So, there are different machine learning paradigms that we will talk about.

1) "Supervised Learning"
   - ↳ Learn an input & output map
   - Eg:- [Classification : categorical output.
     Regression : continuous output.]

2) "Unsupervised Learning"
   - ↳ Discover patterns in the data
   - Eg:- [Clustering : cohesive grouping
     Association : frequent cooccurence]

3) "Reinforcement Learning"
   - ↳ [Learning control]

So the first one is supervised learning where we learn an input to output map. So, we are given some kind of an input, & it could be a description of the patients who come to the clinic & output that has to be produced is whether the patient has a certain disease or not, so they had to learn this kind of an input to output map.. or the input could be some kind of question & then the output would be the answers to the question, or it could be a true or false question. I give you a description of the question & you have to give me true or false as the output.

In supervised learning what we essentially do is, learn a mapping from the input to the required output. if the output you are looking for happens to be categorical output like whether he has

P.T.O

a disease or does not have a disease or whether the answer is true or false then the supervised learning problem is called the [classification problem.]

& If the output happens to be a continuous value like, how long will the product last before it fails or what is the expected rain fall tomorrow, so these kind of problems would be called as regression problems.

These are supervised learning problems, where the output is a continuous value & these are called regression problems.

We will look into more details in classification & regression as we go on.

Second class of problems are known as Unsupervised learning problems; where the goal is really not to produce an output in response to an input, but given a set of data, we have to discover patterns in the data, right.

That is more of an unsupervised learning, there is no desired output we are looking for, we are more interested in finding patterns in the data.

So clustering, is one of the unsupervised learning tasks, where we are interested in finding the cohesive groups among the input patterns. for eg. I may be looking at customers, who come to my shop & I want to figure out if there are categories of customers like may be college students could be one category, IT professionals could be another category & so on. so forth. So when we are looking at that kind of grouping. in my data, we would call that a clustering task.

So another popular unsupervised learning paradigm is known as the "Association rule" mining or frequent pattern mining, where we are interested in finding a frequent co-occurrence of items in the data that is given to us.
So whenever "A" comes to my shop, "B" also comes to my shop. So there are the pairs of co-occurrence so I can always say that okay if I see A then there is very likely that B is also in my shop somewhere. So I can learn these kind of associations b/w data.
We will look into it in more details later.
There are many different variants on supervised & unsupervised learning, but these are the main ones that we look at.

The third form of learning which is called reinforcement learning it is neither "supervised" nor "unsupervised."
& typically these are problems where we are learning to control the behaviour of the system.

Like I said for every task, we need to have some kind of performance measure.

"Task"	"Measure"
1) Classification	error
2) Regression	error
3) Clustering	scatter / purity
4) Associations	support / confidence
5) Reinforcement learning	cost / reward

So, if we are looking at the classification task, the performance measure is going to be the classification error.

So we will talk about many many different performance measures in the duration of this course, But the typical performance measure, you would want to use is the classification error, it's like how many of the items or how many of the patients did I get incorrect. so, how many of them who are not having the disease are predicted by the model to have the disease, & how many of them, that had the disease that I missed. So, that would be one of the measures that could be used. & that would be the measure that we want to use but we will see later that often that is not possible to actually learn directly with respect to the measure.

Likewise for regression also, we have the prediction error as a measure, suppose I say it is going to rain like 23 mm & then it ends up raining .49 cm, so that is a huge prediction error &

In terms of clustering, this becomes little more tricky. to define the performance measures. We'll donot know what is a good clustering algorithm because we donot know how to measure the quality of clusters.

So people come up with all different kind of measures & to one of the more popular ones is a scatter or spread of the cluster that essentially tells you how spread out the points are that belong to a single group. If you remember we are supposed to find cohesive groups. So if the group is not that cohesive it's not all of them are not together then you would say the clustering is of a poorer quality & if you have other ways of measuring things like I was telling you, like, if you know that people are college students

then you can figure out that how many or what fraction of your clusters are college students.

So you can do these kinds of external evaluations, so one means that people use popularly there is purity. & in the Association rule mining, we use variety of measures called 'support & confidence' that takes a little bit of work to explain support & confidence. So, I will defer it.

Now in the Reinforcement learning topic, if we remember, I told you it is learning to control, so you are going to have a cost for controlling the system & also the measure here is cost & you would like to minimize the cost that you are going to accrue while controlling the system.

"So these are the basic machine learning topics"

There are several challenges, when you are trying to build a machine learning model

Challenges :-

1) How good is a model?
2) How do I choose a model?
3) Do I have enough data,
4) Is the data of sufficient quality.
   - Error in data eg: Age = 885; noise in low resolution images, 'missing values'
5) How confident can I be of the results.
6) Am I describing the data correctly?
   - Are Age & income enough? Should I look at gender also?
   - How should I represent age? As a number, or as young, middle age, old?

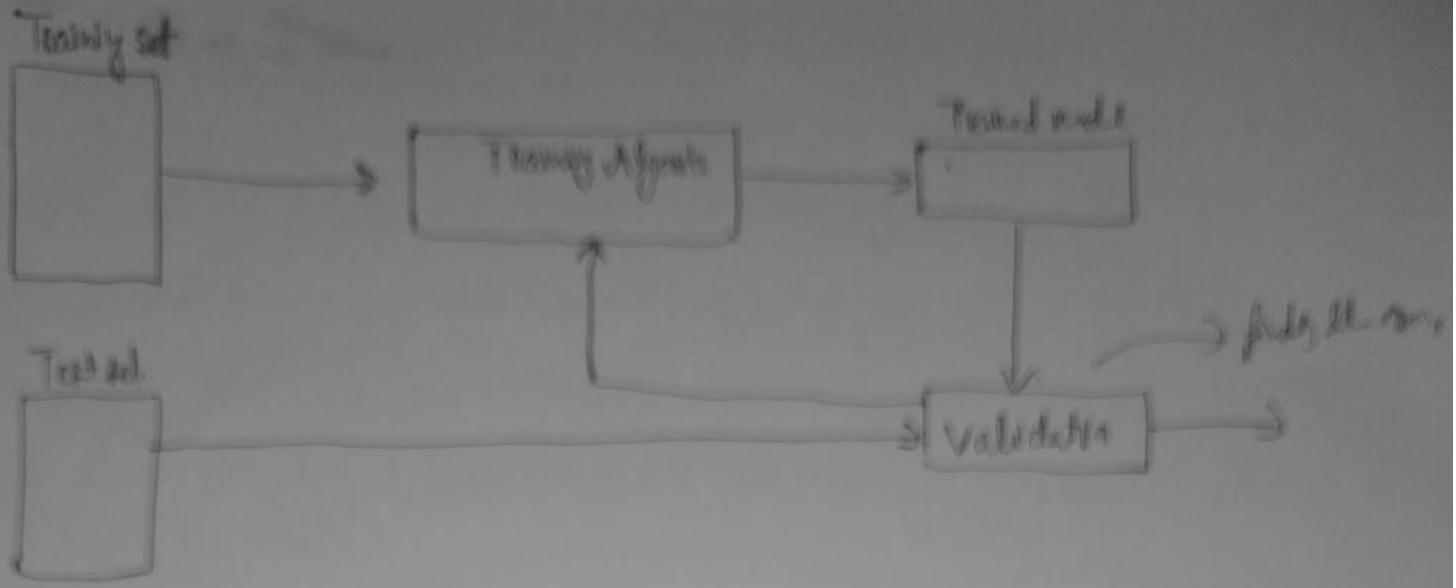
So, the first thing that we have to think about, how good is a model that you have learned, so I talked about a few measures, but often those are not sufficient there are other practical considerations that come into play & we will look at some of those towards the middle or in upcoming lectures.

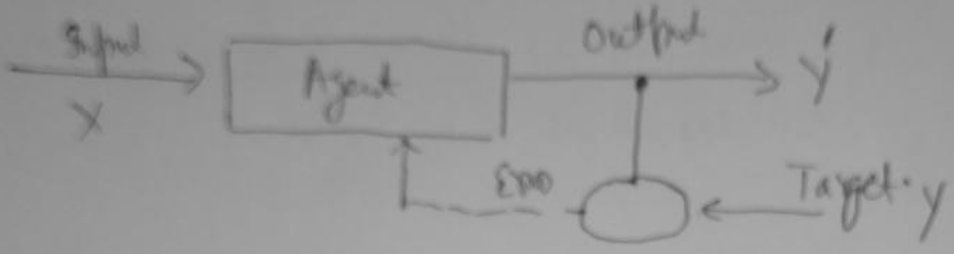
Bulk of the time would be spent on answering, how do I choose a model; so, given some kind of data, which will be the experience that we are talking about, to given this experience, how would I choose a model, that somehow learns what I want to do, do how that improves itself with experience & so, on. do how do I actually find the parameters of the model that gives us the right answer.

So, this is what we will spend much of our time in this course, & there are whole bunch of other things that you really have to answer to be able to build a useful 'data analytics or data mining solutions', questions like do I have enough data or do I have enough experience to say that my model is good.

If the data of sufficient quality, there could be errors in data, suppose I have medical data & "a" is recorded as 225 now what does that mean, it could be 225 days in which case it is a reasonable number, if could be 22.5 years, again is a reasonable no. or 22.5 months is reasonable but if it is 225 years, it's not a reasonable no. which means there is something wrong in the data. So how do you handle things like these, or how do images or missing values...

Since it is a ML course, & it is primarily concerned about the algorithms of machine learning, & the maths & the intuition behind them, & not necessarily about the questions of building a practical system based on this.





ML Pipeline :->

1> Gathering Dataset

2> Preprocessing Data (Standardization & Normalization)

3> Divide the dataset into
   ↳ Training data
   &
   ↳ Testing data

4> Using Training data to train the Model

5> Using Testing data to validate the Results