INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: TY BSc

Subject : Machine Learning & Predictive Analytics

Subject Code:

Chapter: Unit 2 Chp 1

Chapter Name: Introduction to Machine Learning - 1

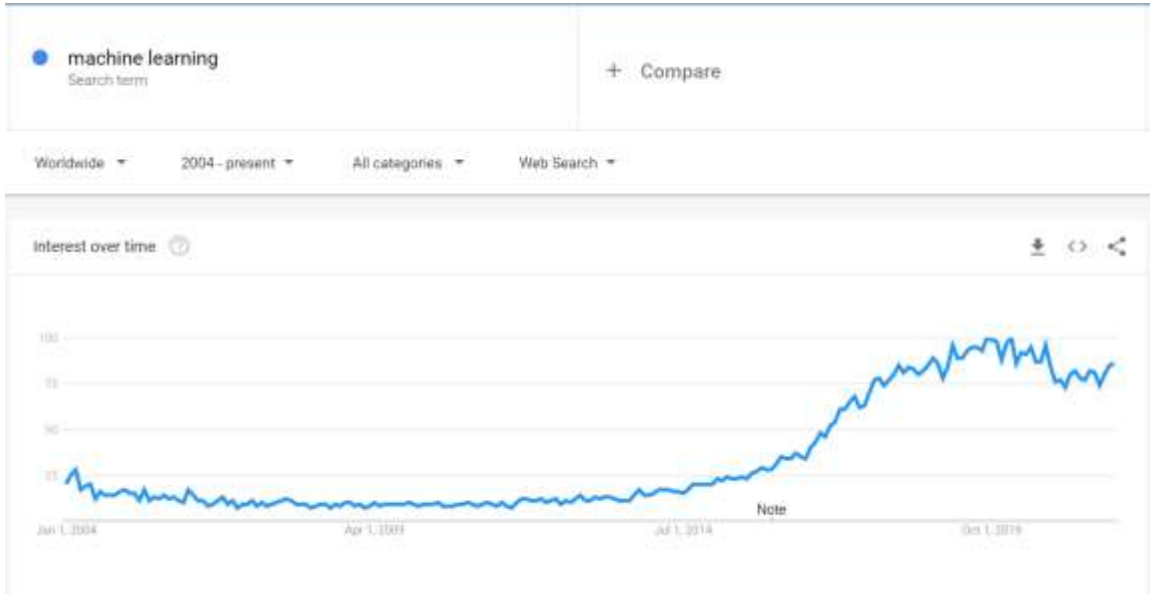
1.1 Machine Learning

By know, you all know why we humans learn? Right?
-We learn in order to ***gain knowledge*** about things unknown to us, in order to use that knowledge towards some activity that in return ***benefits*** us.

Then why would a Machine Learn?

- Simple, we humans can only learn and remember so much that we need machines to *do our work* for us. Machines learn because we command them to learn.

So, the natural question is ***what exactly do we mean by machine learning?***
Let us find out….

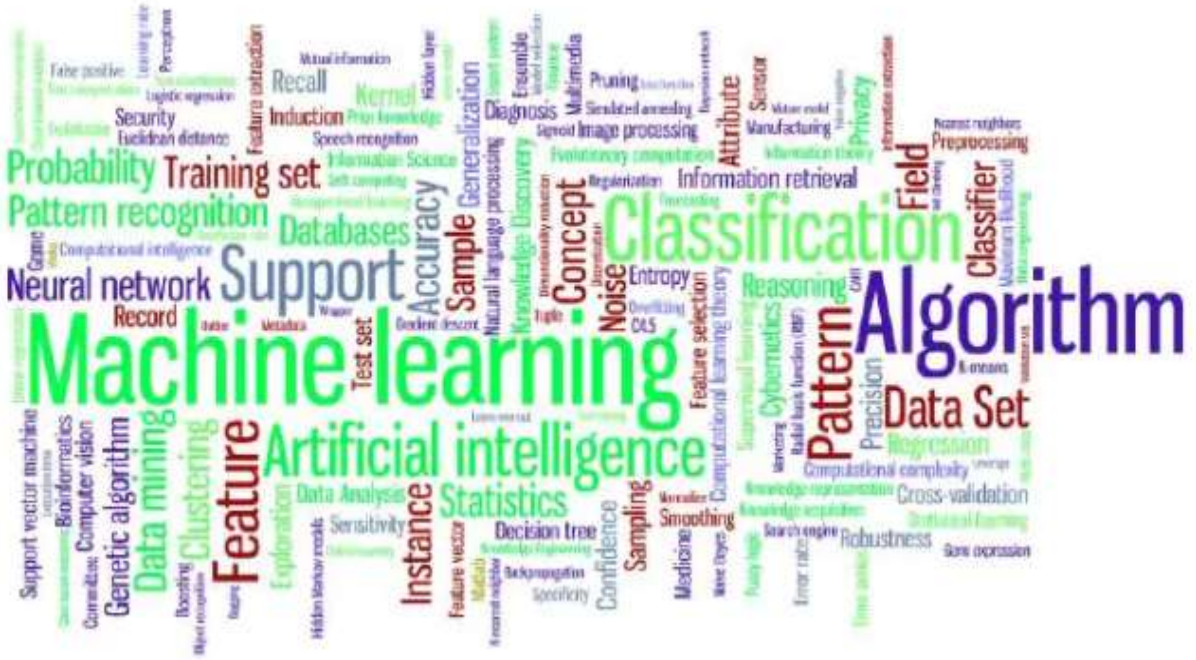


Interest in Machine Learning – Graph of Google Trends

1.2 What is Machine Learning?

- Learning at it's core is a ***process*** in which system improves with experience.
- ML concerns itself with programs that automatically ***improve*** their performance through experience

  – Herbert Simon
- In simpler words, a computer algorithm developed & applied to data in order to generate information in order to find the ***hidden patterns*** and solve a specific problem is Machine/Deep Learning.



1.2 What can Machine Learning do?

- As mentioned earlier , ML can be used to analyze existing data , generate new data , and create software that interacts with humans and if it learns well can ***mimic human intelligence***.
- Commonly solved problems using Machine Learning

  - Consumer segmentation and targeting based on buying patterns

    Target, a US retailer was successful in building a model that would estimate if a customer was pregnant and their due dates. Using this model, Target sent coupons for relevant products to its customer (*Read the Forbes link in the notes)*

  - Forecasting Election Results

    Every year, new broadcasters forecast winners & losers of the elections even before the official vote counting begins. They use both traditional exit poll methods with new age algorithms to make their prediction as accurate as possible

  - Image Recognition

    If you show a computer enough examples of faces, it will be able to recognize faces on its own. Meta uses this when it asks you to tag the people in a photo.

1.2 What can Machine Learning do?

- Credit Rating and Fraud Detection

CRISIL & other such credit rating agencies have implemented ML techniques in order to analyze credit card ***spend patterns*** and assign a Credit Rating to consumers. Banks & other Financial Institutions have further made propriety software that help them ***detect defaults*** on a loan or a fraudulent transaction.

- Risk Classification in Motor Insurance

Insurance Companies have started using ***in-car monitoring devices*** to collect data from the actual user in order to classify the risk more accurately and provide different rates of ***premium*** to different customers.

- False Tax Declarations

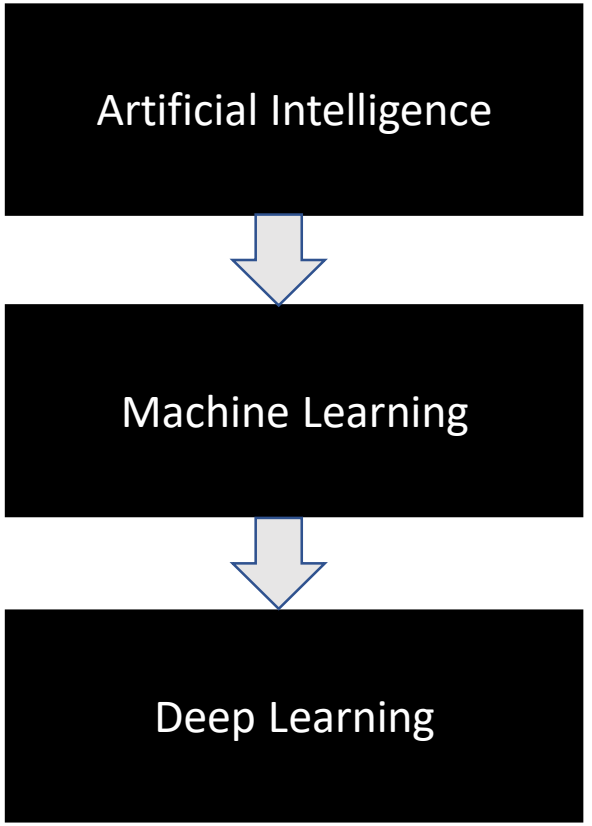
Various government departments involved in collection of taxes from businesses & citizens have implemented various techniques that can identify ***tax theft*** and allow human intervention & audit. In future, they plan to implement such techniques further to target scrutiny towards tax evasions.

1.3 AI, ML & Deep Learning

Artificial Intelligence is a branch of computer science dealing with the simulation of intelligent behavior in computers. *Weak AI* can only fulfil a particular function whereas *Strong AI* comes with all abilities of a human being (*Like Data from Star Trek*).

Machine Learning is a branch of AI but we are ever able to create *Strong AI*, it will be using ML techniques.

Deep Learning is a branch of ML Inspired by our brain. Our brain works using a network of ***neurons*** where a neuron firing triggers the firing of neurons its connected to starting a chain reaction. In deep learning, neurons are replaced by a *node*.



1.4 Various Machine Learning Categories

- So, we have learnt that ML can be used to solve *difficult for humans'* type of problems, but problems can come in different *shapes & sizes (types).* As a result, we have divided various ML algorithms into 4 major categories:

  - Supervised Learning
  - Unsupervised Learning
  - Semi Supervised Learning
  - Reinforcement Learning

- Examples: Customer Segmentation is a business problem solved using Unsupervised learning whereas fraud detection is a problem that can be solved using Supervised Learning

1.4 Supervised Learning

- Let's say you work at a BNPL (*Buy Now Pay Later*) start-up and our tasked with marking accounts that are likely to default in the next billing cycle.
- The Accounts Department has sent across a dataset regarding all accounts active for the last 12 months. You Boss is expecting you to make your ***predictions*** known to him by the end of the week. *What do you do?*

- **You use Supervised Machine Learning Algorithms to your advantage.**
- Supervised ML is associated with ***prediction*** models where the ***output is specified***, and the machine is given a ***SPECIFIC AIM***. Model will try to set parameters in order to achieve *best* prediction.
- Often times, we will set the *LOSS* ***FUNCTION*** along with the target & the explanatory variables.

1.4 Supervised Learning – An Example

- Sam, an avid cricket fan, is recruited by the BCCI as a Data Analyst in the cricket team, to analyze ***player performances*** and help in team selection for the upcoming South Africa test series.
- BCCI has provided you with the ***data*** regarding all international matches and domestic matches the players participated in. The data includes runs scored, balls faced, strike rates, wickets fetched and other appropriate metrics.

The Question is how should we approach a problem like this?

1.4 Supervised Learning – An Example

- *Step 1*: Now , we need to *evaluate* player performances, so we need to create a ***Evaluation Score*****,** which is **function** of all different fields available and required to evaluate the performance of a player. It could be a function of *runs scored, strike rate, number of wickets taken by the player, economy rate, catches taken* etc.

  - In order to finalize this ***FUNCTION***, we will need ***expert*** help from long time coaches & managers.

  - The Team Doctor might also want the current health status of the players to be included in the ***Evaluation Score***.

  - If the ***Evaluation Score ,*** is set-up incorrectly, it would lead to the entire model created rather useless for any purpose. So appropriate ***care*** should be taken in order to set this up.

- Once this ***FUNCTION is set-up***, we can move to the next step, fetching the appropriate data.

1.4 Supervised Learning – An Example

- *Step 2:* As an analyst, the IT Department will usually provide you the entire data set and you are required to pick the appropriate data and ***clean*** it for your particular use.
- We will again use some common knowledge and fetch the appropriate data from given data.

- As we are working on the *South African **TEST** Tour,* it is safe for us to ***ignore*** the data of T20s as it wouldn't be relevant for our purpose.

- While we ***can't ignore*** the ODI data, we can plan to give it ***less weight*** than the Test Data. Similarly, we should be giving more weight to matches played in *South Africa* more than matches played in the *Subcontinent*.

What other data changes can you think of we should be making?

1.4 Supervised Learning – An Example

- Once we have cleaned & set-up the data appropriately, we can *calculate* the ***EVALUATION SCORE*** for a player for each match and set-up a ***regression*** model to ***predict*** the player's evaluation score based on the pitch condition, weather & the opponent.
- We can then use the expected pitch condition, weather & opponent as *South Africa* to estimate the player's evaluation score and ***DECIDE*** on their selection on the tour and the specific matches if needed.

1.4 Supervised Learning - Sub-Types

- **REGRESSION**

  - Regression Algorithms are used if there is a *relationship* between the input variable and the output variable.

  - It is used for the prediction of *continuous* variables, such as Weather forecasting, market trends etc.

- **CLASSIFICATION**

  - Similar to regression algorithms, there exists a relationship between the input variables and the output variable. Here, the relationship is non-linear in nature.

  - It is used for the prediction of *categorical variables* with 2 classes such as Male / Female, True / False etc.

1.4 Supervised Learning – Advantages

- With the help of supervised learning, the model can ***predict*** the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the ***classes*** of objects.
- Supervised learning model helps us to solve various ***real-world*** problems such as fraud detection, spam filtering etc.

1.4 Supervised Learning – Disadvantages

- Supervised learning models are suitable for ***SIMPLE*** tasks but fail to work for handling complex tasks.
- Supervised learning *CANNOT* predict the correct output if the test data is different from the training dataset.
- Training requires lots of computation ***TIME***. The time spent increases fast when the size & complexity of the data increases.
- Supervised learning can only be performed when we have *ENOUGH KNOWLEDGE* about the classes of the object

1.4 Supervised Learning - Summary

- *In short, Supervised Machine Learning is :*

  - Uses historic data , including variables $(x_1, x_2, x_3 \ldots x_n)$ *and the target (*y *or* $f(x_1, x_2 .. x_n)$)

  - The parameters are calculated for all variables be to used in prediction.

  - The target is clearly ***KNOWN*** *and the model is optimized to estimate the target and any change in the target would mean a change in the model.*

The popular Supervised Machine Learning techniques –

- *Linear Regression*
- *Polynomial Regression*
- *Bayesian Linear Regression*
- *Logistic Regression*
- *Decision Trees*
- *Support Vector Machines*
- *Random Forests*

1.5 Unsupervised Learning

- Previously we learnt *supervised* ML in which models are trained using ***labelled data*** under the supervision of training data.
- *What if we DON'T have labeled data & need to find the hidden patterns from the given dataset?*

  Well, to solve such problems, we use **UNSUPERVISED** ML learning algorithms.

- It is defined as *"Unsupervised learning is a type of ML in which models are trained using UNLABELED dataset and are allowed to act on that data WITHOUT any supervision."*

- Unsupervised learning ***cannot*** be directly applied to a regression or classification problem because there is ***no output variable*** for the given input variables.
- The ***GOAL*** here is to find the underlying structure of dataset, group that data according to similarities and represent the dataset in a compressed format.

1.6 Unsupervised Learning - An Example

- A very simple use case for *unsupervised learning* is IMAGE RECOGNITION.

  Open *Google Photos* on your mobile phone
  You will be able to see that Google has segregated your photos into different categories.
  All your photos containing you will be grouped together. You will also find images grouped by places.

- *So how do you think Google did this?*

  Engineers at Google built a program that can look at a photo can recognise its ***unique features***.
  Next time an image with similar unique features comes up, Google will recognise them as ***similar***.
  This process of grouping similar things together is called ***CLUSTERING***.

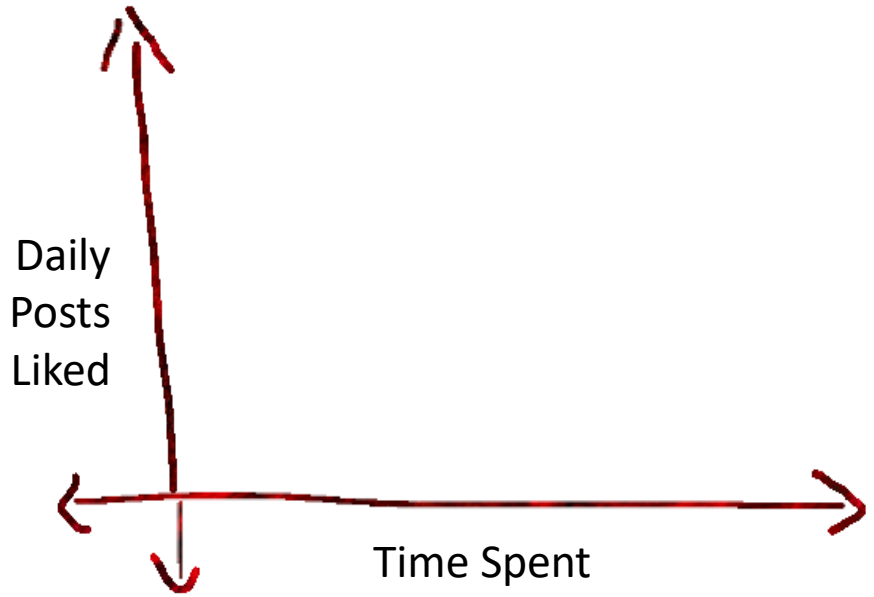
- Lets consider another scenario … *Meta*, a social media company, wants to target its users with more relevant Ads so that it can ***increase engagement*** for its advertisers.
- *Meta*, has appointed Analytics firm where you are working in order to help during the process.
- The idea is to find customers with ***similar backgrounds***, social media history and purchase history so that ***advertisements can be targeted*** for products user will like and buy.
- *Meta,* has agreed to provide you with all the data that you require.

1.6 Unsupervised Learning - An Example

Let's start with a very simple version of the problem –

- For each customer, *Meta* has provided you ***daily time SPENT*** and ***daily posts LIKED*** fields and asked you to **segment** the customers.

You decided that before you do anything it's a great idea to make a **scatterplot** of both fields as it will help you visualize



Scatterplot has helped us in some ***identification***.

Clearly, customers can be *divided* into 2 groups. One group, where they spend *less* time on social media and other that spends *more* time on social media.

Question is :

1. *What **can we do** if there are **more than 2** fields provided to us?*
2. *What if Meta asks us to recognise 4 customer segments?*

1.6 Unsupervised Learning - Sub-Types

- **CLUSTERING**

  - Clustering is a method of ***grouping*** the objects into clusters such that objects with ***most similarities*** remains into a group and has less or no similarities with the objects of another group.

  - Cluster Analysis finds the ***commonalities*** between the data objects and categorizes them as per the *presence & absence* of those commonalities.

- **ASSOCIATION**

  - An association rule is an unsupervised learning method which is used for finding the ***relationships*** between variables in the large database.

  - It determines the set of items that ***occurs together*** in the dataset.

  - A typical example of Association rule is ***Market Basket Analysis***.

1.6 Unsupervised Learning – A Comparison

Advantages	Disadvantages
1. Unsupervised learning is used for ***more complex*** tasks as compared to supervised learning because in unsupervised learning, we ***don't*** have/ need labelled input data	1. Unsupervised learning is intrinsically ***more difficult*** than supervised learning as it does not have corresponding output.
2. Unsupervised learning is preferable as it is ***easy to get*** unlabelled data in comparison to labelled data.	2. The result of the unsupervised learning algorithm might be ***less accurate*** as input data is not labelled and algorithms do not know the exact output in advance

1.6 Unsupervised Learning - Summary

- Unsupervised learning is helpful for finding insights from the data, ***not clearly visible*** to human eye.
- *Unsupervised learning* is much similar as a *human learns* to think by their ***own experiences***, which makes it closer to real AI.
- Unsupervised learning works on ***unlabelled & uncategorized*** data which makes it more important.
- *Common Unsupervised ML Algorithms are:*
  - *K Means Clustering*
  - *KNN (K Nearest Neighbours)*
  - *Hierarchal Clustering*
  - *Anomaly Detection*
  - *Principal Component Analysis*
  - *Anamoly Detection*
  - *Singular Value Decomposition*
  - *Neural Networks.*

1.5 Process of Machine Learning

1. Identification of the Target
2. Fetching of Data relevant to us – Cleaning it appropriately & formatting it as required
3. Formation of Hypothesis function
4. Checking the performance of the hypothesis function
5. Updating it & rechecking the performance
6. Do it till we get the best result