edureka!

MASTERING MACHINE LEARNING WITH EDUREKA

Getting Started with Machine Learning

Don't just Learn it, Master it!

TABLE OF CONTENTS

TABLE OF CONTENTS

Chapter 1

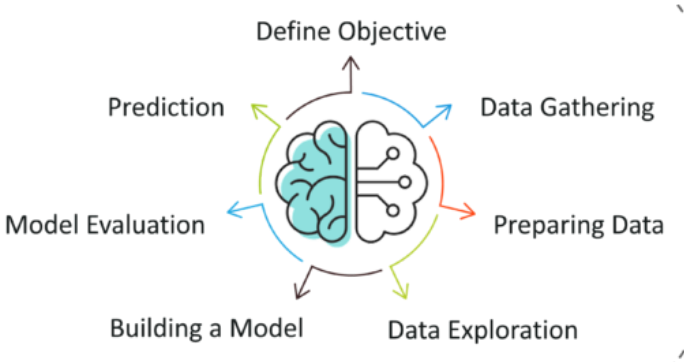
INTRODUCTION TO MACHINE LEARNING

Undoubtedly, Machine Learning is the most in-demand technology in today's market. Its applications range from self-driving cars to predicting deadly diseases such as ALS. The term Machine Learning was first coined by Arthur Samuel in the year 1959. If you browse through the net, searching for 'what is Machine Learning', you'll get at least 100 different definitions. However, the very first formal definition was:

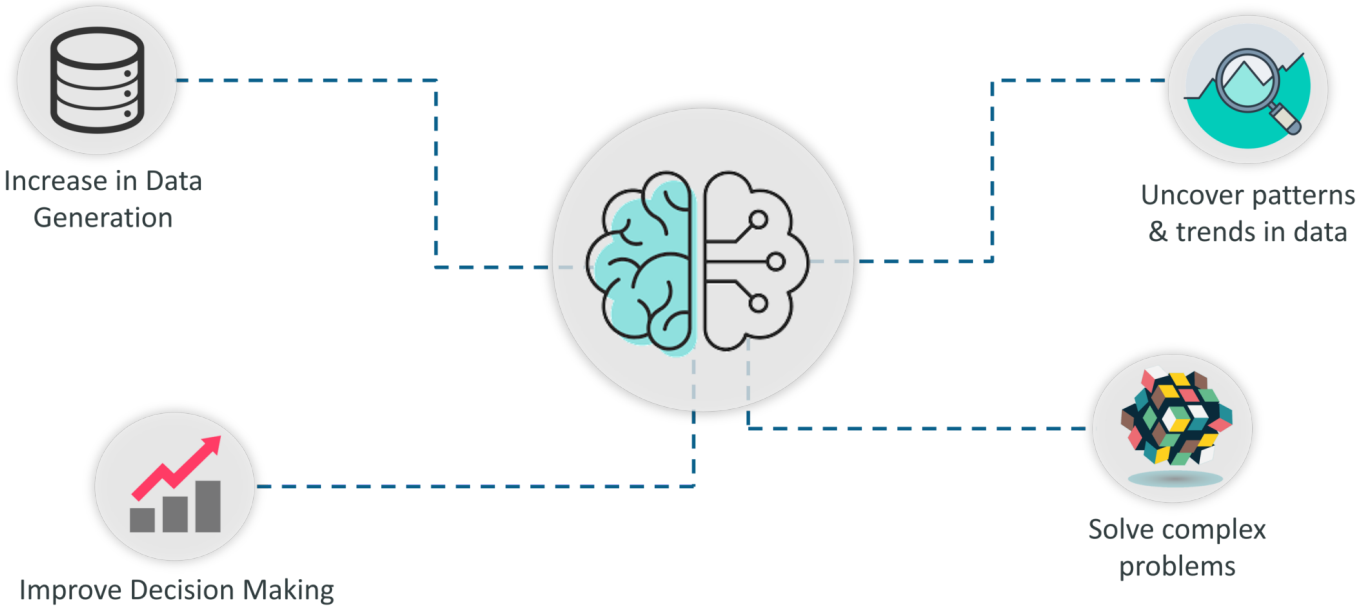
> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."
>
> — *Tom M. Mitchell*

1.1 What is Machine Learning?

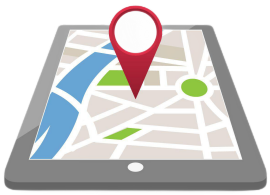
In simple terms, Machine Learning is a subset of Artificial Intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed to do so. In this sense, it is the practice of getting machines to solve problems by gaining the ability to think. The Machine Learning process involves building a Predictive model that can be used to find a solution for a problem statement. The shown image represents the steps in the Machine Learning process:



1.2 Need for Machine Learning



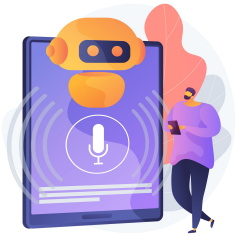
MACHINE LEARNING APPLICATIONS



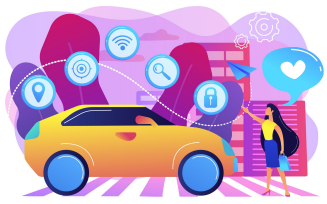
Traffic Alerts

Social Media

Product Recommendation



Virtual Personal Assistants



Self Driving Cars

Google Translate

Chapter 2

MACHINE LEARNING BASICS

Before we dive deeper, let's get familiar with some of the most commonly used terminologies in Machine Learning.

1. **ALGORITHM**
   A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model.

2. **MODEL**
   A model is the main component of Machine Learning and is trained by using an algorithm. It maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.

3. **PREDICTOR VARIABLE**
   It is a feature(s) of the data that can be used to predict the output.

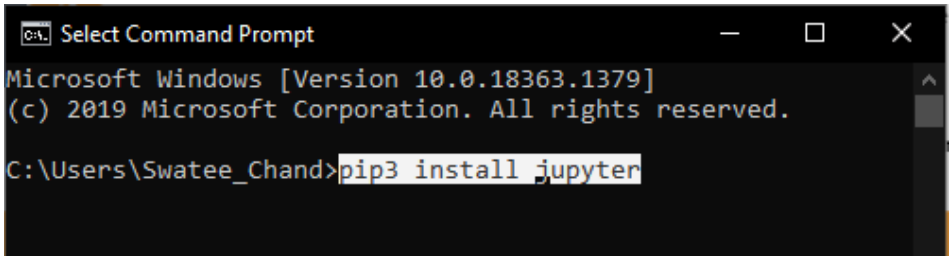
4. **RESPONSE VARIABLE**
   It is the feature or the output variable that needs to be predicted by using the predictor variable(s).

5. **TRAINING DATA**
   The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

6. **TESTING DATA**
   After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

Jupyter Notebook Installation

In order to implement the Machine Learning codes, you will be needing an IDE. Here we will be working with Jupyter. Installation of Jupyter notebook is very simple. All you need to do is type "**pip install jupyter**" on your command prompt.



Chapter 3

PYTHON FUNDAMENTALS FOR MACHINE LEARNING

Machine Learning has become the talk of the town as it yields the most promising careers for the future. We make the computer learn based on past experiences through the data stored or create algorithms that make the computer learn by itself. The programming language that mostly everyone chooses? It's Python. Why? It is because of the support for these domains with the in-built libraries such as Pandas, Scikit-Learn, NumPy and so many more. Here we will talk about a few of such libraries.

3.1 NumPy

NumPy is a Python package that stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object. Python NumPy arrays provide tools for integrating C, C++, etc. It is also useful in linear algebra, random number capability, etc. NumPy array can also be used as an efficient multi-dimensional container for generic data.

1-D NUMPY ARRAY

```
import numpy as np
a=np.array([1,2,3])
print(a)
```

OUTPUT

```
[1 2 3]
```

MULTI-D NUMPY ARRAY

```
import numpy as np
a=np.array([(1,2,3),(4,5,6)])
print(a)
```

OUTPUT

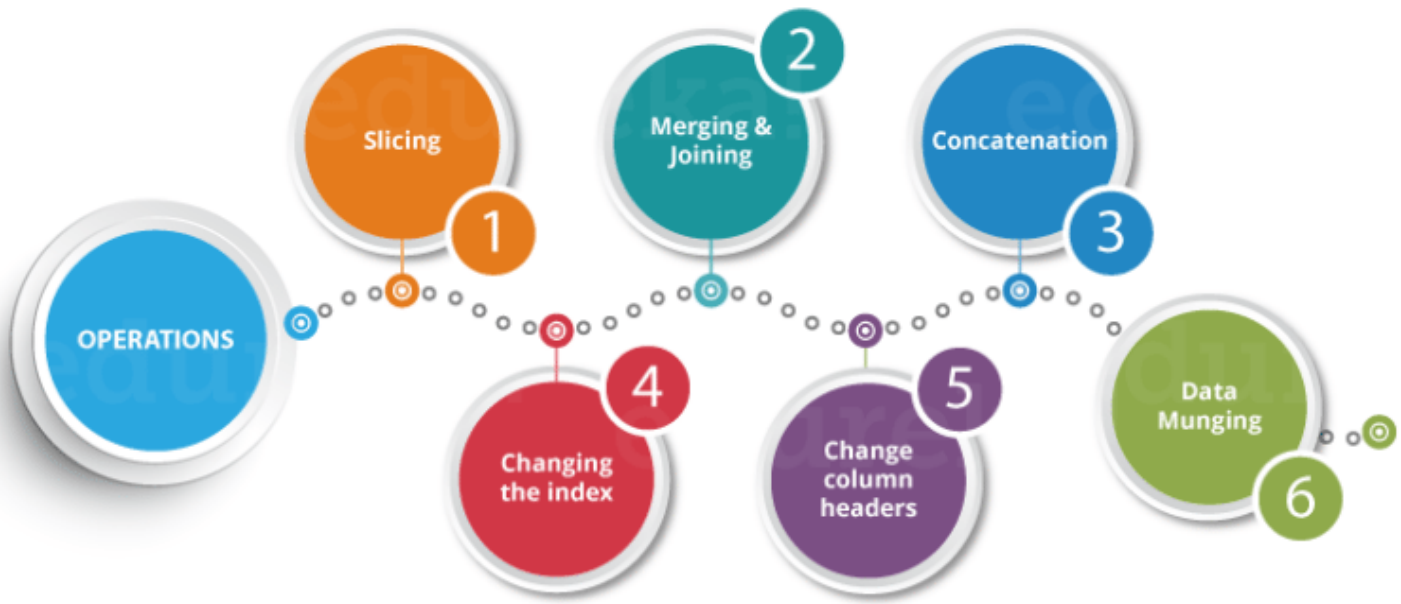
```
[[ 1 2 3]
[4 5 6]]
```

3.2 Pandas

Pandas is an open-source software library that is built on top of NumPy. It is used for data manipulation, analysis and cleaning. Python pandas is well suited for different kinds of data, such as:

1. Tabular data with heterogeneously-typed columns
2. Ordered and unordered time series data
3. Arbitrary matrix data with row & column labels
4. Unlabelled data
5. Any other form of observational or statistical data sets

PYTHON PANDAS OPERATIONS



3.3 Scikit Learn

Scikit learn is a library used to perform Machine Learning in Python. Scikit learn is an open-source library that is licensed under BSD and is reusable in various contexts, encouraging academic and commercial use. It provides a range of supervised and unsupervised learning algorithms in Python. It consists of popular algorithms and libraries. Apart from that, it also contains the following packages:

01 NumPy	02 Matplotlib	03 SciPy
NumPy is a Python package that stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object.	matplotlib.pyplot is a plotting library used for 2D graphics in Python programming language. It can be used in Python scripts, shell and web application servers.	SciPy is an open-source Python library used to solve scientific & mathematical problems. It is built on the NumPy extension and helps in data manipulation and visualization.

To implement Scikit learn, we first need to import the above packages. You can download these two packages using the command line or if you are using PyCharm, you can directly install it by going to your setting in the same way you do it for other packages.

Chapter 4

MACHINE LEARNING CLASSIFICATION

A machine can learn to solve a problem by following any three approaches covered in this chapter.

4.1 Supervised Learning

Supervised Learning is a technique in which we teach or train the machine using data that is well labeled. To understand Supervised Learning, let's consider an analogy. As kids we all needed guidance to solve math problems. Our teachers helped us understand what addition is, and how it is done. Similarly, you can think of Supervised Learning as a type of Machine Learning that involves a guide. The labeled data set is the teacher that will train you to understand patterns in the data. Here, the labeled data set is the training data set. Now with having a basic understanding of what Supervised Learning is, let's also understand what makes this kind of learning important.

IMPORTANCE OF SUPERVISED LEARNING

- Learning gives the algorithm experience which can be used to output the predictions for new unseen data
- Experience also helps in optimizing the performance of the algorithm
- Real-world computations can also be taken care of by the Supervised Learning algorithms

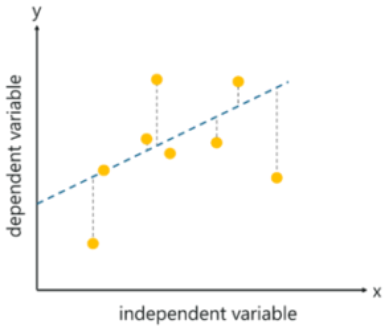
TYPES OF SUPERVISED LEARNING

4.1.1 Regression

Regression is the kind of Supervised Learning that learns from the Labeled Datasets and is then able to predict a continuous-valued output for the new data given to the algorithm. It is used whenever the output required is a number such as money or height etc. Some popular Supervised Learning algorithms are discussed below:

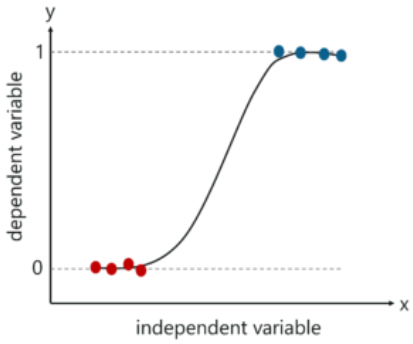
1 LINEAR REGRESSION

This algorithm assumes that there is a linear relationship between the 2 variables, Input (X) and Output (Y), of the data it has learnt from. The Input variable is called the Independent Variable and the Output variable is called the Dependent Variable. When unseen data is passed to the algorithm, it uses the function, calculates and maps the input to a continuous value for the output.



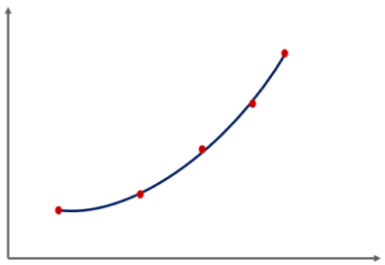
2 LOGISTIC REGRESSION

This algorithm predicts discrete values for the set of Independent variables that have been passed to it. It does the prediction by mapping the unseen data to the logit function that has been programmed into it. The algorithm predicts the probability of the new data and so it's output lies between the range of 0 and 1.



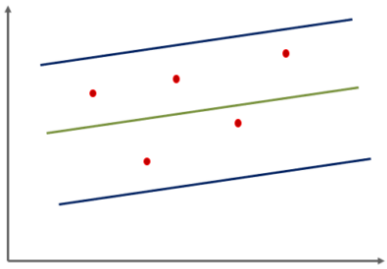
3 POLYNOMIAL REGRESSION

Polynomial Regression is a method used to handle non-linear data. Non-linearly separable data is basically when you cannot draw out a straight line to study the relationship between the dependent and independent variables.



4 SUPPORT VECTOR REGRESSION

For Support Vector Machine Regression or SVR, we identify a hyperplane with maximum margin such that the maximum number of data points are within those margins. It is quite similar to the support vector machine classification algorithm.



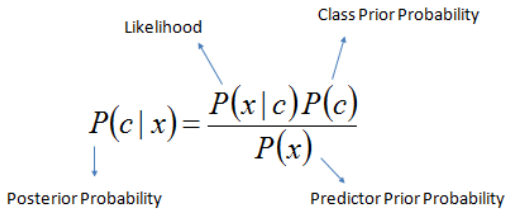
4.1.2 Classification

Classification, on the other hand, is the kind of learning where the algorithm needs to map the new data that is obtained to any one of the 2 classes that we have in our dataset. The classes need to be mapped to either 1 or 0, which in real-life translated as 'Yes' or 'No', 'Rains' or 'Does Not Rain' and so forth. The output will be either one of the classes and not a number as it was in Regression. Some of the most well-known algorithms are discussed below:

1 NAIVE BAYES CLASSIFIER

This is a classification technique based on an assumption of independence between predictors or what's known as Bayes' theorem. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Building a Bayesian model is simple and particularly functional in the case of enormous data sets. Along with simplicity, Naive Bayes is known to outperform sophisticated classification methods as well.

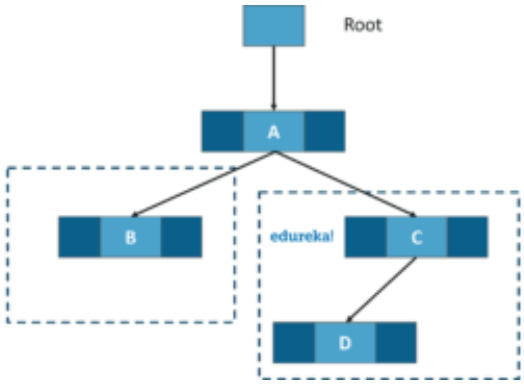
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x), and P(x|c). The expression for Posterior Probability is as follows.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

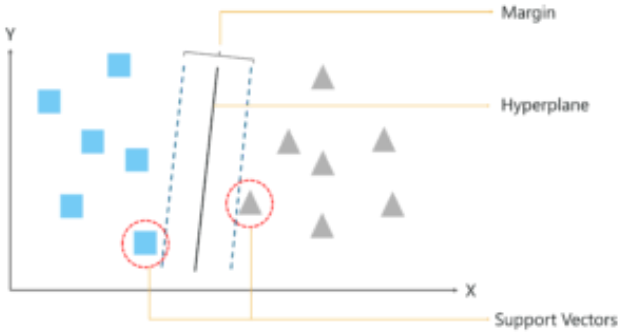
2 DECISION TREE

Decision Trees classify based on the feature values. They use the method of Information Gain and find out which feature of the dataset gives the best of information, make that as the root node and so on till they are able to classify each instance of the dataset. Every branch in the Decision Tree represents a feature of the dataset. They are one of the most widely used algorithms for classification.



3 SUPPORT VECTOR MACHINES (SVM)

It is a Supervised Learning Machine Learning classification algorithm that has become extremely popular nowadays owing to its extremely efficient results. Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible.



4.2 Unsupervised Learning

Unsupervised Learning can be thought of as self-learning where the algorithm can find previously unknown patterns in datasets that do not have any sort of labels. It helps in modeling probability density functions, finding anomalies in the data, and much more. To give you a simple example, think of a student who has textbooks and all the required material to study but has no teacher to guide. Ultimately, the student will have to learn by himself or herself to pass the exams. This sort of self-learning is what we have scaled into Unsupervised Learning for machines.

IMPORTANCE OF UNSUPERVISED LEARNING

- It works on datasets that are unlabeled and find patterns that would previously not be known to us.
- The patterns obtained are helpful if we need to categorize the elements or find an association between them.
- They can also help detect anomalies and defects in the data which can be taken care of by us.

TYPES OF UNSUPERVISED LEARNING

CLUSTERING

ASSOCIATION

4.2.1 Clustering

Clustering is the type of Unsupervised Learning where you find patterns in the data that you are working on. It may be the shape, size, color, etc. which can be used to group data items or create clusters. Some popular algorithms in Clustering are discussed below:

1

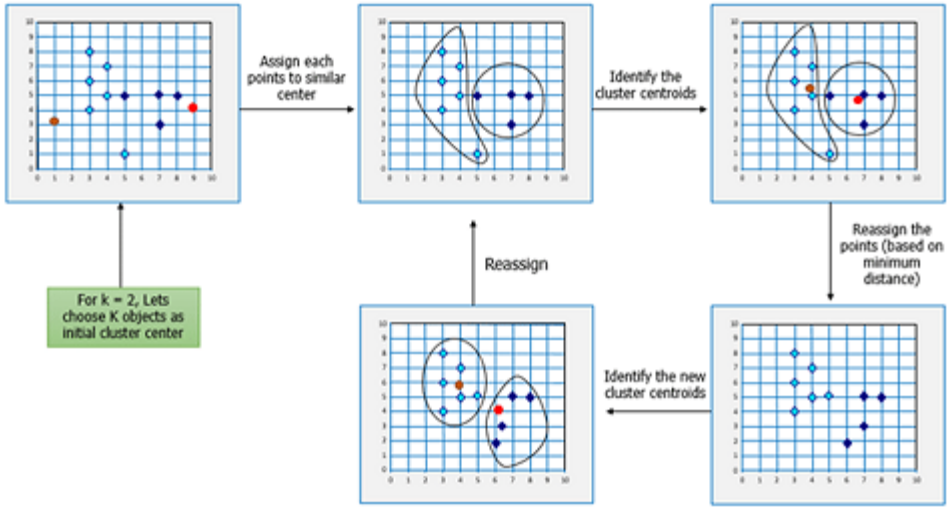
HIERARCHICAL CLUSTERING

This algorithm builds clusters based on the similarity between different data points in the dataset. It goes over the various features of the data points and looks for the similarity between them. If the data points are found to be similar, they are grouped together. This continues until the dataset has been grouped which creates a hierarchy for each of these clusters.

2

K-MEANS CLUSTERING

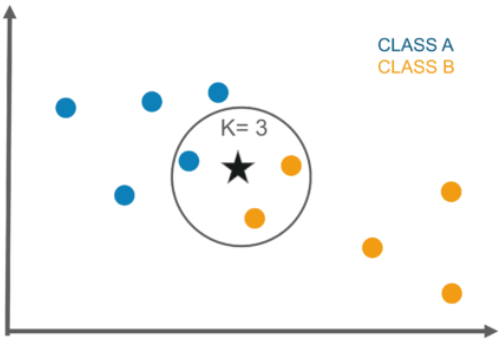
This algorithm works step-by-step where the main goal is to achieve clusters that have labels to identify them. The algorithm creates clusters of different data points which are as homogenous as possible by calculating the centroid of the cluster and making sure that the distance between this centroid and the new data point is as little as possible. The smallest distance between the data point and the centroid determines which cluster it belongs to while making sure the clusters do not interlay with each other. The centroid acts like the heart of the cluster. This ultimately gives us the cluster which can be labeled as needed.



3

K-NN CLUSTERING

This is probably the simplest of the Machine Learning algorithms, as the algorithm does not really learn but rather classifies the new data point based on the datasets that have been stored by it. This algorithm is also called a lazy learner because it learns only when the algorithm is given a new data point. It works well with smaller datasets as huge datasets take time to learn.



4.2.2 Association

Association is the kind of Unsupervised Learning where you find the dependencies of one data item to another data item and map them such that they help you profit better. Some popular algorithms in Association Rule Mining are discussed further:

1

APRIORI ALGORITHM

The Apriori Algorithm is a breadth-first search based which calculates the support between items. This support basically maps the dependency of one data item with another which can help us understand what data item influences the possibility of something happening to the other data item. For example, bread influences the buyer to buy milk and eggs. So that mapping helps increase profits for the store. That sort of mapping can be learnt using this algorithm which yields rules as for its output.

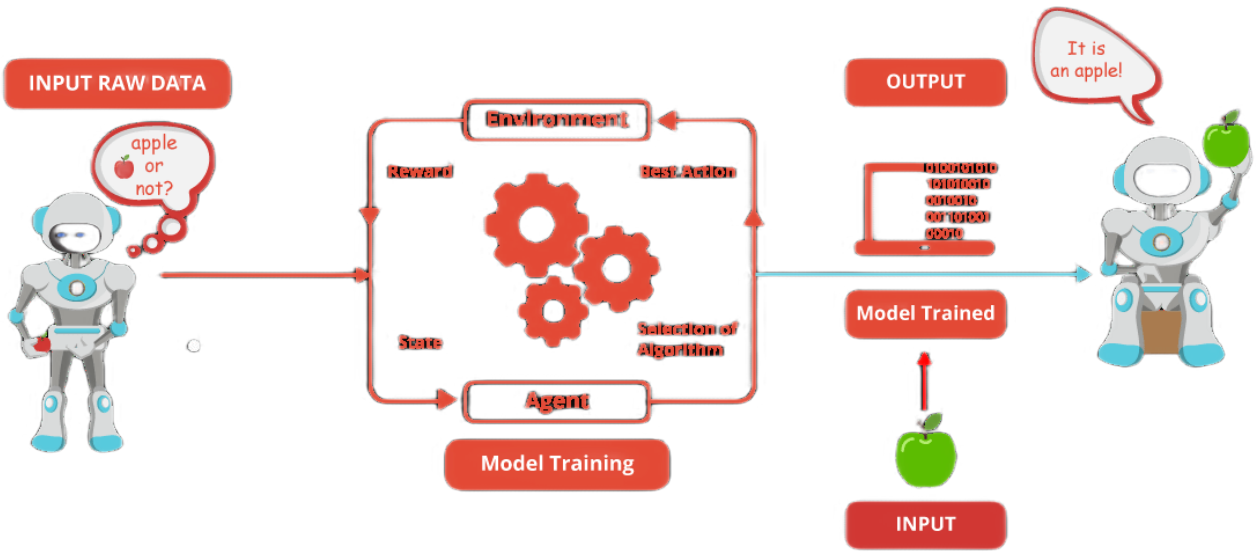
2

FP-GROWTH ALGORITHM

The Frequency Pattern (FP) algorithm finds the count of the pattern that has been repeated, adds that to a table and then finds the most plausible item and sets that as the root of the tree. Other data items are then added into the tree and the support is calculated. If that particular branch fails to meet the threshold of the support, it is pruned. Once all the iterations are completed, a tree with the root to the item will be created which are then used to make the rules of the association. This algorithm is faster than Apriori as the support is calculated and checked for increasing iterations rather than creating a rule and checking the support from the dataset.

4.3 Reinforcement Learning

Reinforcement Learning is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. And again once trained it gets ready to predict the new data presented to it.



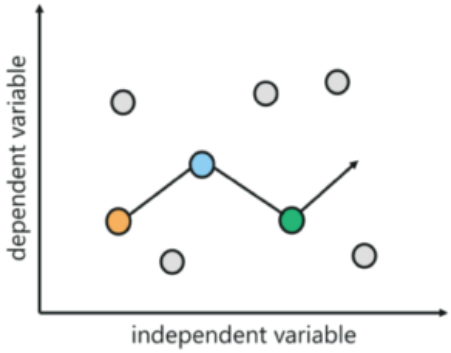
Chapter 5

ADVANCED MACHINE LEARNING CONCEPTS

This chapter will introduce you to some of the advanced concepts of Machine Learning.

5.1 Regularization in Machine Learning

In general, Regularization means to make things regular or acceptable. This is exactly why we use it for applied Machine Learning. In the context of Machine Learning, Regularization is the process that regularizes or shrinks the coefficients towards zero. In simple words, Regularization discourages learning a more complex or flexible model, to prevent overfitting.



REGULARIZATION TECHNIQUES

01 Ridge Regression	02 Lasso Regression
This Regularization technique performs L2 regularization. It modifies the RSS by adding the penalty (shrinkage quantity) equivalent to the square of the magnitude of coefficients.	This Regularization technique performs L1 regularization. It modifies the RSS by adding the penalty (shrinkage quantity) equivalent to the sum of the absolute value of coefficients.

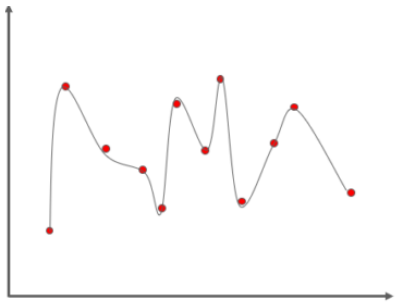
5.2 Overfitting and Underfitting

Building a Machine Learning model is not just about feeding the data, there are a lot of deficiencies that affect the accuracy of any model. Overfitting and Underfitting in Machine Learning are such deficiencies that hinder the accuracy as well as the performance of the model.

1

OVERFITTING IN MACHINE LEARNING

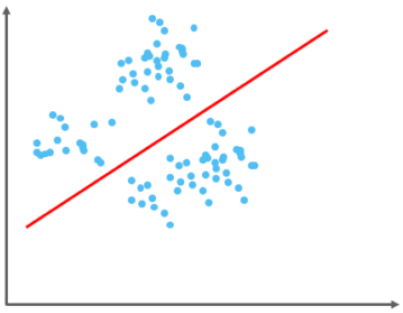
A statistical model is said to be overfitted when we feed it a lot more data than necessary. To make it relatable, imagine trying to fit into oversized apparel. When a model fits more data than it actually needs, it starts catching the noisy data and inaccurate values in the data. As a result, the efficiency and accuracy of the model decreases.



2

UNDERFITTING IN MACHINE LEARNING

In order to avoid overfitting, we could stop the training at an earlier stage. But it might also lead to the model not being able to learn enough from training data, that it may find it difficult to capture the dominant trend. This is known as underfitting. The result is the same as overfitting, inefficiency in predicting outcomes.



5.3 Dimensionality Reduction Technique

Machine Learning in general works wonders when the dataset provided for training the machine is large and concise. Usually having a good amount of data helps us build a better predictive model since we have more data to train the machine with. However, using a large data set has its pitfalls. The biggest pitfall is the curse of dimensionality. It turns out that in large dimensional datasets, there might be lots of inconsistencies in the features or lots of redundant features. This increases the computation time and makes data processing and EDA more convoluted. To get rid of the curse of dimensionality, a process called dimensionality reduction was introduced. Dimensionality reduction techniques can be used to filter only a limited number of significant features needed for training and this is where PCA comes in.

5.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that enables you to identify correlations and patterns in a data set. It helps in transforming it into a data set of significantly lower dimensions without loss of any important information. The main idea behind PCA is to figure out patterns and correlations among various features in the data set. On finding a strong correlation between different variables, a final decision is made about reducing the dimensions of the data in such a way that the significant data is still retained. Such a process is very essential in solving complex data-driven problems that involve the use of high-dimensional data sets.

Chapter 6

PRACTICE DATASETS FOR MACHINE LEARNING

Datasets are an integral part of Machine Learning and NLP (Natural Language Processing). Without training datasets, Machine Learning algorithms would not have a way to learn text mining, text classification, or how to categorize products. This chapter will provide you a list of free datasets which you can use to practice Machine Learning.

DATASETS FOR GENERAL MACHINE LEARNING

In this context, "general" is referred to as Regression, Classification, and Clustering with relational data.

WINE QUALITY
Properties of red and white Vinho Verde wine samples from the north of Portugal. The goal here is to model wine quality based on some physicochemical tests.

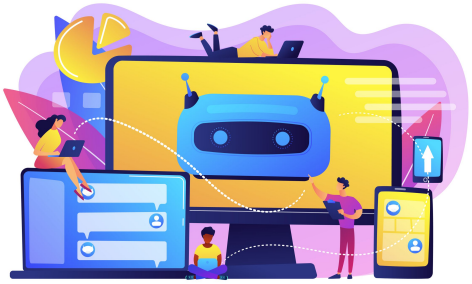
CREDIT CARD DEFAULT
Predicting credit card default is a valuable use for machine learning. This dataset includes payment history, demographics, credit, and default data.

DATASETS FOR NATURAL LANGUAGE PROCESSING

NLP is all about text data. And for data like text, the datasets need to have real-world applications so that sanity checks can be performed easily.

AMAZON REVIEWS
It contains approximately 35 million reviews from Amazon spanning 18 years. Data includes user information, product information, ratings, and text review.

NEWS GROUP CLASSIFICATION
Collection of almost 20,000 newsgroup documents, partitioned evenly across 20 newsgroups. It is great for practicing topic modeling and text classification.



FINANCE & ECONOMICS DATASETS FOR MACHINE LEARNING

Financial quantitative records are kept for decades, hence this industry is perfectly suited for machine learning.

QUANDL
A great source of economic and financial data that is useful to bui models to predict stock prices or economic indicators.

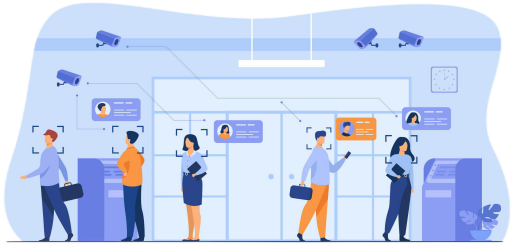
IMF DATA
The International Monetary Fund (IMF) publishes data on international finances, foreign exchange reserves, debt rates, commodity prices, and investments.

IMAGE DATASETS FOR COMPUTER VISION

Image datasets are useful to train a wide range of computer vision applications, like medical imaging technology, face recognition, and autonomous vehicles.

IMAGENET
This de-facto image dataset for new algorithms is organized according to the WordNet hierarchy, where each node is depicted by hundreds and thousands of images.

GOOGLE'S OPEN IMAGES
A collection of around 9 million URLs to images annotated with labels spanning over 6,000 categories under Creative Commons.

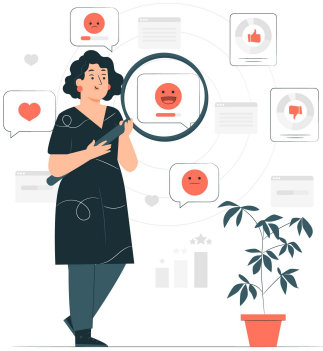


SENTIMENT ANALYSIS DATASETS FOR MACHINE LEARNING

Sentiment Analysis can be defined as a systematic analysis of online expressions.

IMDB REVIEWS
Dataset for binary sentiment classification. It features 25,000 movie reviews.

SENTIMENT140
Uses 160,000 tweets with emoticons pre-removed.



Chapter 7

FREQUENTLY ASKED INTERVIEW QUESTIONS

Machine Learning is a buzzword in the technology world right now and so is the need for Machine Learning Professionals is high in demand and this surge is due to evolving technology and the generation of huge amounts of data aka Big Data. This chapter covers the questions that will help you in your Machine Learning Interviews and open up various career opportunities as a **Machine Learning aspirant**.

1. What are the different types of Machine Learning?
2. How would you explain Machine Learning to a school-going kid?
3. How does Deep Learning differ from Machine Learning?
4. Explain Classification and Regression.
5. What do you understand by Selection Bias?
6. What do you understand by Precision and Recall?
7. Explain false negative, false positive, true negative and true positive with a simple example.
8. What is a Confusion Matrix?
9. What is the difference between Inductive and Deductive learning?
10. How is KNN different from K-means clustering?
11. What is ROC curve and what does it represent?
12. What's the difference between Type I and Type II error?
13. Is it better to have too many false positives or too many false negatives? Explain.
14. Which is more important to you – Model Accuracy or Model Performance?
15. What is the difference between Gini Impurity and Entropy in a Decision Tree?
16. What is the difference between Entropy and Information Gain?
17. What is Overfitting? And how do you ensure you're not overfitting with a model?
18. Explain Ensemble learning technique in Machine Learning.
19. What is bagging and boosting in Machine Learning?
20. How would you screen for outliers and what should you do if you find one?
21. What are Collinearity and Multicollinearity?
22. What do you understand by Eigenvectors and Eigenvalues?
23. What is A/B Testing?
24. What is Cluster Sampling?
25. Running a binary classification tree algorithm is quite easy. But do you know how the tree decides on which variable to split at the root node and its succeeding child nodes?
26. Name a few libraries in Python used for Data Analysis and Scientific Computations.
27. Which library would you prefer for plotting in Python language: Seaborn or Matplotlib or Bokeh?
28. How are NumPy and SciPy related?

CAREER GUIDANCE

WHO IS A MACHINE LEARNING PROFESSIONAL?
Machine Learning Professional is specialized in developing Machine Learning algorithms that can "learn" from or adapt to the data and make predictions.

Machine Learning Engineers

Machine Learning Engineers work in close collaboration with Data Scientists. While Data Scientists extract meaningful insights from large datasets and communicate the information to business stakeholders, Machine Learning Engineers ensure that the models used by Data Scientists can ingest vast amounts of real-time data for generating more accurate results.

Computer Vision Engineer

As a **Computer Vision Engineer**, you use software to handle the processing and analysis of large data populations, and your efforts support the automation of predictive decision-making efforts

Software Developer/Engineer

Software Developers/Engineers with specialization in AI/ML are the creative minds behind intelligent computer programs. Their main job is to develop efficient ML algorithms and applications.

Robotics Engineers

Robotics engineers are responsible for designing, testing, and building robots that are productive and safe to operate as well as economical to purchase and maintain.

Human-Centered Machine Learning Designer

Machine Learning has an exclusive branch that is dedicated to designing ML algorithms centered around humans. Hence, the name Human-Centered Machine Learning. **Human-Centered Machine Learning Designers** are responsible for creating intelligent systems that can "learn" the preferences and behavior patterns of individual humans through information processing and pattern recognition.

NLP Scientist

NLP Scientists are primarily responsible for designing and developing machines and applications that can learn the patterns of speech of a human language and also translate spoken words into other languages.

EDUREKA MACHINE LEARNING TRAINING

MACHINE LEARNING MASTERS PROGRAM

Weekend/Weekday | Live Class/SP | 24 x 7 Technical Assistance

www.edureka.co/masters-program/machine-learning-engineer-training

PGD IN AI AND MACHINE LEARNING

Weekend | Live Class | E&ICT Academy NITW | Edureka

www.edureka.co/executive-programs/machine-learning-and-ai

MACHINE LEARNING TRAINING

Weekend | Live Class | 24 x 7 Technical Assistance

www.edureka.co/machine-learning-certification-training

PYTHON PROGRAMMING TRAINING

Weekend | Live Class | 24 x 7 Technical Assistance

www.edureka.co/python-programming-certification-training

LEARNER'S REVIEWS

SS Reddy

★★★★★

Edureka's PGP helped me get on the right path. Live lectures by experts working in the field helped us understand real time application areas and scopes while learning. Overall, I am happy that Edureka was part of my learning journey.

Yogendra SR

★★★★★

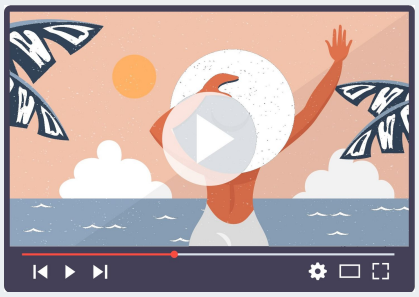
Awesome the way of teaching and support. And best part is Edureka most focus on practice, practice makes a man perfect.

Ankit Sharma

★★★★★

Its very simple in learning and interesting too. The way our instructor is teaching us is simply awesome. The thing which I like the most about Edureka is its Support service,as I have got all my queries answered by them on time. Thank you Edureka :)

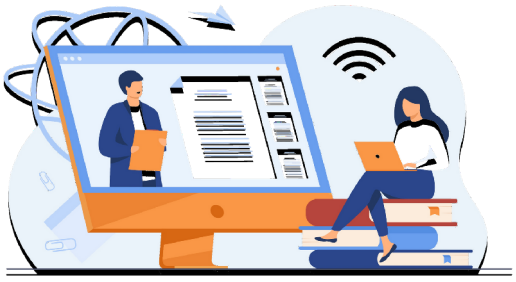
Free Resources



3000+
Video Tutorials on
YouTube

2500+ Technical
Blogs

Active
Community



30+
Free Monthly
Webinars

About Us

There are countless online education marketplaces on the internet. And there's us. We are not the biggest. We are not the cheapest. But we are the fastest growing. We have the highest course completion rate in the industry. We aim to become the largest online learning ecosystem for continuing education, in partnership with corporates and academia. To achieve that we remain ridiculously committed to our students. Be it constant reminders, relentless masters or 24 x 7 online technical support - we will absolutely make sure that you run out of excuses to not complete the course.

Contact Us

IndiQube ETA, 3rd Floor,
No.38/4,
Adjacent to Dell EMC2,
Dodanekundi,
Outer Ring Road, Bengaluru,
Karnataka - 560048

IN: 08035068112 | US: +1415 697 0520
www.instagram.com/edureka.co/
www.facebook.com/edurekaIN
www.linkedin.com/company/edureka/
www.youtube.com/user/edurekaIN
t.me/s/edurekaupdates
twitter.com/edurekaIN
in.pinterest.com/edurekaco/

News & Media

INDIA TODAY

Edureka partners with **NIT Warangal** to upskill IT professionals in **AI and Machine Learning**

Deloitte.

Edureka (Brain4ce Education Solutions) tops Deloitte Tech Fast 50 2014 rankings

edureka!