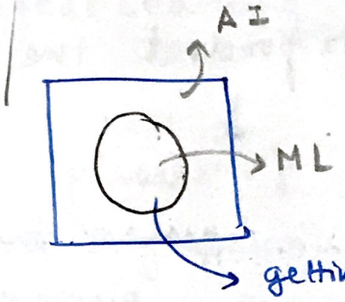
* Introduction to Machine Learning

↓

methodology in which we want machines to learn from past experiences / data so that they can handle the cases coming in future.

↓

more like learning models or patterns, and based on that predicting outputs to future tests / inputs.



* AI vs ML

↓

making machines smarter so that they can take smart decisions.

↓

decision taking can majorly happen in two ways:

a) specifically being programmed to handle data

b) learning from past data (i/o) and developing apt. models to predict or decide the output if the input is hit or encountered again.

* Applications:

(i) recommendation systems.
(ii) specific feed or posts.
(iii) advertisements.
(iv) opinion mining
(v) data analytics
(vi) self driving cars.
(vii) oil reservoir prediction
(viii) seismic data use.

* Data mining: finding relevant info. from a huge pile of data

↓

can be solved using i) ML
ii) other data analytics techniques

(50) * Types of Machine Learning

(i) Supervised learning
(ii) Unsupervised learning
(iii) Reinforcement learning.

supervised:
(i) specific datasets with inputs (past data) and their outputs, and for future inputs, we want to predict the output.
↓
a) Housing price problem
b) tumor types

↓
Y: clearly marked for past data

(ii) Unsupervised : no clear markings of o/p.
↓
example. putting similar people in the same group on facebook for content recommendation
↓
i.e huge amounts of data, and we want to cluster them into meaningful parts
↓
finding patterns in data.

(iii) Reinforcement : way humans learn, based on incentives
↓
receive positive and negative feedback on the basis of actions taken for a given situations.
↓
for example computer chess game. If a set of moves results in a win, then,

the moves are correlated to the feedback, otherwise with negative.

↓

As the m/c gains more experience, it gathers an understanding of what moves are logical and what moves are not.

*) Data storage was expensive in the past → not now

↓

every digital entry is / can be stored.

↓

This has led to the progress in machine learning

** Supervised Learning → past exp. labelled, from i/p to o/p.

↓

(clear demarcation)

a) Regression → o/p is in a cont range

↑) House price specific prediction

b) classification

↓

placing elements into groups / classes

based on attributes. o/p is one of few classes.

↓

(a) cancer tumor type

(b) news classified

(c) image type / Monkey

(digit recognition)

House price problem → specific / exact amount

↓

(regression)

↓

range

a) < 100K

b) 1M > → > 100K

c) > 1M

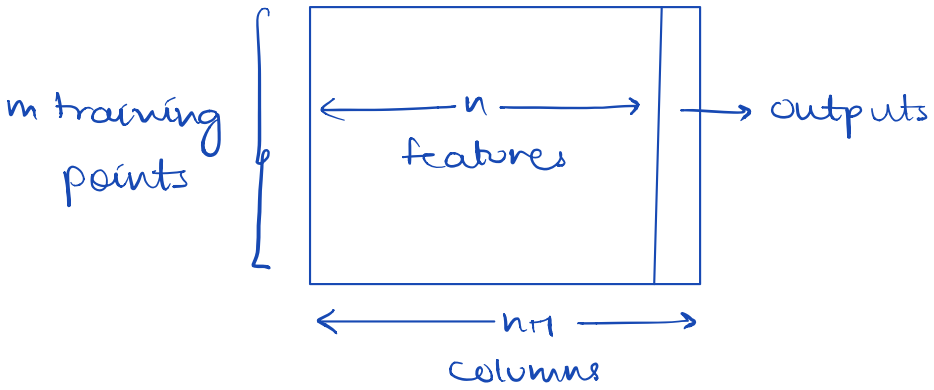
} classes

↓

classification

*) Steps for supervised learning : we are provided training data, the distribution of which we try to model.

↓



features provide description of factors which may impact the o/p in some manner

↓

training data has m data points with n features and 1 output column.

↓

Dimension : $m * (n+1)$

↓

1) Finding data : finding relevant data for the problem is the first step. for example recommender systems for restaurants.

Data on what kind of restaurants, what kind of food people like.

↓

Multiple sources may be available for example company database, forms, etc.

↓

Depending on the problem, we need to figure out where to arrange data from.

2) Data loading and cleaning

↓

one source, no issues

Multiple sources, then issues arise

↓

for example inconsistancy in the features present in dataset from two different resources

↓

consistency needs to be achieved by either taking common features or finding a way to fill the values for the set that doesn't have the feature.

↓

may have to handle strings, NANs. → case by case decision

↓

may require adding / removing columns as well.

a) for example, used car database may contain buy and sell columns. These can be used to derive age column, which may turn out to be a useful feature for the problem.

b) for example, Titanic Dataset, dealing with predicting survival of people, name is not a very useful feature and can be therefore dropped during analysis

↓

whatever helps / aids the analysis of the data should be done.

→ Pick the algorithm: Intermediate step of picking the algorithm which would be used to get the model for the underlying data.

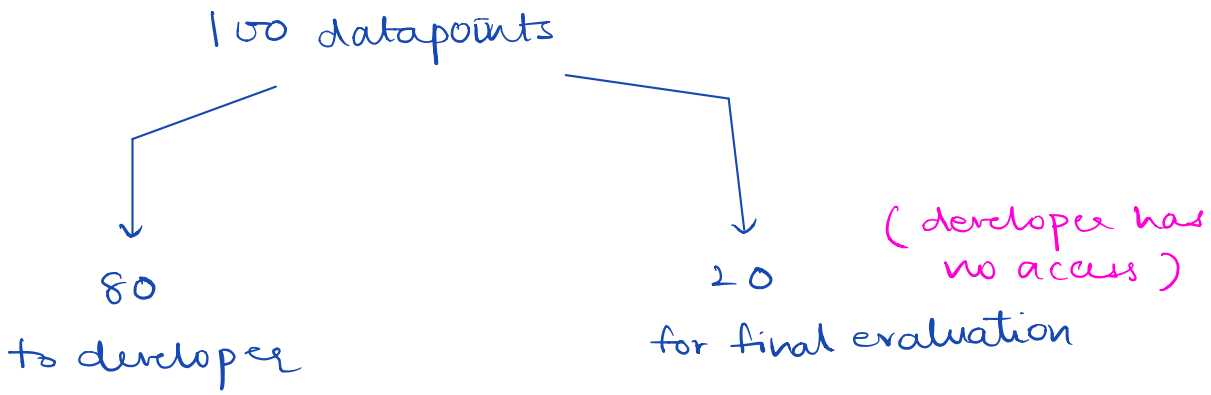
3) Training data used to learn a model

4) Testing data used to test the efficiency of the model based on different metrics. like accuracy.

↓

Can be used to do comparisons between different algorithms' performance.

*) Walk through



Option1: we could use the entire data for training and use the same data to test the model. This model may perform extraordinarily well, but if a data point that is not present in the dataset is given, then the model might falter big time. The probability of encountering new data is high in the real world, and hence the training process should be catered around that.

Option 2 : Dividing our training data (80 datapoints) into two parts called train and test. 7 : 3 is a common ratio used for this split. Now the model is trained on 56 points and then tested on our end on the remaining 24 data points.

↓

The performance of different algorithms on test data can be used for picking the algorithm.

*) The phenomenon of giving good performance on the training data, but giving bad performance on test data is called overfitting and should be avoided by all means.

↓

Here the model is trying to remember the data.

*) We'll be given X_{cv} for which we generate a labels file that will be tested for the correct labels.