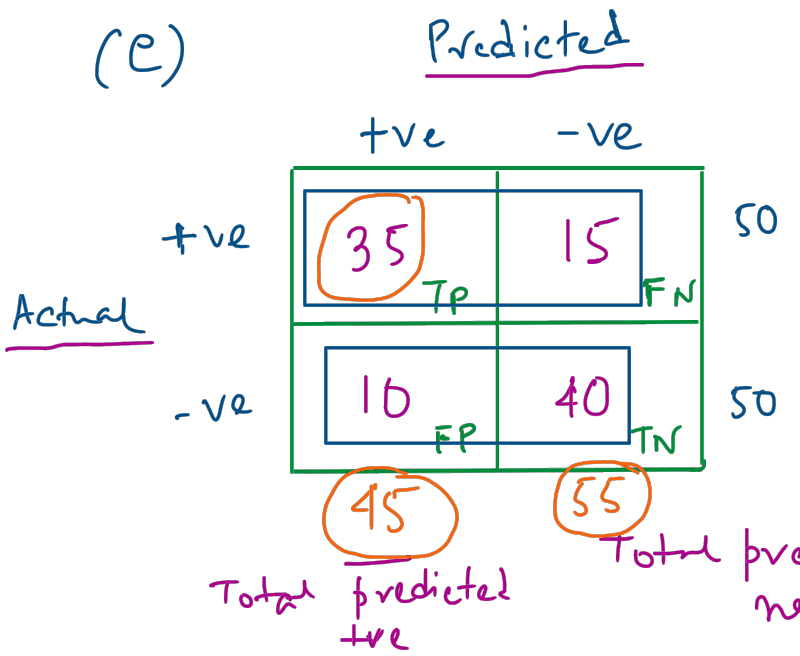
- Consider a binary classification problem where there are only two classes: **Positive (or True)** and **Negative (or False)**.

		Classifier Outcome	
		Positive	**Negative**
Actual Labels	**Positive**	True Positive	False Negative
	Negative	False Positive	True Negative

- Confusion matrix is a nice way to describe classifier's performance.
- **True Positives (TP)** are the cases when classifier gives *positive* output for a positive test sample. **True Negatives (TN)** are the cases when classifier gives *negative* output for actual negative test samples.
- **False Positives (FP)** is when the outcome is incorrectly classified as "positive", when it is in fact "negative".
- **False Negatives (FN)** is when the outcome is incorrectly classified as "negative", when it is in fact "positive".

- True Positives (TP) and True Negatives (TN) are correct classifications. Where as False Positive (FP) and False Negatives (FN) are misclassifications. We want to design a classifier whose **misclassification rate** is as low as possible.

(e)



100 datapoints

Miss-classifications = FP + FN = 25

Total data = TP + TN + FP + FN = 100

misclassification rate = $\frac{\#\text{ misclassification}}{\#\text{ data}}$

$= \frac{FP + FN}{TP + TN + FP + FN}$ (25%, 0.25)

Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$ = 75 %. 0.75

Accuracy = 1 − misclassification rate

Precision = How many predicted true samples are actually true $= \frac{35}{45}$

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

Recall = How many actual true events has been predicted as true $= \frac{35}{50}$

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

EVALUATION OF BINARY CLASSIFIER

$$Misclassification\ Rate/Misclassification\ Error = \frac{FP + FN}{TP + TN + FP + FN}$$

Other Performance Measures obtained from Confusion Matrix:

		Classifier Outcome	
		Positive	**Negative**
Actual Labels	**Positive**	True Positive	False Negative
	Negative	False Positive	True Negative

- **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$ = 1 – Misclassification Error
- **Precision (P) :** How many predicted true events are actually true

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

- **Recall (R) :** How many actual true events are correctly classified

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

- **F-Measure (F) :** Harmonic mean of Precision and Recall

$$F = \frac{2PR}{P+R}$$

Harmonic mean :- (HM)

a, b

(AM) arithmatic mean = $\frac{a+b}{2}$

(GM) geometric mean = $\sqrt{ab}$

$$HM = \left(\frac{\frac{1}{a}+\frac{1}{b}}{2}\right)^{-1} = \frac{2}{\frac{1}{a}+\frac{1}{b}}$$

$$= \frac{2}{\frac{a+b}{ab}} = \frac{2ab}{a+b}$$

$a = 2, b = 8$

$AM = 5$

$GM = 4$

$HM = 3.2$

$AM \geq GM \geq HM$

$$F_1 = \frac{2PR}{P+R} = \frac{2}{\frac{1}{P}+\frac{1}{R}}$$

S_1: $P = 0.9$, $R = 0.2$

$$F_1 = \frac{2 \times 0.9 \times 0.2}{1.1} = 0.33$$

S_2 $P = 0.3$ $R = 0.8$, $F_1 = \frac{2 \times 0.3 \times 0.8}{1.1} = 0.436$

S_3 $P = 0.7$ $R = 0.7$, $F_1 = \frac{2 \times 0.7^2}{2 \times 0.7} = 0.7$.

$$F_\beta = \frac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \qquad \beta > 0$$

F-beta score

if $\beta = 1$ then $F_\beta = F_1 = \frac{2PR}{P+R}$

if β is higher ($\beta > 1$) then more weightage to recall

$\beta < 1$ then more weightage to precision

$$F_2 = \frac{5PR}{4P+R} \qquad F_{0.5} = \frac{5PR}{P+4R}$$

Thank You