**Machine Learning**

Machine learning is a subset of Artificial Intelligence and also a data analytics technique, which learn from experience. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Deep learning is a subset of (specialized form) machine learning.

**Why Machine Learning Matters**

With the rise in big data, machine learning has become a key technique for solving problems in areas, such as:

* Computational finance, for credit scoring and algorithmic trading
* Computational biology, for tumour detection, drug discovery, and DNA sequencing
* Energy production, for price and load forecasting
* Automotive, aerospace, and manufacturing, for predictive maintenance
* Natural language processing, for voice recognition applications
* Image processing and computer vision, for face recognition, motion detection, and object detection

**More Data, More Questions, Better Answers**

Machine learning algorithms find natural patterns in data that generate insight and helps the industry practitioners make better decisions and predictions. They are used every day to make critical decisions in medical diagnosis, stock trading, energy load forecasting, and more.

For example, media sites rely on machine learning to sift through millions of options to give you song or movie recommendations. Retailers use it to gain insight into their customers’ purchasing behaviour.

**When Should we Use Machine Learning?**

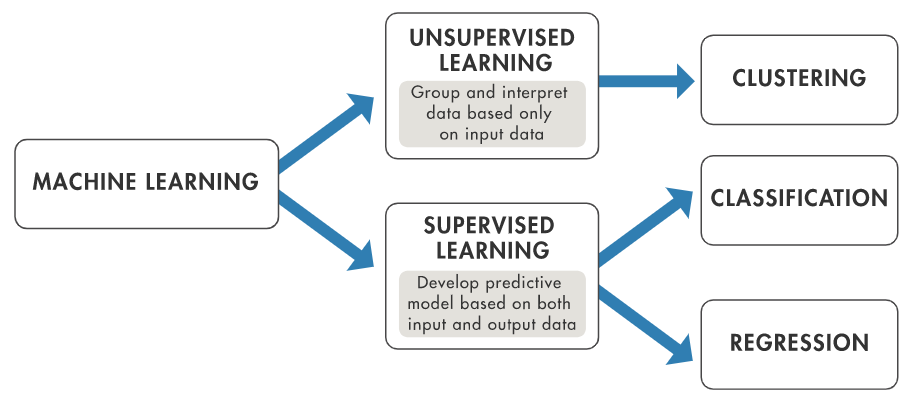
Consider using machine learning when you have a complex task or problem involving a large amount of data and lots of variables, but no existing formula or equation.

For example, machine learning is a good option if you need to handle the following situations:

* The rules of a task are constantly changing - as in fraud detection from transaction records
* The nature of the data keeps changing, and the program needs to adapt - as in automated trading, energy demand forecasting, and predicting shopping trends
* Hand-written rules and equations are too complex - as in face recognition and speech recognition

**How Machine Learning Works**

Machine learning uses two types of techniques: **Supervised learning**, which trains a model on known input and output data so that it can predict future outputs, and **Unsupervised learning**, which finds hidden patterns or intrinsic structures in input data.



**Supervised Learning**

Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying to predict.

**Supervised learning** uses **Regression** and **Classification** techniques to develop predictive models.

**Regression techniques** predict continuous responses - for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading.

Use regression techniques if you are working with a data range or if the nature of your response is a real number, such as temperature or the time until failure for a piece of equipment.

Common regression algorithms include linear model, nonlinear model, stepwise regression, Gradient Descent Regression, Support Vector Regression, Ridge and Lasso Regressions.

**Classification techniques** predict discrete responses - for example, whether an email is genuine or spam, or whether a tumour is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring.

Use classification if your data can be tagged, categorized, or separated into specific groups or classes. For example, applications for hand-writing recognition use classification to recognize letters and numbers. In image processing and computer vision, unsupervised pattern recognition techniques are used for object detection and image segmentation.

Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbour, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.

**Using Supervised Learning to Predict Heart Attacks:** Suppose clinicians want to predict whether someone will have a heart attack within a year. They have data on previous patients, including age, weight, height, and blood pressure. They know whether the previous patients had heart attacks within a year. So, the problem is combining the existing data into a model that can predict whether a new person will have a heart attack within a year.

**Unsupervised Learning**

Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labelled responses.

**Clustering** is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis and market research.

For example, if a cell phone company wants optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers. A phone can only talk to one tower at a time, so the team uses clustering algorithms to design the best placement of cell towers to optimize signal reception for groups, or clusters, of their customers.

Common algorithms for performing clustering include k-means and k-medoids, Apriori algorithms, hierarchical clustering, Gaussian mixture models and hidden Markov models.

**How Do You Decide Which Machine Learning Algorithm to Use?**

Choosing the right algorithm can seem overwhelming—there are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning.

There is no best method or one size fits all. Finding the right algorithm is partly just trial and error - even highly experienced data scientists can’t tell whether an algorithm will work without trying it out. But algorithm selection also depends on the size and type of data you’re working with, the insights you want to get from the data, and how those insights will be used.

**List of Common Machine Learning Algorithms Used in Industries**

List of commonly used machine learning algorithms and these algorithms can be applied to almost any data problems:

* Linear Regression
* Logistic Regression
* Decision Tree
* Time Series - ARMA, ARIMA, Auto ARIMA
* SVM
* Naive Bayes
* kNN
* K-Means
* Random Forest
* ANN
* Dimensionality Reduction Algorithms
* Apriori Algorithms
* Gradient Boosting algorithms (Ensemble models)
* GBM
* XGBoost
* LightGBM
* CatBoost

**Type of Regression methods**

1. Linear Regression (Simple Linear Regression / Multi Regression)

* Dependant variable is continuous
* Relation between dependant and independent variables are linear in nature

2. Polynomial Regression

* Relation between dependant and independent variables are non-linear
* Method: lm

3. Ordinary Least Squares Regression

* Same as linear regression

4. Quantile Regression

* Extn. Of linear regression. Quantile = percentile.
* When to use: if the variability of dependant variable increases with the increase of independent variable - this violates assumption of linear regression (normal errors with constant variance)

5. Ridge and Lasso Regressions

* Regularize (shrink) co-efficients; estimated co-efficients are pushed towards ‘0’
* Ridge: meta-parameter is ‘alpha’
* Lasso: meta-parameter is ‘lambda’
* Use Ridge when features are not highly correlated and when you want to perform feature selection as Lasso only selects one from bunch of highly correlated features
* Library: glmnet
* Model: cv.glmnet

6. Ordinal Regression

* Used when dependant variable is ordinal in nature
* Ordinal variable – which follows order (ex. Very good, good, fair, bad, very bad)
* Ex. Survey responses
* Library: ordinal

7. Poisson Regression

* Used when dependant variable has count data and has Poisson distribution
* Poisson distribution: how many times an event is likely to occur in particular time
* Count data should not be whole numbers or negative
* Ex. No. of calls in customer care
* Method: glm; family: poisson

8. Support Vector Regression

* It tries to fit error in certain threshold (margin)
* Ex. Financial forecasting
* Library: e1071

9. Gradient Descent Regression

* Regression model using Gradient Descent methodology
* It is used to minimize the cost function (mse / rmse func.s etc)
* Gradient point is the steepest point and which will reduce the error
* Library: gbm
* Method: gbm
* Distribution: gaussian

10. Step wise Regression

* Method of regressing multiple variables while simultaneously removing that aren’t important
* It does regression multiple times
* Method: step(linear\_model)

**Linear Regression**

A linear model (simple / Multi Regression) is an algorithm that makes a prediction by simply computing a weighted sum of the input features plus a bias term ‘c’ (also referred to as the intercept term). By using linear regression model, we build the relationship between a dependent variable ‘y’ and one or more independent variables ‘x’.

A multiple linear regression model with ‘k’ predictor variables x1, x2, ..., xk and a response y, can be written as:

**y = c + m1x1+m2x2+ ……+mkxk + e**

where y = Dependent Variable

c = Intercept or bias term

x1, x2 … xn= Independent Variables

m1, m2 … mk = Regression coefficients

e = Residual

**Note: Simple Linear Regression yhat = c + mx + e**

**Four key assumptions associated with a linear regression model**

1. Linearity: The relationship between ‘x’ and the mean of ‘y’ is linear
2. Homoscedasticity: The variance of residual is the same for any value of ‘x’
3. Independence: Variables are independent of each other
4. Normality: For any fixed value of ‘x’, ‘y’ is normally distributed

**Most commonly known performance evaluation metrics include:**

**R-squared (R2),** which is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between the observed outcome values and the predicted values by the model. The Higher the R-squared, the better the model.

**Root Mean Squared Error (RMSE),** which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, the RMSE is the square root of the mean squared error (MSE), which is the average squared difference between the observed actual outcome values and the values predicted by the model. So, MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE). The lower the RMSE, the better the model.

**Residual Standard Error (RSE),** also known as the model sigma, is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model. In practice, the difference between RMSE and RSE is very small, particularly for large multivariate data.

**Mean Absolute Error (MAE),** like the RMSE, the MAE measures the prediction error. Mathematically, it is the average absolute difference between observed and predicted outcomes, MAE = mean(abs(observeds - predicteds)). MAE is less sensitive to outliers compared to RMSE.

**Where is Linear Regression Used?**

**Evaluating Trends and Sales Estimates** - *Linear regressions can be used in business to evaluate trends and make estimates or forecasts.*

If a company’s sales have increased steadily every month for the past few years, conducting a linear analysis on the sales data with monthly sales on the y-axis and time on the x-axis would produce a line that that depicts the upward trend in sales. After creating the trend line, the company could use the slope of the line to forecast sales in future months.

**Analysing the Impact of Price Changes** *- Linear regression can also be used to analyse the effect of pricing on consumer behaviour.*

If a company changes the price on a certain product several times, it can record the quantity it sells for each price level and then performs a linear regression with quantity sold as the dependent variable and price as the explanatory variable. The result would be a line that depicts the extent to which consumers reduce their consumption of the product as prices increase, which could help guide future pricing decisions.

**Assessing Risk** - *Linear regression can be used to analyse risk.*

A health insurance company might conduct a linear regression plotting number of claims per customer against age and discover that older customers tend to make more health insurance claims. The results of such an analysis might guide important business decisions made to account for risk.

**Advantages of using Linear Regression**

* Simple implementation and less complexity
* Performance on linearly separable datasets
* Overfitting can be reduced by regularization

**Disadvantages of using Linear Regression**

* Prone to underfitting
* Sensitive to outliers
* Linear Regression assumes that the data is independent

= = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =

**Polynomial Regression**

**Polynomial Regression** is a form of linear regression in which the relationship between the independent variable ‘x’ and dependent variable ‘y’ is modelled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of ‘x’ and the corresponding conditional mean of ‘y’, denoted E(y|x)

**Why Polynomial Regression**

* There are some relationships that a researcher will hypothesize is curvilinear. Clearly, such type of cases will include a polynomial term.
* Inspection of residuals. If we try to fit a linear model to curved data, a scatter plot of residuals (Y axis) on the predictor (X axis) will have patches of many positive residuals in the middle. Hence in such situation it is not appropriate.
* An assumption in usual multiple linear regression analysis is that all the independent variables are independent. In polynomial regression model, this assumption is not satisfied.

**Uses of Polynomial Regression**

These are basically used to define or describe non-linear phenomenon such as:

* Growth rate of tissues.
* Progression of disease epidemics
* Distribution of carbon isotopes in lake sediments

The basic goal of regression analysis is to model the expected value of a dependent variable ‘y’ in terms of the value of an independent variable ‘x’. In simple regression, we used following equation -

**y = mx + c + e**

Here ‘y’ is dependent variable, ‘c’ is y intercept, ‘m’ is the slope and ‘e’ is the error rate.

In many cases, this linear model will not work out for example if we analysing the production of chemical synthesis in terms of temperature at which the synthesis take place in such cases, we use quadratic model

**y = c + m1x + m2\*x^2 +m3\*X^3+ e**

In general, we can model it for nth value.

**y = c + m1x + m2\*x^2 +....+ mn\*x^n**

Since regression function is linear in terms of unknown variables, hence these models are linear from the point of estimation. Hence through Least Square technique, let’s compute the response value that is y.

**Advantages of using Polynomial Regression**

* Broad range of function can be fit under it.
* Polynomial basically fits wide range of curvature.
* Polynomial provides the best approximation of the relationship between dependent and independent variable.

**Disadvantages of using Polynomial Regression**

* These are too sensitive to the outliers.
* The presence of one or two outliers in the data can seriously affect the results of a nonlinear analysis.
* In addition, there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression

**Polynomial Regression for Non-Linear Data**

Non-linear data is usually encountered in daily life. Consider some of the equations of motion as studied in physics.

* **Projectile Motion**: The height of a projectile is calculated as h = -½ gt2 +ut +ho
* **Equation of motion under free fall**: The distance travelled by an object after falling freely under gravity for ‘t’ seconds is ½ g t2.
* **Distance travelled by a uniformly accelerated body**: The distance can be calculated as ut + ½gt2

where,

g = acceleration due to gravity

u = initial velocity

ho = initial height

a = acceleration

In addition to these examples, Non-linear trends are also observed in the growth rate of tissues, the progress of disease epidemic, black body radiation, the motion of the pendulum etc. These examples clearly indicate that we cannot always have a linear relationship between the independent and dependent attributes. Hence, linear regression is a poor choice for dealing with such nonlinear situations. This is where Polynomial Regression comes to our rescue.

Polynomial Regression is a powerful technique to encounter the situations where a quadratic, cubic or a higher degree nonlinear relationship exists. The underlying concept in polynomial regression is to add powers of each independent attribute as new attributes and then train a linear model on this expanded collection of features.

**Categorical variable:** Categorical variables contain a finite number of categories or distinct groups. Categorical data might not have a logical order. For example, categorical predictors include gender, material type, and payment method.

**Discrete variable**: Discrete variables are numeric variables that have a countable number of values between any two values. A discrete variable is always numeric. For example, the number of customer complaints or the number of flaws or defects.

**Continuous variable**: Continuous variables are numeric variables that have an infinite number of values between any two values. A continuous variable can be numeric or date/time. For example, the length of a part or the date and time a payment is received.