LOVELY PROFESSIONAL UNIVERSITY

REPORT on Machine LEARNING

Submitted by

Muteen Mushatq & **Aditya Khajuria**

Registration No : 12001755 , **12014057**

Programme Name : Btech. CSE (3rd Year)

School of Computer Science & Engineering

Lovely Professional University, Phagwara

INTRODUCTION

What is machine learning?

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Recommendation engines are a common use case for machine learning. Other popular uses include fraud detection, spam filtering, malware threat detection, business process automation (BPA) and Predictive maintenance.

Why is machine learning important?

Machine learning is important because it gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

What are the different types of machine learning?

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches:
supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

- **Supervised learning:** In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the

algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

- **Unsupervised learning:** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.
- **Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.
- **Reinforcement learning:** Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

TECHNOLOGY LEARNT

It had 24 units which was further divided into chapters and then topics so during my whole 6 week course I learned the following:

INTRODUCTION TO MACHINE LEARNING

Overview Of Machine Learning

Definition of Machine Learning: Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM. He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed ". However, there is no universally

accepted definition for machine learning. Different authors define the term differently. We give below two more definitions.

Define an API

An **API (Application Programming Interface)** is a collection of packages, a package is the collection of classes, interfaces and sub-packages. A sub-package is a collection of classes interfaces and sub sub packages etc.

- Machine learning is programming computers to optimize a performance criterion using example data or past experience . We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data.
- The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Features of Machine Learning

Following are some of the important features that make Machine Learning the first choice of software architects.

Step 1: Pre-processing of Data

A data mining technique that involves transforming raw data into an understandable format.

Each algorithm works differently and has different data requirements. For example, some algorithms need numeric features to be normalized, and some do not. Then there's the complication of text, which needs to be split into words and phrases, and in some languages, such as Japanese, that's really difficult! Look for an automated machine learning platform that knows how to best prepare data for each different algorithm, recognizes and prepares text, and follows best practice for data partitioning.

Step 2: Feature Engineering

Feature engineering is the process of altering the data to help machine learning algorithms work better, which is often time-consuming and expensive. While some feature engineering requires domain knowledge of the data and business rules, most feature engineering is generic.

Look for an automated machine learning platform that can automatically engineer new features from existing numeric, categorical, and text features. You will want a system that knows which algorithms benefit from extra feature engineering and which don't, and only generates features that make sense given the data characteristics.

Step 3: Diverse Algorithms

Every dataset contains unique information that reflects the individual events and characteristics of a business. Due to the variety of situations and conditions, one algorithm cannot successfully solve every possible business problem or dataset. Because of this, we need access to a diverse repository of algorithms to test against our data, in order to find the best one for our particular data.

Look for an automated machine learning platform that has dozens, or even hundreds of algorithms. Ask how often new algorithms are added.

Step 4: Algorithm Selection

Having hundreds of algorithms available at your fingertips is great, but unless you are more patient than I am, you don't have time to try each and every one of those algorithms on your data. Some algorithms aren't suited to your data, some are not suited to your data sizes, and some are extremely unlikely to work well on your data.

Look for an automated machine learning platform that knows which algorithms make sense for your data and runs only those. That way you will get better algorithms, faster.

Step 5: Training and Tuning

It's quite standard for machine learning software to train the algorithm on your data. After all, you wouldn't want to manually do Newton-Raphson iteration would you? Probably not. But, often there's still the hyperparameter tuning to worry about. Then you want to do feature selection, to improve both the speed and accuracy of a model.

Look for an automated machine learning platform that uses smart hyperparameter tuning, not just brute force, and knows the most important hyperparameters to tune for each algorithm. Check whether the platform knows which features to include and which to leave out, and which feature selection method works best for different algorithms.

Step 6: Ensembling

In data science jargon, teams of algorithms are called "ensembles" or "blenders." Each algorithm's strengths balance out the weaknesses of another. Ensemble models typically outperform individual algorithms because of their diversity.

Look for an automated machine learning platform that finds the optimal algorithms to blend together, includes a diverse range of algorithms, and tunes the weighting of the algorithms within each blender.

Step 7: Head-to-Head Model Competitions

You won't know in advance which algorithm performs best on your data. So, you need to compare the accuracy and speed of different algorithms on your data, regardless of which programming language or machine learning library they came from. You can think of it as being like a competition amongst the models, where the best model wins!

Look for an automated machine learning platform that builds and trains dozens of algorithms, compares the results, and ranks the best algorithms based on your needs. The platform should compare accuracy, speed, and individual predictions.

Step 8: Human-Friendly Insights

Albert Einstein once said, "If you can't explain it simply, you don't understand it well enough." Over the past few years machine learning and artificial intelligence have made massive strides forward in predictive power, but at the price of complexity. It is not enough for a machine learning solution to score well on only accuracy and speed. You also have to trust the answers it is giving. In regulated

industries, you have to justify the model to the regulator. And in marketing, you need to align the marketing message with the audience the model has chosen.

Look for an automated machine learning platform that explains model decisions in a humaninterpretable manner. The platform should show which features are most important for each model and show the patterns fitted for each feature. Ask whether the platform can provide worked examples, including the key reasons why a prediction is either high or low. Check whether the platform automatically writes detailed model documentation, and how well that documentation complies with your regulator's requirements.

Step 9: Easy Deployment

A recent Harvard Business Review article described a team of analysts that built an impressive predictive model, using the latest in machine learning algorithms. But, the business lacked the infrastructure needed to directly implement the trained model in a production setting, and the model was "too complex for the IT team to reproduce".

Look for an automated machine learning platform that offers easy deployment, including one-click deploy, that can be operated by a business person. Ask how many deployment options are available, whether models can be deployed on your standard system hardware, and whether the platform pretests exported scoring code to ensure it generates the same answers as in training. Also, check whether the vendor has a large technical support team located all around the world that can provide data science and engineering support 24 hours per day.

Step 10: Model Monitoring and Management

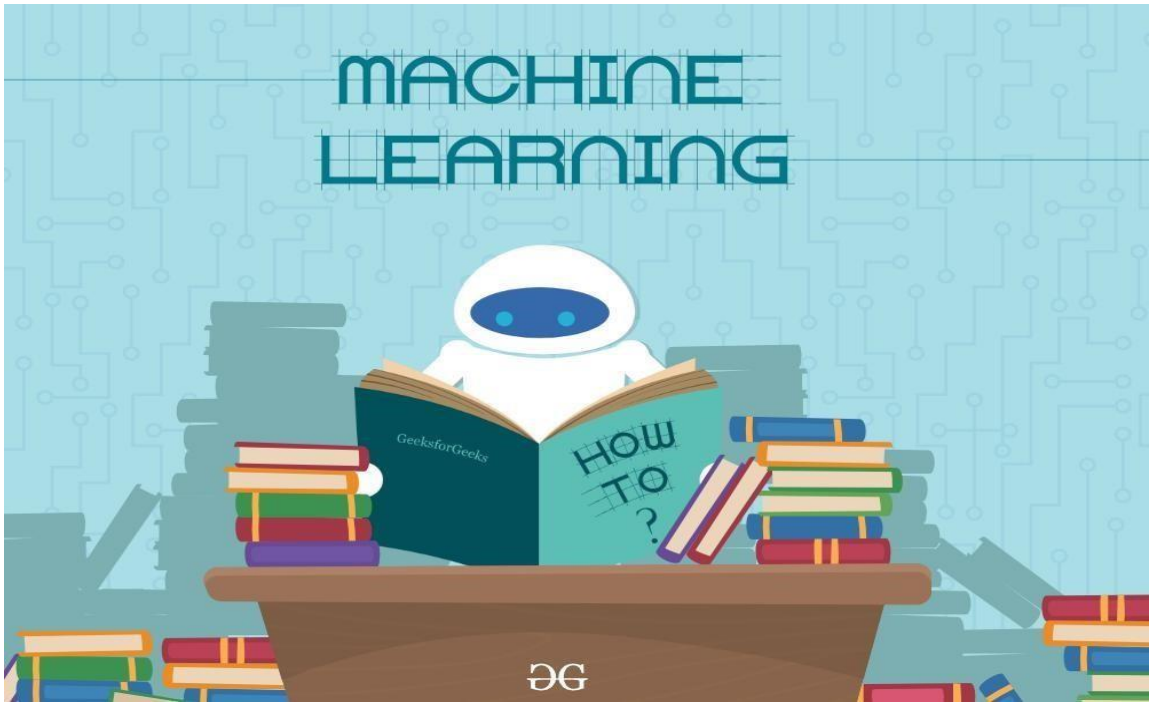
In a constantly changing world, your AI applications need to keep up to date with the latest trends. Look for an automated machine learning platform that proactively identifies when a model's performance is deteriorating over time, making it easy to compare predictions to actual results, simplifying the task of training a new model on the latest data.

.

Machine Learning

Machine learning (ML) refers to a system's ability to acquire, and integrate knowledge through largescale observations, and to improve, and extend itself by learning new knowledge rather than by being programmed with that knowledge. ML techniques are used in intelligent tutors to acquire new knowledge about students, identify their skills, and learn new teaching approaches. They improve teaching by repeatedly observing how students react and generalize rules about the domain or student. The role of ML techniques in a tutor is to independently observe and evaluate the tutor's actions. ML tutors customize their teaching by reasoning about large groups of students, and tutorstudent interactions, generated through several components. A *performance element* is responsible for making improvements in the tutor, using perceptions of tutor/student interactions, and knowledge about the student's reaction to decide how to modify the tutor to perform better in the future. ML techniques are used to identify student learning strategies, such as, which activities do students select most frequently and in which order. Analysis of student behavior leads to greater student learning outcome by providing tutors with useful diagnostic information for generating feedback.Steps For Compiling and Executing Programs

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: **The ability to learn**. Machine learning is actively being used today, perhaps in many more places than one would expect.



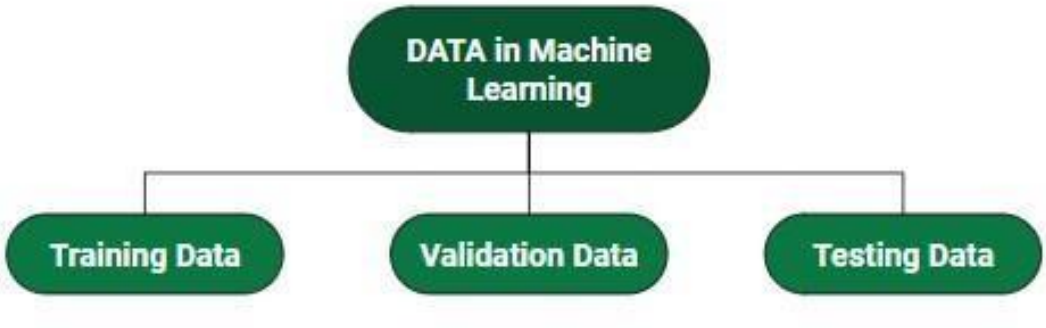
DATA: It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, Artificial Intelligence. Without data, we can't train any model and all modern research and automation will go

in vain. Big Enterprises are spending lots of money just to gather as much certain data as possible.
Example: Why did Facebook acquire WhatsApp by paying a huge price of $19 billion?
The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information of their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

INFORMATION: Data that has been interpreted and manipulated and has now some meaningful inference for the users.
KNOWLEDGE: Combination of inferred information, experiences, learning, and insights. Results in awareness or concept building for an individual or organization.

How we split data in Machine Learning?

- **Training Data:** The part of data we use to train our model. This is the data that your model actually sees(both input and output) and learns from.
- **Validation Data:** The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.
- **Testing Data:** Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values(without seeing actual output). After prediction, we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.



Consider an example:

There's a Shopping Mart Owner who conducted a survey for which he has a long list of questions and answers that he had asked from the customers, this list of questions and answers is **DATA**. Now every time when he wants to infer anything and can't just go through each and every question of thousands of customers to find something relevant as it would be time-consuming and not helpful. In order to reduce this overhead and time wastage and to make work easier, data is manipulated through software,

calculations, graphs, etc. as per own convenience, this inference from manipulated data is **Information**. So, Data is a must for Information. Now **Knowledge** has its role in differentiating between two individuals having the same information. Knowledge is actually not technical content but is linked to the human thought process.

Before starting your programming, make sure you have one text editor in place and you have enough experience to write a computer program, save it in a file, and finally execute it.

Different Forms of Data

- **Numeric Data** : If a feature represents a characteristic measured in numbers , it is called a numeric feature.
- **Categorical Data :** A categorical feature is an attribute that can take on one of the limited , and usually fixed number of possible values on the basis of some qualitative property . A categorical feature is also called a nominal feature.
- **Ordinal Data** : This denotes a nominal variable with categories falling in an ordered list . Examples include clothing sizes such as small, medium , and large , or a measurement of customer satisfaction on a scale from “not at all happy” to “very happy”.

Properties of Data

1. **Volume:** Scale of Data. With the growing world population and technology at exposure, huge data is being generated each and every millisecond.
2. **Variety:** Different forms of data – healthcare, images, videos, audio clippings.
3. **Velocity:** Rate of data streaming and generation.
4. **Value:** Meaningfulness of data in terms of information that researchers can infer from it.
5. **Veracity:** Certainty and correctness in data we are working on.

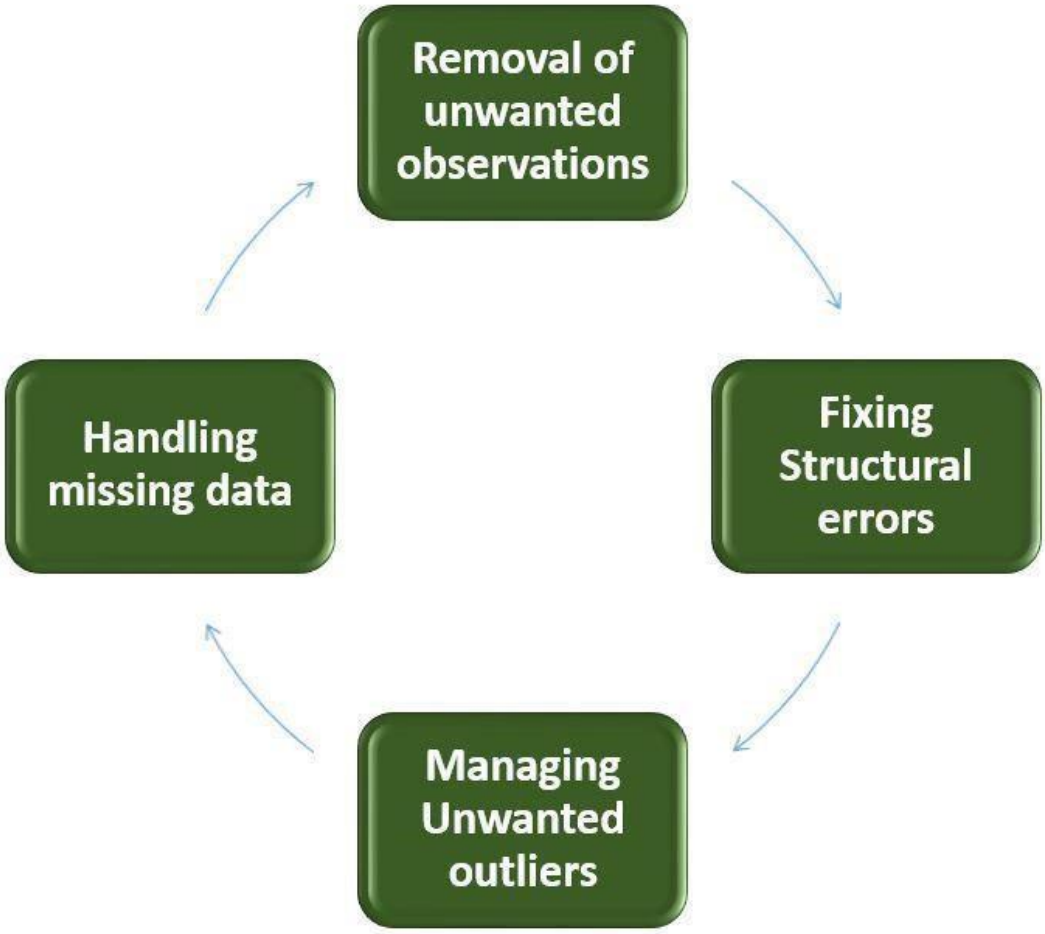
Some facts about Data

- As compared to 2005, 300 times i.e. 40 Zettabytes (1ZB=10^21 bytes) of data will be generated by 2020.
- By 2011, the healthcare sector has a data of 161 Billion Gigabytes
- 400 Million tweets are sent by about 200 million active users per day
- Each month, more than 4 billion hours of video streaming is done by the users.
- 30 Billion different types of content are shared every month by the user.
- It is reported that about 27% of data is inaccurate and so 1 in 3 business idealists or leaders don’t trust the information on which they are making decisions.

The above-mentioned facts are just a glimpse of the actually existing huge data statistics. When we talk in terms of real-world scenarios, the size of data currently presents and is getting generated each and every moment is beyond our mental horizons to imagine.

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data cleaning. Professional data scientists usually invest a very large portion of their time in this step because of the belief that **"Better data beats fancier algorithms"**. If we have a well-cleaned dataset, there are chances that we can get achieve good results with simple algorithms also, which can prove very beneficial at times especially in terms of computation when the dataset size is large.
Obviously, different types of data will require different types of cleaning. However, this systematic approach can always serve as a good starting point.

Steps involved in Data Cleaning:



1. **Removal of unwanted observations**

   This includes deleting duplicate/ redundant or irrelevant values from your dataset. Duplicate observations most frequently arise during data collection and Irrelevant observations are those that don't actually fit the specific problem that you're trying to solve.

   - Redundant observations alter the efficiency by a great extent as the data repeats and may add towards the correct side or towards the incorrect side, thereby producing unfaithful results.
   - Irrelevant observations are any type of data that is of no use to us and can be removed directly.

1. **Fixing Structural errors**

The errors that arise during measurement, transfer of data, or other similar situations are called structural errors. Structural errors include typos in the name of features, the same attribute with a different name, mislabelled classes, i.e. separate classes that should really be the same, or inconsistent capitalization.

For example, the model will treat America and America as different classes or values, though they represent the same value or red, yellow, and red-yellow as different classes or attributes, though one class can be included in the other two classes. So, these are some structural errors that make our model inefficient and give poor quality results.

2. **Managing Unwanted outliers**

   Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. Generally, we should not remove outliers until we have a legitimate reason to remove them.
   Sometimes, removing them improves performance, sometimes not. So, one must have a good reason to remove the outlier, such as suspicious measurements that are unlikely to be part of real data.

3. **Handling missing data**

   Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important. The two most common ways to deal with missing data are:

   - Dropping observations with missing values.
     - The fact that the value was missing may be informative in itself.
     - Plus, in the real world, you often need to make predictions on new data even if some of the features are missing.
   - Imputing the missing values from past observations.
     - Again, "missingness" is almost always informative in itself, and you should tell your algorithm if a value was missing.
     - Even if you build a model to impute your values, you're not adding any real information. You're just reinforcing the patterns already provided by other features.

Missing data is like missing a puzzle piece. If you drop it, that's like pretending the puzzle slot isn't there. If you impute it, that's like trying to squeeze in a piece from somewhere else in the puzzle. So, missing data is always an informative and an indication of something important. And we must be aware of our algorithm of missing data by flagging it. By using this technique of flagging and filling, you are essentially allowing the algorithm to estimate the optimal constant for missingness, instead of just filling it in with the mean.

Some data cleansing tools

- Openrefine

- Trifacta Wrangler
- TIBCO Clarity
- Cloudingo
- IBM Infosphere Quality Stage

Conclusion:

So, we have discussed four different steps in data cleaning to make the data more reliable and to produce good results. After properly completing the Data Cleaning steps, we'll have a robust dataset that avoids many of the most common pitfalls. This step should not be rushed as it proves very beneficial in the further process.

When Regression is chosen?

A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used, the simplest is linear regression. It tries to fit data with the best hyperplane which goes through the points.

Regression Analysis is a statistical process for estimating the relationships between the dependent variables or criterion variables and one or more independent variables or predictors. Regression analysis explains the changes in criteria in relation to changes in select predictors. The conditional expectation of the criteria is based on predictors where the average value of the dependent variables is given when the independent variables are changed. Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.

Types of Regression:

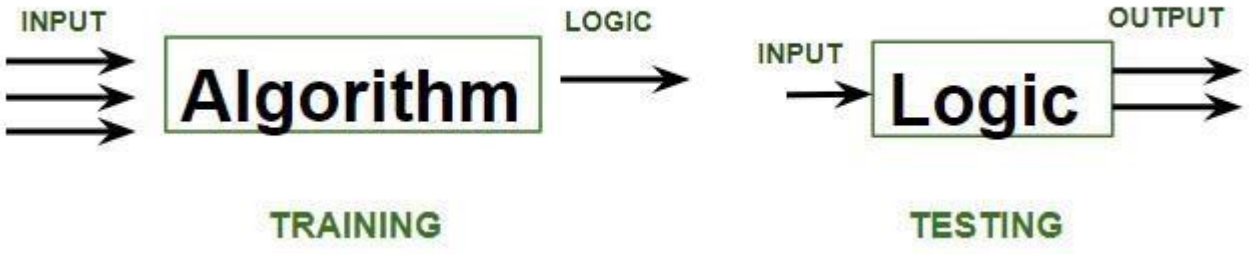
- **Linear regression** is used for predictive analysis. Linear regression is a linear approach for modelling the relationship between the criterion or the scalar response and the multiple predictors or explanatory variables. Linear regression focuses on the conditional probability distribution of the response given the values of the predictors. For linear regression, there is a danger of overfitting. The formula for linear regression is: Y' = bX + A.
- **Polynomial regression** is used for curvilinear data. Polynomial regression is fit with the method of least squares. The goal of regression analysis is to model the expected value of a dependent variable y in regards to the independent variable x. The equation for polynomial regression is: l = .
- **Stepwise regression** is used for fitting regression models with predictive models. It is carried out automatically. With each step, the variable is added or subtracted from the set of explanatory variables. The approaches for stepwise regression are forward selection, backward elimination, and bidirectional elimination. The formula for stepwise regression is .

- **Ridge regression** is a technique for analyzing multiple regression data. When multicollinearity occurs, least squares estimates are unbiased. A degree of bias is added to the regression estimates, and as a result, ridge regression reduces the standard errors. The formula for ridge regression is .
- **Lasso regression** is a regression analysis method that performs both variable selection and regularization. Lasso regression uses soft thresholding. Lasso regression selects only a subset of the provided covariates for use in the final

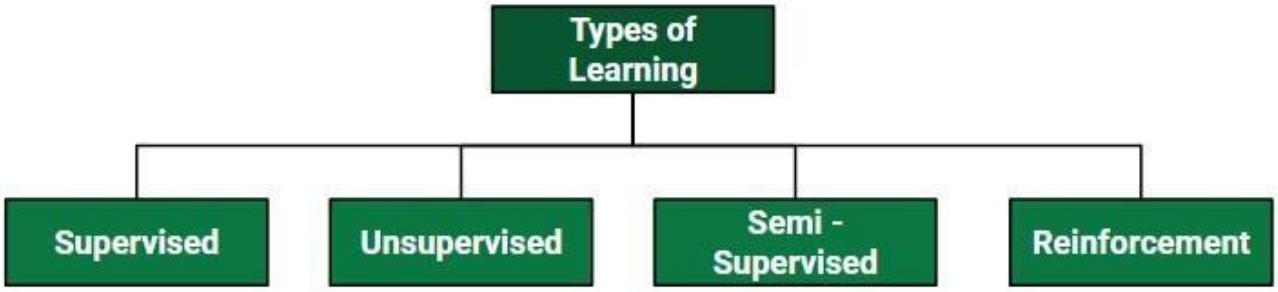
  model. Lasso regression is .
- **ElasticNet regression** is a regularized regression method that linearly combines the penalties of the lasso and ridge methods. ElasticNet regression is used for support vector machines, metric learning, and portfolio optimization. The

  penalty function is given by: .
  Below is the simple implementation:

ML | Types of Learning – Supervised Learning



A machine is said to be learning from **past Experiences**(data feed-in) with respect to some class of **tasks** if its **Performance** in a given Task improves with the Experience. For example, assume that a machine has to predict whether a customer will buy a specific product let's say "Antivirus" this year or not. The machine will do it by looking at the **previous knowledge/past experiences** i.e the data of products that the customer had bought every year and if he buys Antivirus every year, then there is a high probability that the customer is going to buy an antivirus this year as well. This is how machine learning works at the basic conceptual level.



Supervised learning is when the model is getting trained on a labelled dataset. A **labelled** dataset is one that has both input and output parameters. In this type of learning both training and validation, datasets are labelled as shown in the figures below.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

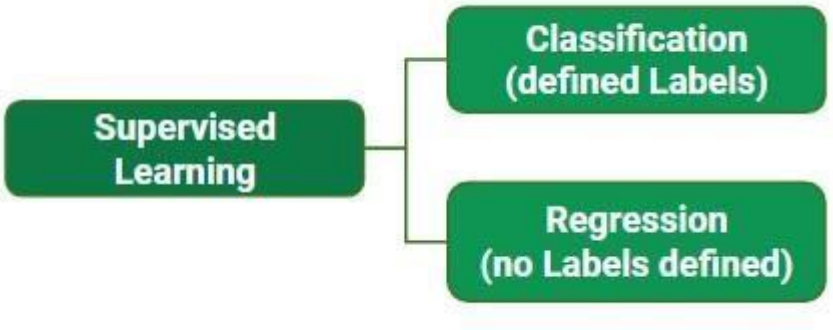
Figure B: REGRESSION

Both the above figures have labelled data set as follows:

- **Figure A:** It is a dataset of a shopping store that is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age, and salary.
  Input: Gender, Age, Salary
  Output: Purchased i.e. 0 or 1; 1 means yes the customer will purchase and 0 means that the customer won't purchase it.
- **Figure B:** It is a Meteorological dataset that serves the purpose of predicting wind speed based on different parameters.
  Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction
  Output: Wind Speed

Training the system: While training the model, data is usually split in the ratio of 80:20 i.e. 80% as training data and the rest as testing data. In training data, we feed input as well as output for 80% of data. The model learns from training data only. We use different machine learning algorithms(which we will discuss in detail in the next articles) to build our model.

Learning means that the model will build some logic of its own.
Once the model is ready then it is good to be tested. At the time of testing, the input is fed from the remaining 20% of data that the model has never seen before, the model will predict some value and we will compare it with the actual output and calculate the accuracy.



Types of Supervised Learning:

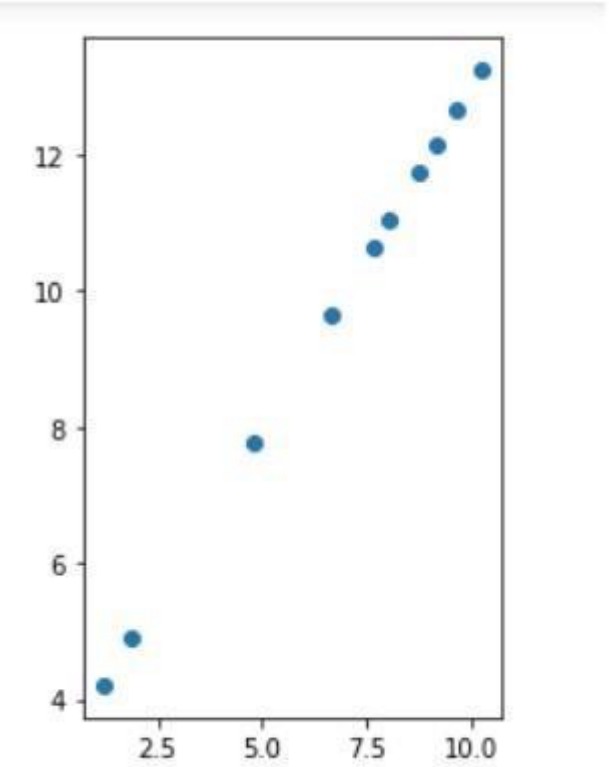
A. Classification: It is a Supervised Learning task where output is having defined labels(discrete value). For example in above Figure A, Output – Purchased has defined labels i.e. 0 or 1; 1 means the customer will purchase, and 0 means that the customer won't purchase. The goal here is to predict discrete values belonging to a particular class and evaluate them on the basis of accuracy.

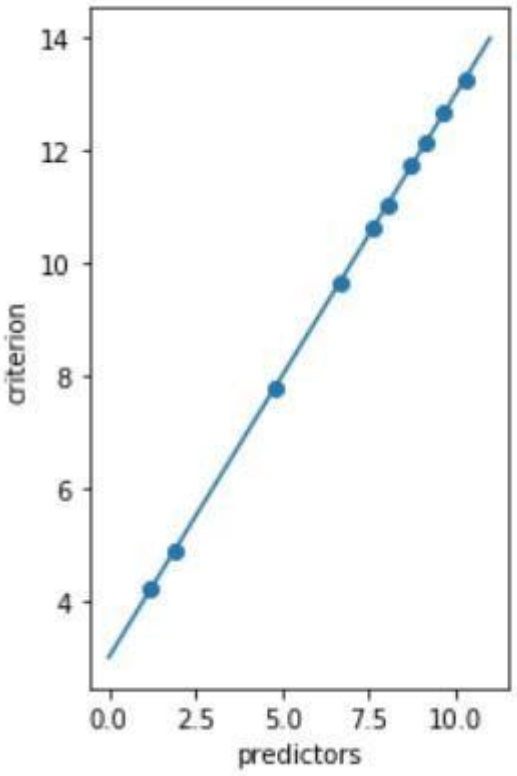
It can be either binary or multi-class classification. In **binary** classification, the model predicts either 0 or 1; yes or no but in the case of **multi-class** classification, the model predicts more than one class. **Example:** Gmail classifies mails in more than one class like social, promotions, updates, and forums.

B. Regression: It is a Supervised Learning task where output is having continuous value. For example in above Figure B, Output – Wind Speed is not having any discrete value but is continuous in a particular range. The goal here is to predict a value as much closer to the actual output value as our model can and then evaluation is done by calculating the error value. The smaller the error the greater the accuracy of our regression model.

Example of Supervised Learning Algorithms:

- Linear Regression
- Logistic Regression
- Nearest Neighbor
- Gaussian Naive Bayes
- Decision Trees
- Support Vector Machine (SVM)
- Random Forest

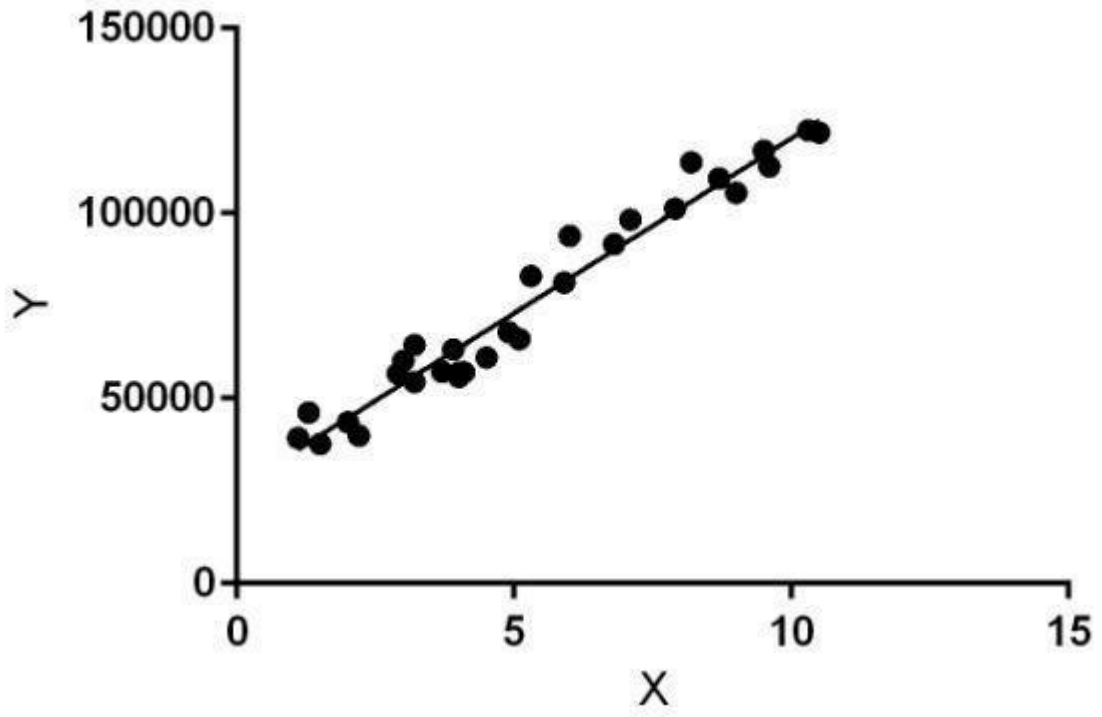




•

ML | Linear Regression

- **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.
- **Hypothesis function for Linear Regression :**

$$y = \theta_1 + \theta_2.x$$

- While training the model we are given :

x: input training data (univariate – one input variable(parameter)) **y:** labels to data (supervised learning)

- When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.

θ_1: intercept **θ_2:** coefficient of x

- Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.
- **How to update θ_1 and θ_2 values to get the best fit line ?**

- **Cost Function (J):**

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and $minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$ true y value (y).

- $$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

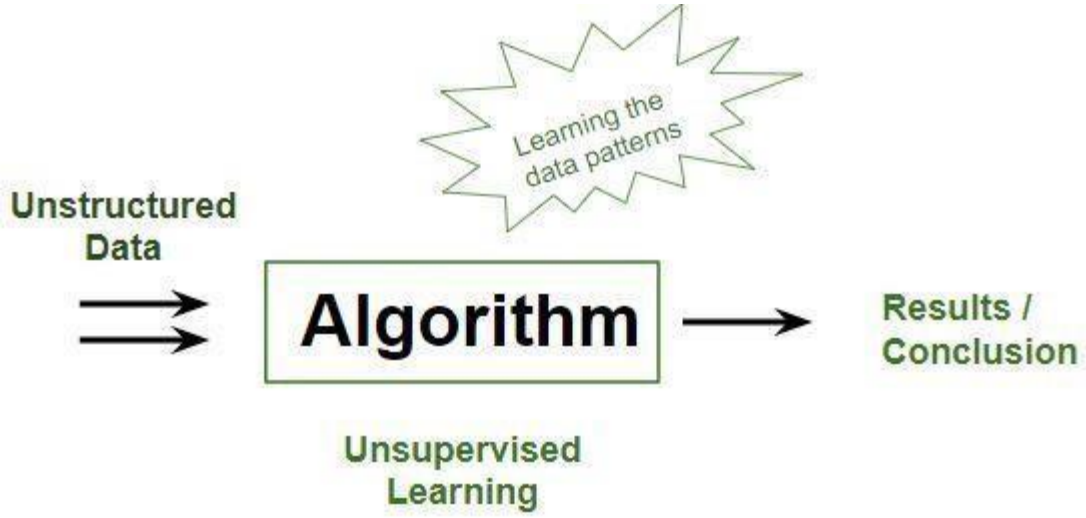
- Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).
- Gradient Descent**:**

To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

ML | Types of Learning

Unsupervised Learning:

Or unsupervised machine learning analyzes and clusters unlabeled datasets using machine learning algorithms. These algorithms find hidden patterns and data without any human intervention, i.e., we don't give output to our model. The training model has only input parameter values and discovers the groups or patterns on its own. Data-set in Figure A is Mall data that contains information about its clients that subscribe to them. Once subscribed they are provided a membership card and the mall has complete information about the customer and his/her every purchase. Now using this data and unsupervised learning techniques, the mall can easily group clients based on the parameters we are feeding in.

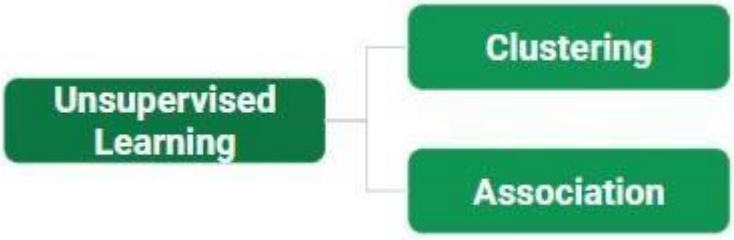


CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure A

The input to the unsupervised learning models is as follows:

- **Unstructured data**: May contain noisy(meaningless) data, missing values, or unknown data
- **Unlabeled data**: Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to the labeled one in the Supervised approach.



Types of Unsupervised Learning are as follows:

- **Clustering:** Broadly this technique is applied to group data based on different patterns, such as similarities or differences, our machine model finds. These algorithms are used to process raw, unclassified data objects into groups. For example, in the above figure, we have not given output parameter values, so this technique will be used to group clients based on the input parameters provided by our data.
- **Association:** This technique is a rule-based ML technique that finds out some very useful relations between parameters of a large data set. This technique is basically used for market basket analysis that helps to better understand the relationship between different products. For e.g. shopping stores use algorithms based on this technique to find out the relationship between the sale of one product w.r.t to another's sales based on customer behavior. Like if a customer buys milk, then he may also buy bread, eggs, or butter. Once trained well, such models can be used to increase their sales by planning different offers.

Some algorithms:

- K-Means Clustering
- DBSCAN – Density-Based Spatial Clustering of Applications with Noise
- BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies Hierarchical Clustering Semi-supervised Learning:

As the name suggests, its working lies between Supervised and Unsupervised techniques. We use these techniques when we are dealing with data that is a little bit labeled and the rest large portion of it is unlabeled. We can use the unsupervised techniques to predict labels and then feed these labels to supervised techniques. This technique is mostly applicable in the case of image data sets where usually all images are not labeled.

Reinforcement Learning:

In this technique, the model keeps on increasing its performance using Reward Feedback to learn the behavior or pattern. These algorithms are specific to a particular problem e.g. Google Self Driving car, AlphaGo where a bot competes with humans and even itself to get better and better performers in Go Game. Each time we feed in data, they learn and add the data to their knowledge which is training data. So, the more it learns the better it gets trained and hence experienced.

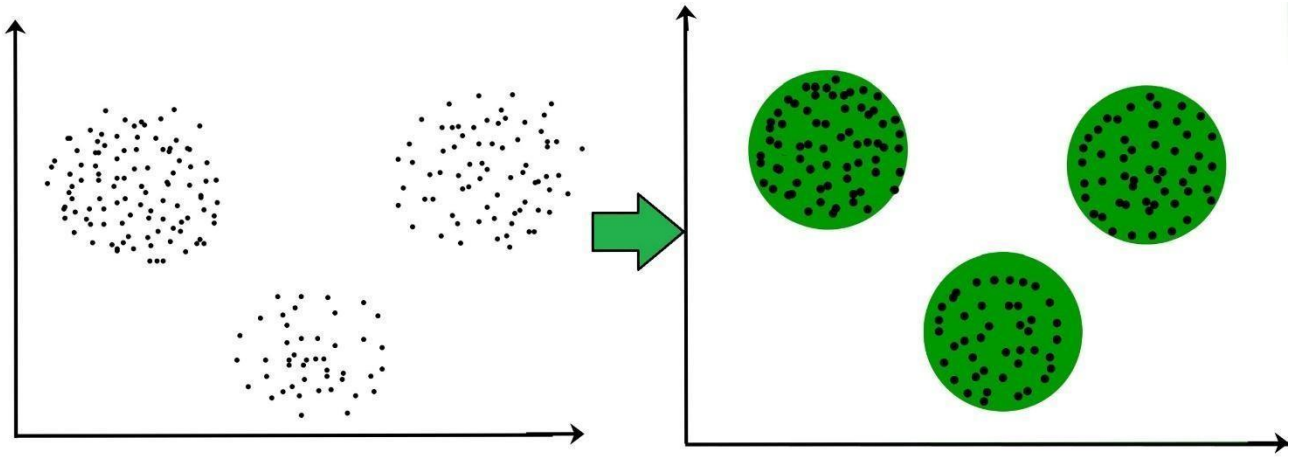
- Agents observe input.
- An agent performs an action by making some decisions.
- After its performance, an agent receives a reward and accordingly reinforces and the model stores in state-action pair of information.
- Temporal Difference (TD)
- Q-Learning
- Deep Adversarial Networks

Clustering in Machine Learning

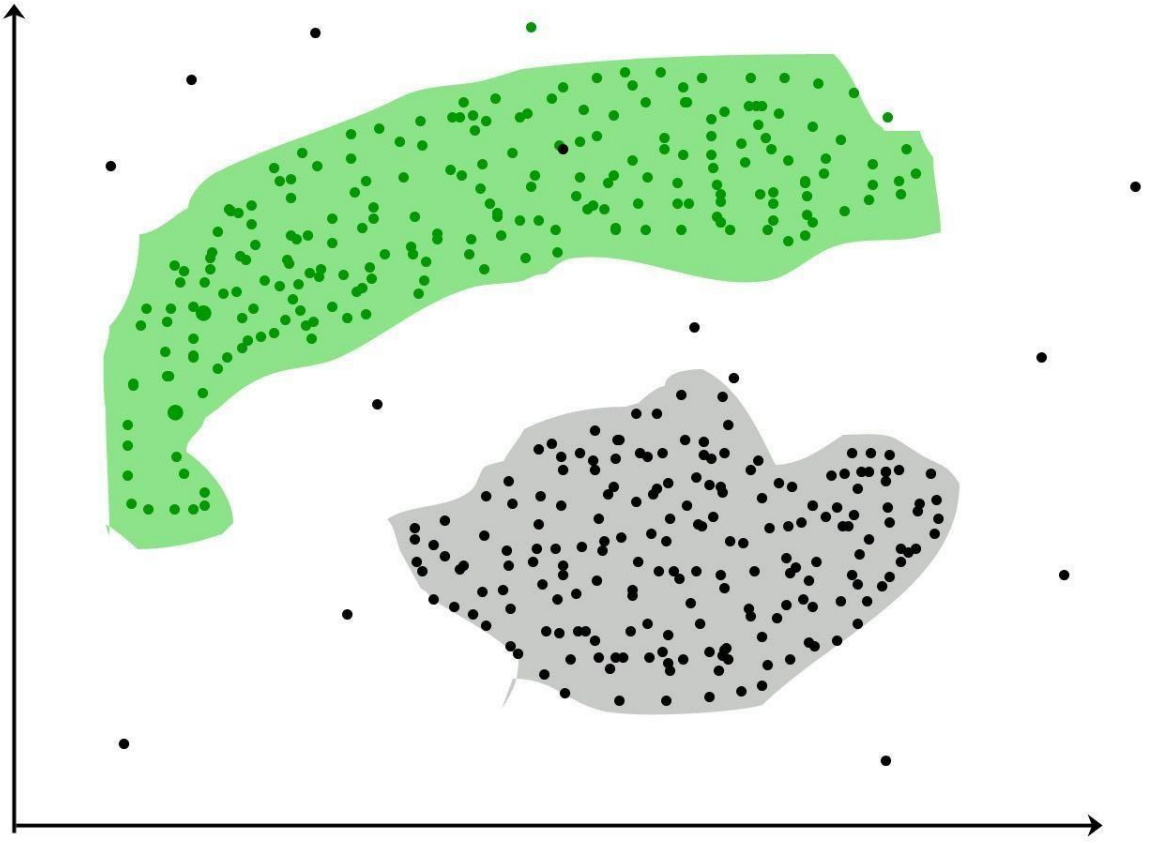
Introduction to Clustering

It is basically a type of *unsupervised learning method*. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



It is not necessary for clusters to be spherical. Such as :



DBSCAN: Density-based Spatial Clustering of Applications with Noise

These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center. Various distance methods and techniques are used for the calculation of the outliers.

Why Clustering?

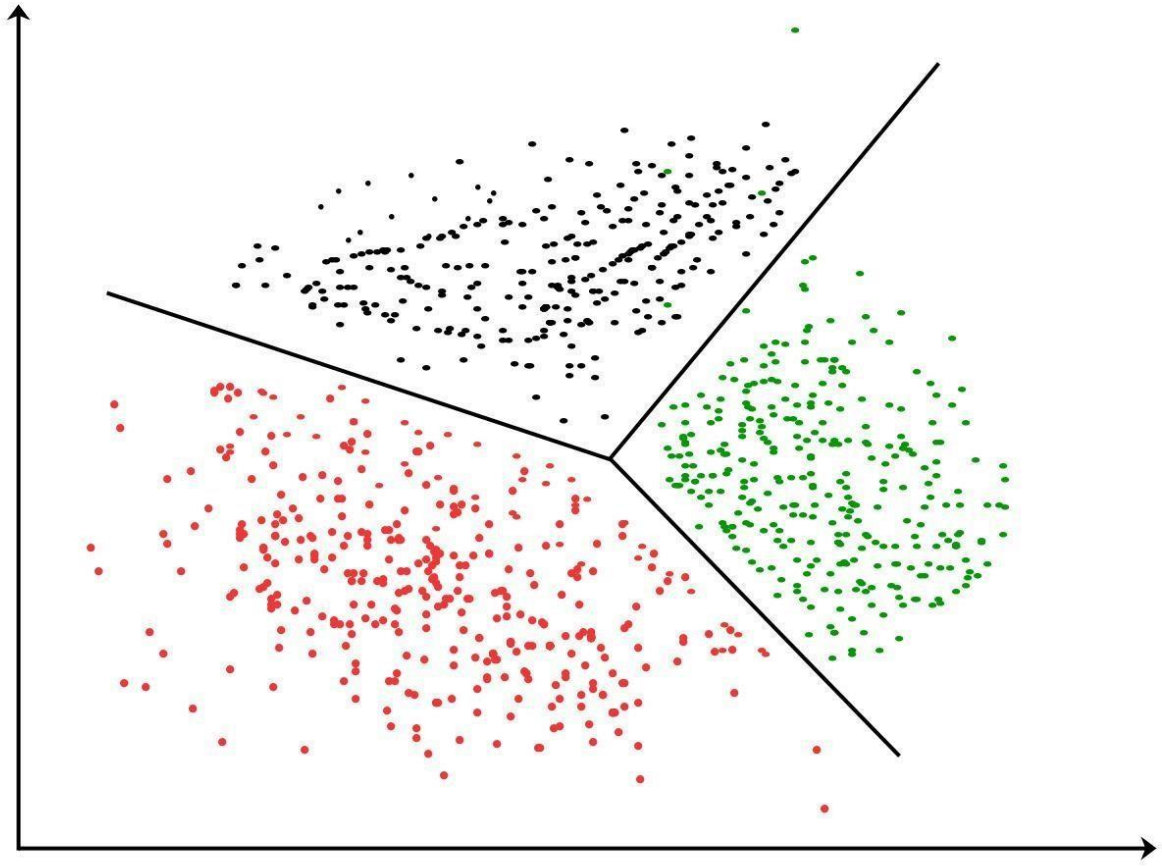
Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Clustering Methods :

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure)*, etc.
- **Hierarchical Based Methods:** The clusters formed in this method form a treetype structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
- **Agglomerative** (bottom-up approach)
- **Divisive** (top-down approach) examples CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), etc.
- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search), etc.
- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest), etc.

Clustering Algorithms :

K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



Applications of Clustering in different fields

- **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
- **Biology:** It can be used for classification among different species of plants and animals.
- **Libraries:** It is used in clustering different books on the basis of topics and information.
- **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.

City Planning: It is used to make groups of houses and to study their values based on their geographical locations and other factors presented.

Continuous Improvement

Machine Learning algorithms are capable of learning from the data we provide. As new data is provided, the model's accuracy and efficiency to make decisions improve with subsequent training. Giants like Amazon, Walmart, etc collect a huge volume of new data every day. The accuracy of finding associated products or recommendation engine improves with this huge amount of training data available.

Automation for everything

A very powerful utility of Machine Learning is its ability to automate various decisionmaking tasks. This frees up a lot of time for developers to use their time to more productive use. For example, some common use we see in our daily life is social media sentiment analysis and chatbots. The moment a negative tweet is made related to a product or service of a Company, a chatbot instantly replies as first-level customer support. Machine Learning is changing the world with its automation for almost everything we can think of.

Trends and patterns identification

This advantage is a no brainer. All of us interested in Machine Learning technology are well aware of how the various Supervised, Unsupervised and Reinforced learning algorithms can be used for various classification and regression problems. We identify various trends and patterns with a huge amount of data using this technology. For example, Amazon analyzes the buying patterns and search trends of its customers and predicts products for them using Machine Learning algorithms.

Wide range of applications

Machine Learning is used in every industry these days, for example from Defence to Education. Companies generate profits, cut costs, automate, predict the future, analyse trends and patterns from the past data, and many more. Applications like GPS Tracking for traffic, Email spam filtering, text prediction, spell check and correction, etc are a few used widely these days. Machine Learning is a branch of Artificial Intelligence, the latest trends and applications can be found in Artificial Intelligence Trends in 2020.

CONCLUSION

Practical knowledge means the visualization of the knowledge, which we read in our books. For this, we perform experiments and get observations. Practical knowledge is very important in every field. One must be familiar with the problems related to that field so that he may solve them and become a successful person.

After achieving the proper goal in life, an engineer has to enter in professional life. According to this life, he has to serve an industry, may be public or private sector or self own. For the efficient work in the field, he must be well aware of the practical knowledge as well as theoretical knowledge.

Due to all above reasons and to bridge the gap between theory and practical, our Engineering curriculum provides a practical training of 45 days. During this period a student work in the industry and get well all type of experience and knowledge about the working of companies and hardware and software tools.

I have undergone my 45 days summer training in 5th sem at SIMPLILEARN. This report is based on the knowledge, which I acquired during my 45 days of summer training.

BIBLIOGRAPHY

- SIMPLILEARN Machine Learning Course
- GeekforGeeks website
- W3schools website