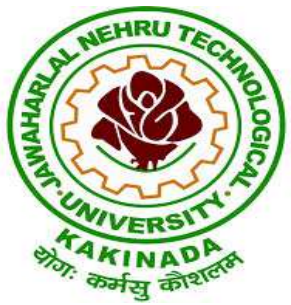
Basics of Machine Learning

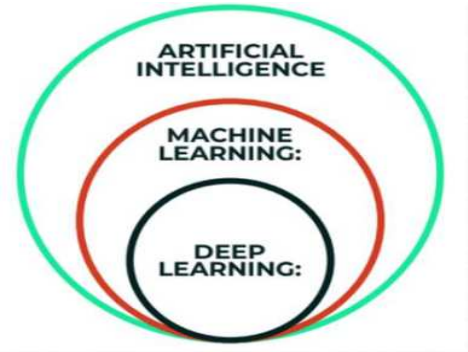
B.T.Krishna
E.C.E.Department
Jawaharlal Nehru Technological University Kakinada
Kakinada, Andhrapradesh,India–533003
Email:tkbattula@gmail.com



Machine Learning I

- Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.
- Machine learning is an subfield of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The term machine learning was first introduced by Arthur Samuel in 1959 for playing checkers on computer.
- The Machine Learning process starts with inputting training data into the selected algorithm. Training data being known or unknown data to develop the final Machine Learning algorithm.
- To test whether this algorithm works correctly, new input data is fed into the Machine Learning algorithm. The prediction and results are then checked.
- If the prediction is not as expected, the algorithm is retrained multiple numbers of times until the desired output is found.
- This enables the Machine Learning algorithm to continually learn on its own and produce the most optimal answer that will gradually increase in accuracy over time.

Machine Learning II

- 

- Artificial Intelligence (AI) is an umbrella discipline that covers everything related to making machines smarter.
- Machine Learning (ML) is commonly used along with AI but it is a subset of AI. ML refers to an AI system that can self-learn based on the algorithm.
- Deep Learning (DL) is a machine learning applied to large data sets.

Labelled and Unlabelled Data

- Data is simply information. It is in generally represented with a table. Each row is a data point and have certain features.

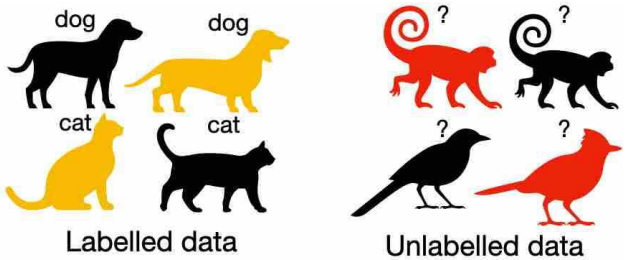


Figure: Labelled and Unlabelled Data

- Labeled data contains meaningful tags and is used in supervised learning, while unlabeled data doesn't contain additional information and is used in unsupervised learning.
- Labeled data requires the additional process of labeling, while unlabeled data is essentially raw data before labeling.
- Labeled data is harder to obtain (there are less datasets available, or you have to label it yourself), whereas unlabeled data is more abundant.

Traditional Programming vs Machine Learning

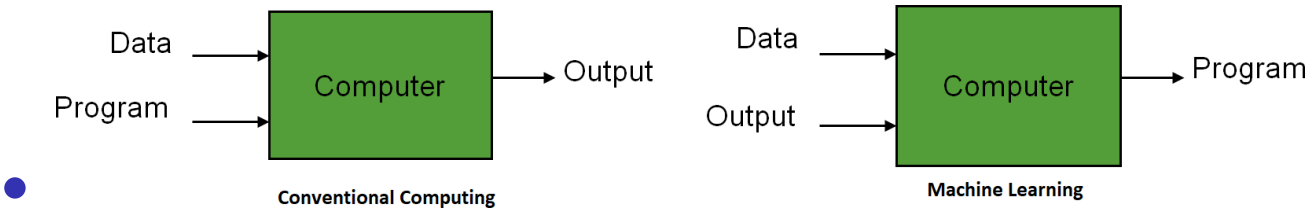


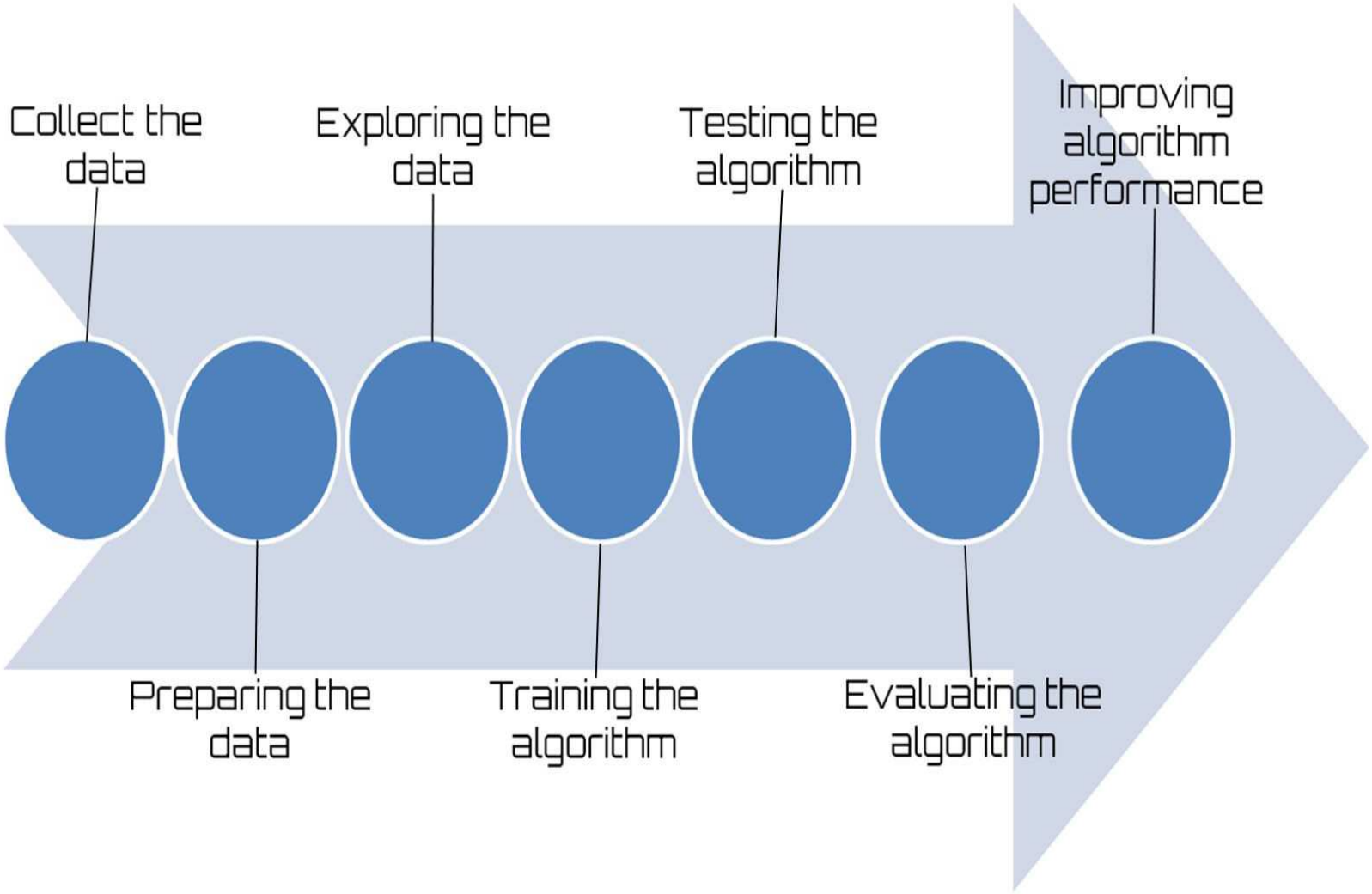
Figure: Difference Between Conventional Computing And Machine Learning

- Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.
- The machine learning framework will be,

$$y = f(x) \tag{1}$$

- where y is **Output**, f is **Prediction function** and x is a **feature**
  - **Training:** given a training set of labeled examples $(x_1, y_1), \ldots, (x_N, y_N)$, estimate the prediction function f by minimizing the prediction error on the training set
  - **Testing:** apply f to a never before seen test example x and output the predicted value $y = f(x)$

Machine Learning Workflow



Machine Learning Workflow I

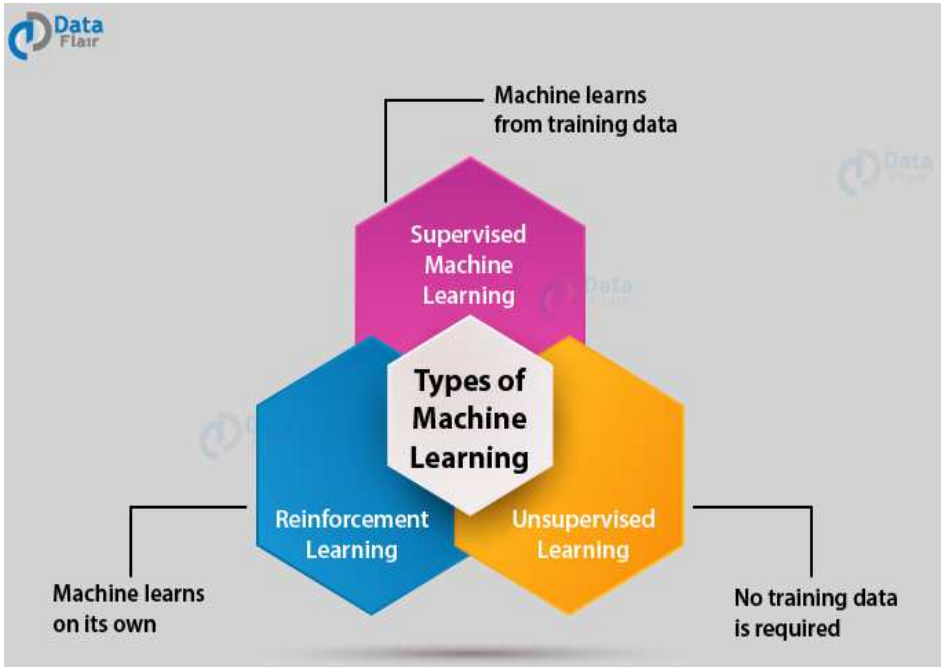
1. **Collect the data:** Data Gathering is the first step of the machine learning life cycle.In practice, it is collected through lengthy procedures that may, for example, derive from measurement campaigns or face-to-face interviews.
2. **Preparing the data:**After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
3. **Exploring the data:**It is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.
4. **Training the algorithm:** In this step, the machine learning begins to work with the definition of the model and the next training. The model starts to extract knowledge from large amounts of data that we had available, and that nothing has been explained so far. For unsupervised learning, there's no training step because you don't have a target value.

Machine Learning Workflow II

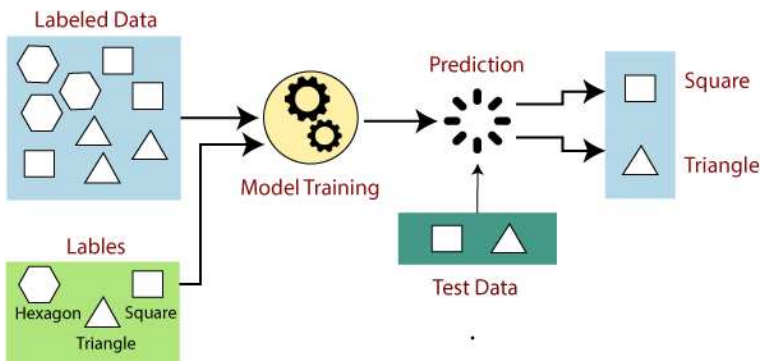
5. **Testing the algorithm:** Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.
6. **Evaluating the algorithm:** We have reached the point where we can apply what has been done so far. We can assess the approximation ability of the model by applying it to real data. The model, preventively trained and tested, is then valued in this phase.
7. **Improving algorithm performance:** Finally we can focus on the finishing steps. We've verified that the model works, we have evaluated the performance, and now we are ready to analyze the whole process to identify any possible room for improvement.

Types of Machine Learning

- The types of machine learning is,
  1. Supervised Learning
  2. Unsupervised Learning
  3. Reinforcement Learning



Supervised Learning I



- In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.
- Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

Supervised Learning II

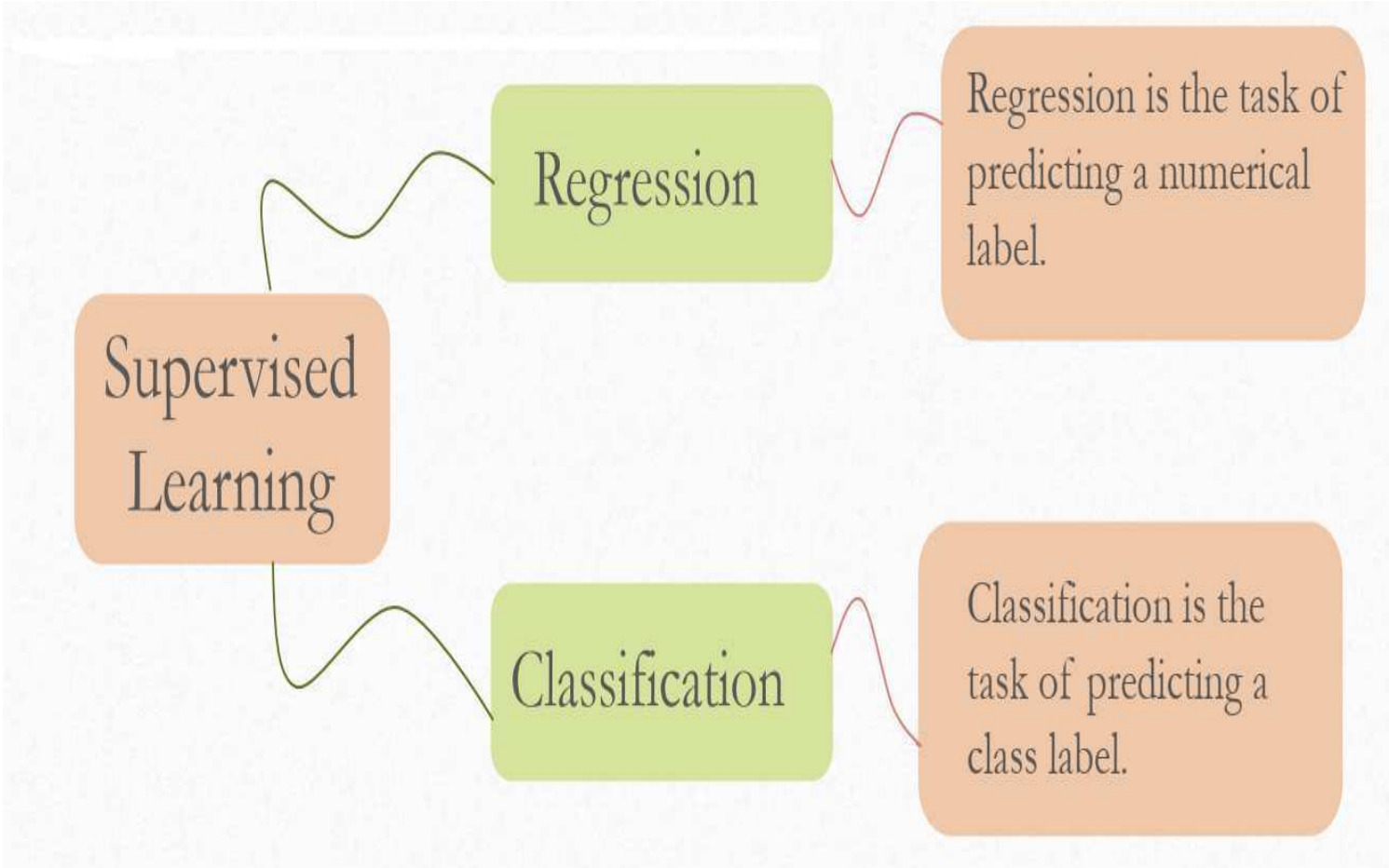


Figure: Categories of Supervised Learning

Supervised Learning

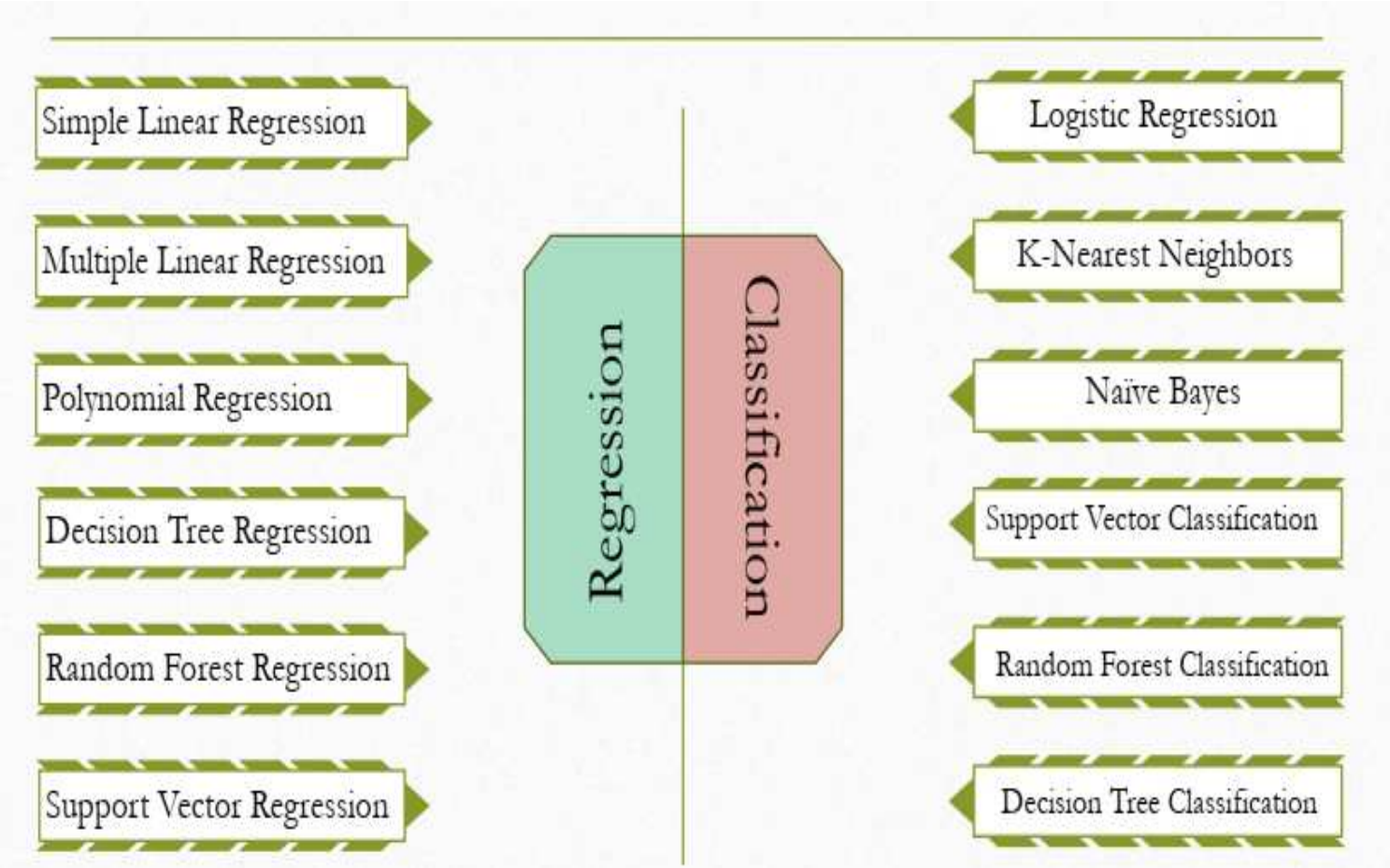


Figure: Supervised Learning Algorithms

Regression

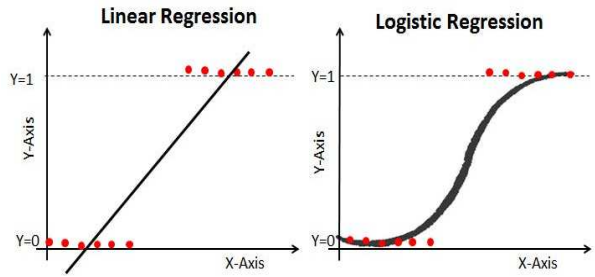
1. Regression is a form of predictive modelling technique which estimates the relationship between a dependent (target) and independent variable(s) (predictor).
2. Thus, it helps market researchers and data analysts to eliminate the worst and evaluate the best set of variables for building effective predictive models
3. There are different types of regression including simple linear regression, multiple linear regression, non-linear regression, etc.
4. The difference between linear and non-linear regression is that the plot of the model gives a curve in a non-linear regression and in linear regression, it gives a line.
5. The difference between simple linear regression and multiple linear regression is that multiple linear regression has more than 1 independent variables, whereas simple linear regression has only 1 independent variable.
6. 

Figure: Regression Techniques

Classification I

- In classification, the goal is to predict a categorical variable
- Naïve Bayes algorithm
  1. Naïve Bayes algorithm is quite similar to the logistic regression algorithm
  2. Naïve Bayes algorithm can be applied using GaussianNB() function from the sub package namely sklearn.naive-bayes.
- k-nearest neighbors algorithm
  1. k-nearest neighbors algorithm is a non-parametric method used for both classification and regression problems
  2. When k-NN is used for regression problems, the prediction is based on the mean or the median of the K-most similar instances.
  3. When k-NN is used for classification, the output is based on the mode. If we are using k and we have an even number of classes, k should be assigned an odd number to avoid a tie and vice versa.
  4. The k-NN gives better results if same scale is used for all the data. The k-NN works well with a small number of input variables, but struggles when the number of inputs is large.
  5. The most important parameter of the k-NN algorithm is k, which specifies the number of neighbor observations that contribute to the output predictions.

Classification II

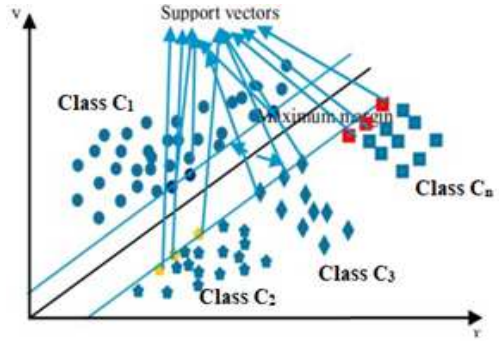
- Support Vector Machines
  1. Support vector machines are powerful models and perform well on a variety of datasets.
  2. SVM require careful pre-processing of data and requires tuning of the data and are hard to inspect. But this model is used, if all the features represent in similar units.
  3. SVM technique is easy to understand and is generally useful for unknown distribution data and non-regular data.
  4. In this algorithm, each data item is plotted as a point in n-dimensional space, where n is number of attributes in the dataset.
  5. 

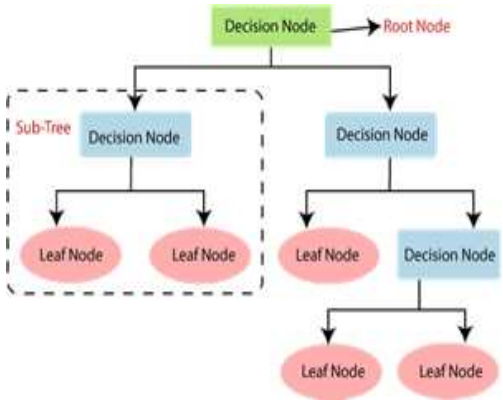
Figure: Support Vector Machine Classifier Algorithm

Classification III

- Decision Tree
  1. Decision Tree is basically a graph that represents choices and their outcomes in form of a tree structure.
  2. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions.
  3. Decision trees are typically drawn upside down such that terminal node (leaves) is at the bottom and root node is at the tops. Root node represents entire population or sample and this further gets divided into two or more homogeneous sets.
  4. Splitting is a process of dividing a node into two or more subnodes. When a subnode splits into further subnodes, then it is called decision node. A node, which is divided into subnodes is called parent node of subnodes; subnodes are the child of parent node.
  5. The decision of making strategic splits heavily affects a trees accuracy. The creation of subnodes increases the homogeneity of resultant subnodes. It then selects the split that results in most homogeneous subnodes.
  6. Decision tree is preferred in cases when there is a high non linearity and complex relationship between dependent and independent variables.

Classification IV

7. Decision tree should not be adopted while doing prediction for dependent continuous variable



8.

Figure: Decision Tree Classifier Algorithm

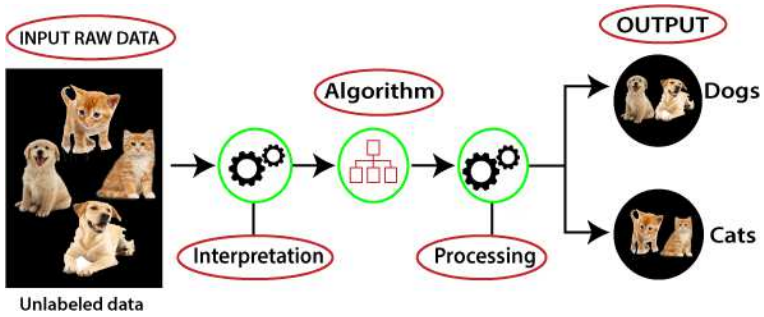
Confusion Matrix I

- A confusion matrix is a matrix (table) that can be used to measure the performance of an machine learning algorithm, usually a supervised learning one.
- Each row of the confusion matrix represents the instances of an actual class and each column represents the instances of a predicted class.
- The name confusion matrix reflects the fact that it makes it easy for us to see what kind of confusions occur in our classification algorithms.

Confusion Matrix		Predicted classes	
		male	female
Actual classes	male	42	8
	female	18	32

- This means that the classifier correctly predicted a male person in 42 cases and it wrongly predicted 8 male instances as female. It correctly predicted 32 instances as female. 18 cases had been wrongly predicted as male instead of female.

Unsupervised Learning I

- 

- Unsupervised learning describes a class of problems that involves using a model to describe or extract relationships in data.
- Compared to supervised learning, unsupervised learning operates upon only the input data without outputs or target variables. As such, unsupervised learning does not have a teacher correcting the model, as in the case of supervised learning.

Unsupervised Learning II

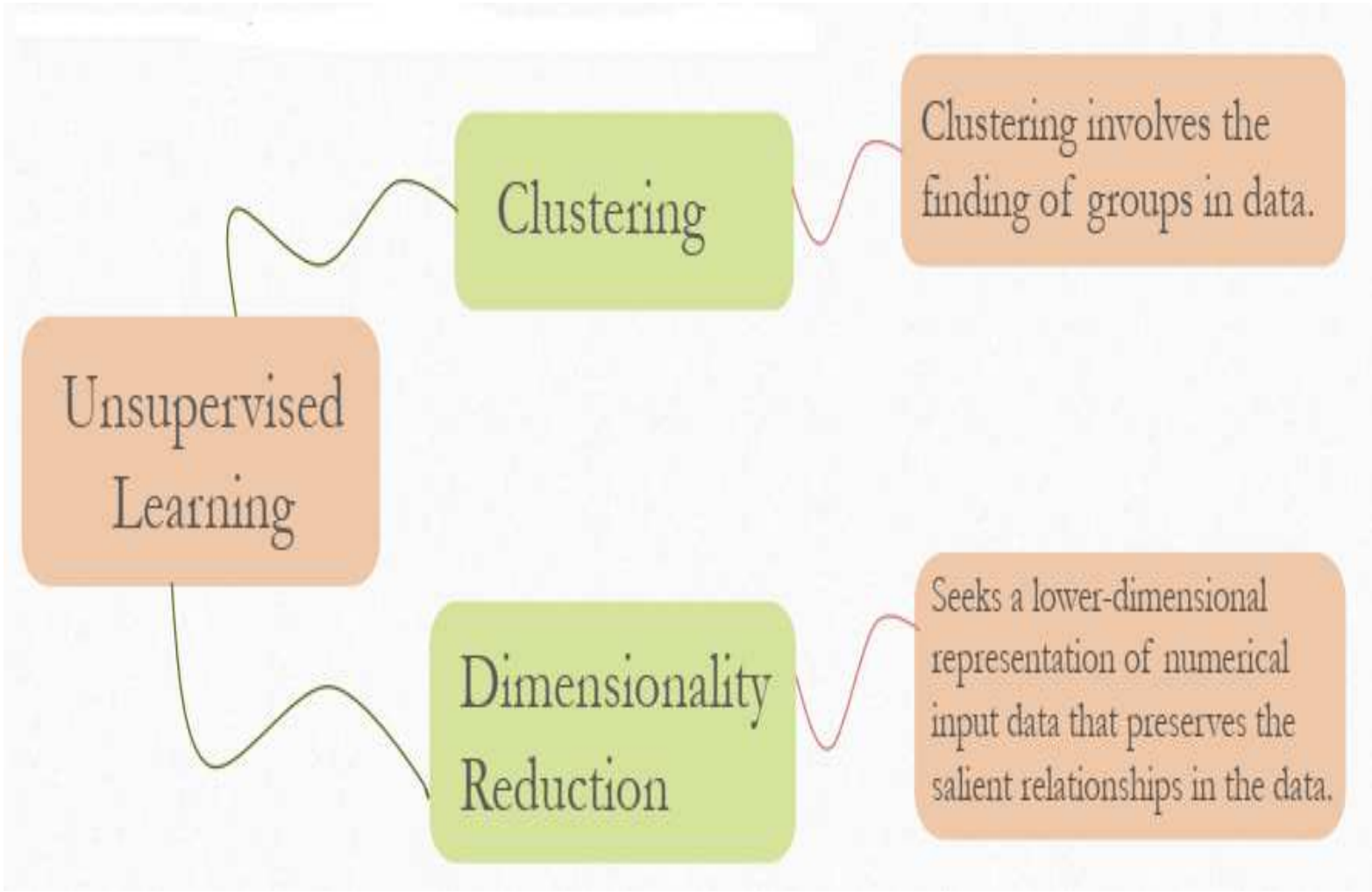
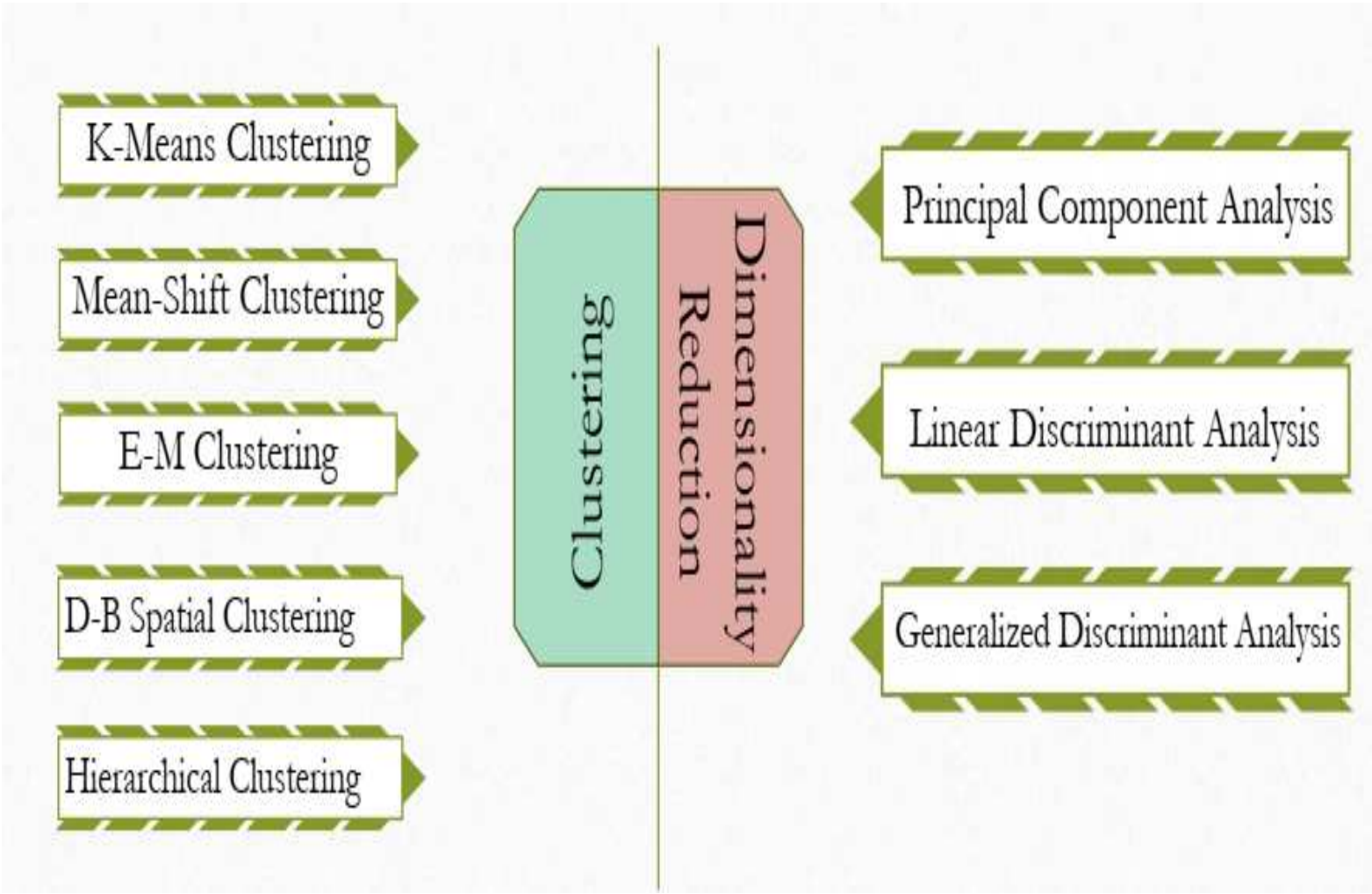


Figure: Categories of Unsupervised Learning

Unsupervised Learning



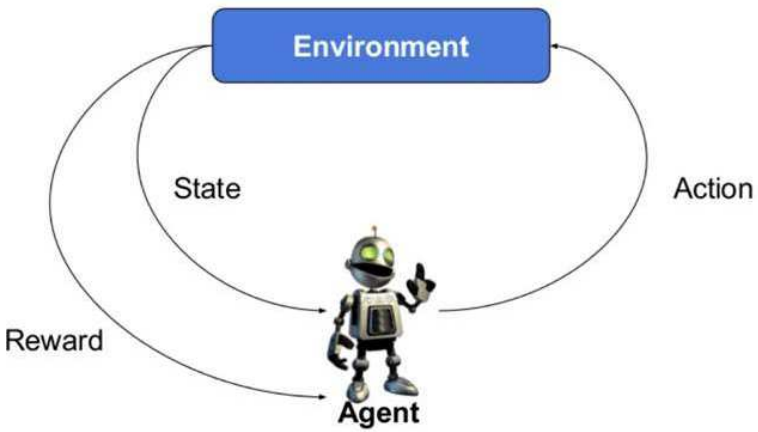
-

Figure: Unsupervised Learning Algorithms

Unsupervised Machine Learning Algorithms

- The two common **dimensionality reduction algorithms** include principle component analysis and factor analysis.
- These algorithms takes input of a high dimensional representation of the data, which consists of many features and produces a output that summarizes the data by grouping essential characteristics and results into fewer factors.
- Factor Analysis (FA) is an exploratory data analysis method used to search important underlying factors or latent variables from a set of observed variables. Factor analysis is a linear statistical model.
- **Clustering** is the process of organizing objects into groups whose members are similar in some manner; it deals with finding a structure in a collection of unlabeled data.
- A cluster is a collection of objects that have similar characteristics between them and are dissimilar to the objects belonging to other clusters.
- There are two forms of clustering: k-means and hierarchical clustering

Reinforcement Learning I

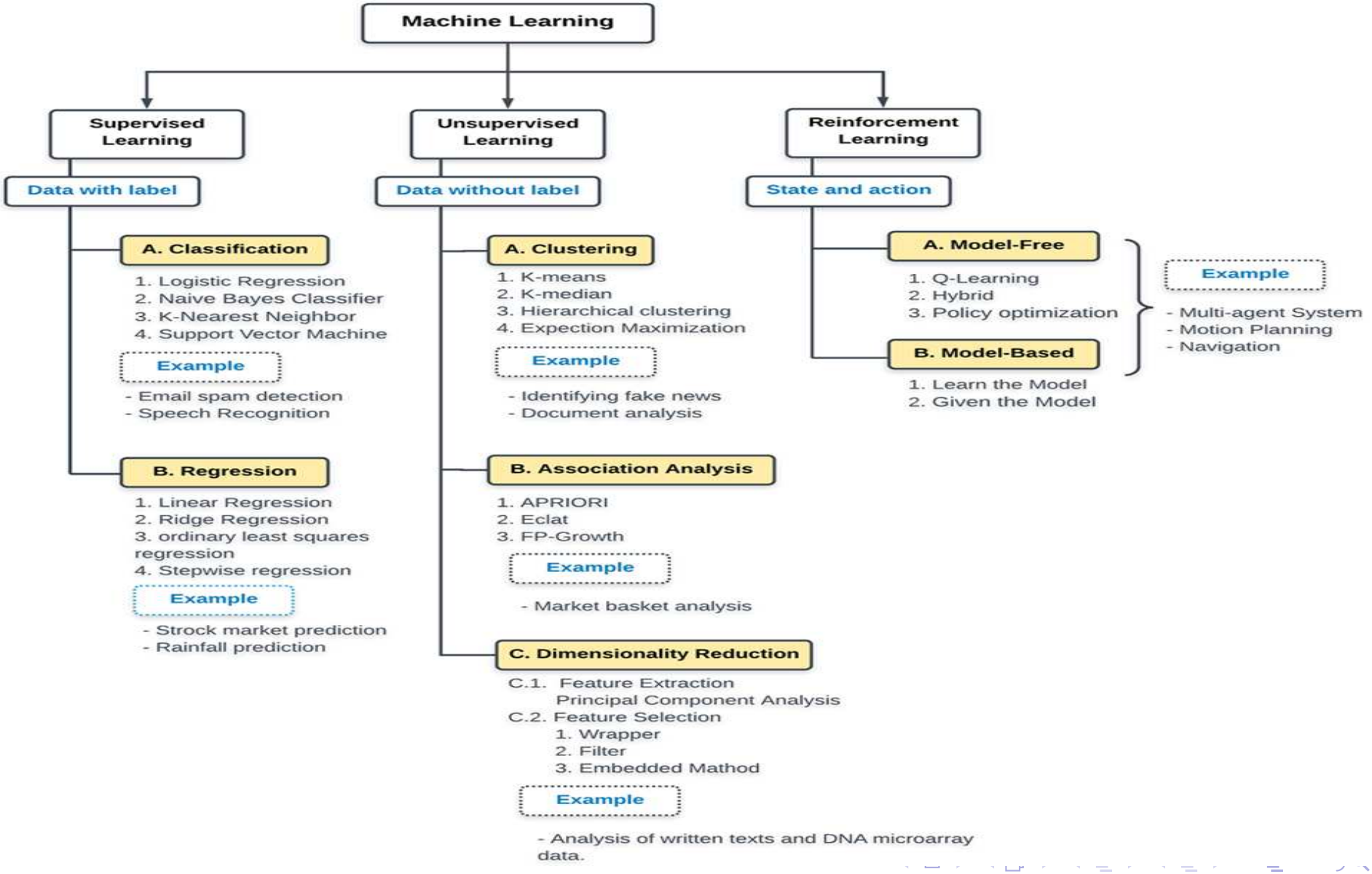


- Reinforcement learning (RL) is learning by interacting with an environment.
- Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions.
- In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.
- Since there is no labeled data, so the agent is bound to learn by its experience only.
- RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as game-playing, robotics, etc.

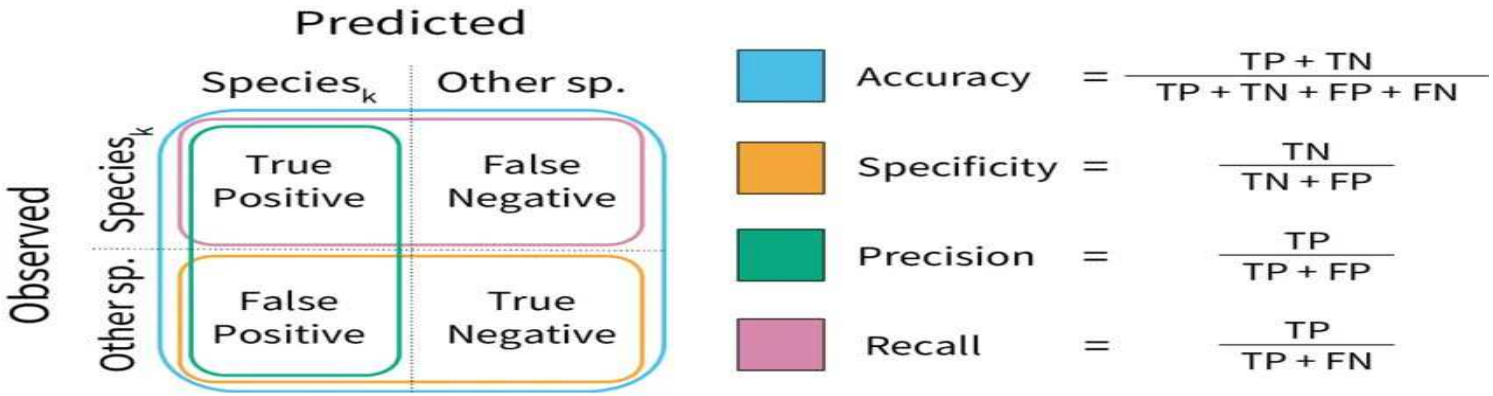
Comparison of Types of Machine Learning

Criteria	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	It is a method in which we teach the machine using labelled data	Machine is trained on an unlabelled data without any guidance	In reinforcement learning the agent interacts with its environment by producing actions and discovers errors or rewards
Type of Problems	Regression and Classification	Association and Clustering	Reward Based
Training	External Supervision	No Supervision	No Supervision
Aim	Aim is to forecast outcomes	Aim is to discover underlying patterns	Aim is to learn series of action
Type of Data	Labelled Data	Unlabelled Data	No Predefined Data
Approach	Map Labelled Input to known output	Understand patterns and discover outputs	Follow Trail and Error Method
Output Feedback	Direct Feedback	No Feedback	Reward System
Popular Algorithms	Linear Regression, Logistic Regression,SVM,K Nearest Neighbour, Random Forest etc.	k-means, Apriori,C-means etc.	Q-Learning, SARSA etc.

Machine Learning Classification



Performance Metrics



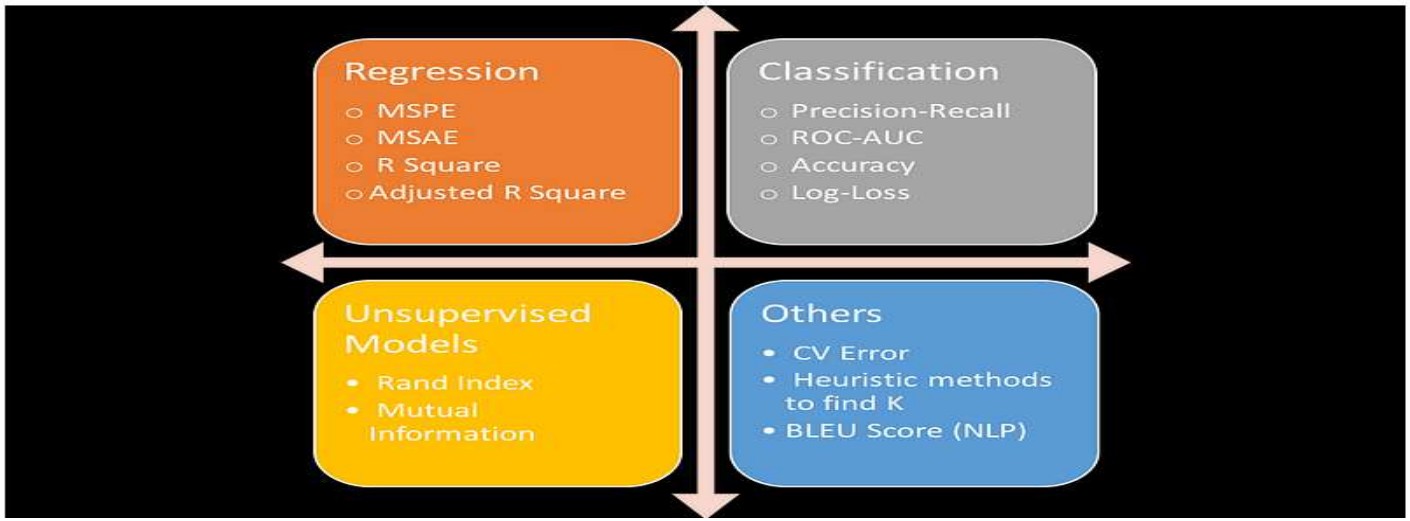
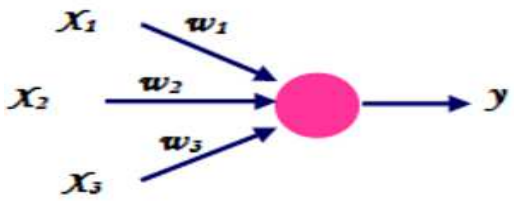


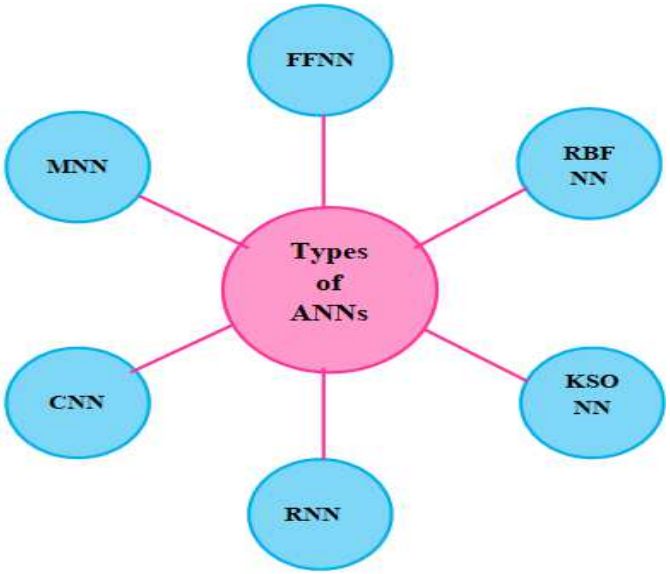
Figure: Performance Metrics of Machine Learning Algorithms

Artificial Neural Networks I

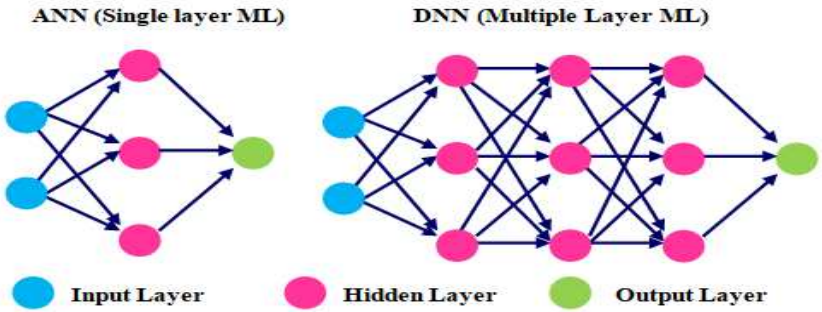


Perceptron model

$$y = \sum_{i=0}^{n} w_i x_i$$



Classification of ANNs



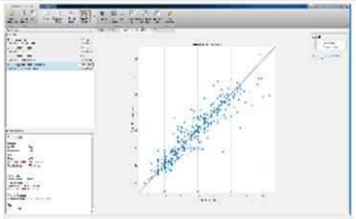
Structure of ANN and DNN

Machine Learning Toolbox

Statistics and Machine Learning Toolbox

Search MathWorks.com

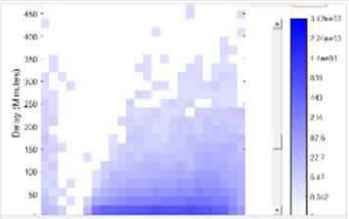
Overview | Features | Code Examples | Videos | Webinars | What's New | Product Pricing | Trial software | Contact sales



Regression Learner App

Train regression models using supervised machine learning

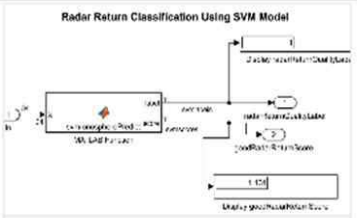
R2017a



Big Data Algorithms

Perform support vector machine (SVM) and Naive Bayes classification, create bags of decision trees, and fit lasso regression on out-of-memory data

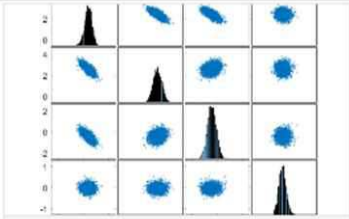
R2017a



Code Generation

Generate C code for prediction by using linear models, generalized linear models, decision trees, and ensembles of classification trees (requires MATLAB Coder)

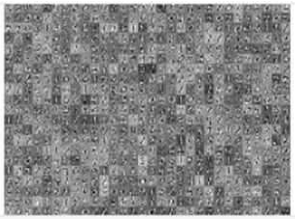
R2017a



Bayesian Statistics

Perform gradient-based sampling using Hamiltonian Monte Carlo (HMC) sampler

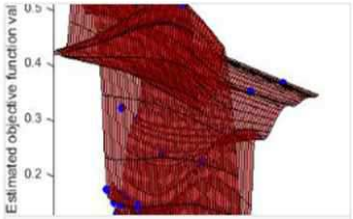
R2017a



Feature Extraction

Perform unsupervised feature learning by using sparse filtering and reconstruction independent component analysis (RICA)

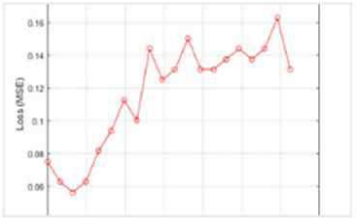
R2017a



Bayesian Optimization

Tune machine learning algorithms by searching for optimal hyperparameters

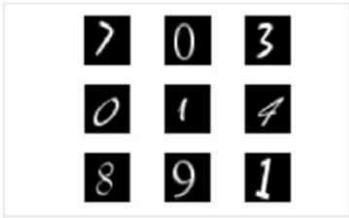
R2016b



Feature Selection

Use neighborhood component analysis (NCA) to choose features for machine learning models

R2016b



Code Generation

Generate C code for prediction by using SVM and logistic regression models (requires MATLAB Coder)

R2016b

Machine Learning Toolbox I

- MATLAB has two specific toolboxes for processing machine learning problems. They are the Statistics and Machine Learning Toolbox and Neural Network Toolbox. While the first solves machine learning problems through statistical techniques and algorithms most widely used in this field, the second is specific to ANNs. In the following sections, we will analyze in detail the features of these tools.
- The Statistics and Machine Learning Toolbox contains all the tools necessary to extract knowledge from large datasets.
  1. Regression techniques, including linear, generalized linear, nonlinear, robust, regularized, ANOVA, repeated measures, and mixed-effects models
  2. Big data algorithms for dimension reduction, descriptive statistics, k-means clustering, linear regression, logistic regression, and discriminant analysis
  3. Univariate and multivariate probability distributions, random and quasi-random number generators, and Markov chain samplers
  4. Hypothesis tests for distributions, dispersion, and location; Design of Experiments (DOE) techniques for optimal, factorial, and response surface designs

Machine Learning Toolbox II

5. Classification Learner app and algorithms for supervised machine learning, including SVMs, boosted and bagged decision trees, KNN, Naive Bayes, discriminant analysis, and Gaussian process regression
6. Unsupervised machine learning algorithms, including kmeans, k-medoids, hierarchical clustering, Gaussian mixtures, and HMMs
7. Bayesian optimization for tuning machine learning algorithms by searching for optimal hyperparameters

Neural Network Toolbox I

- The Neural Network Toolbox provides algorithms, pre-trained models, and apps to create, train, visualize, and simulate neural networks with one hidden layer (called shallow neural network) and neural networks with several hidden layers (called deep neural networks).
- Here is a descriptive list of the key features of this tool.
  1. Deep learning with CNNs (for classification and regression) and autoencoders (for feature learning)
  2. Transfer learning with pre-trained CNNs models and models from the Caffe model zoo
  3. Training and inference with CPUs or multi-GPUs on desktops, clusters, and clouds
  4. Unsupervised learning algorithms, including selforganizing maps and competitive layers
  5. Supervised learning algorithms, including multilayer, radial basis, Learning Vector Quantization (LVQ), time-delay, nonlinear autoregressive (NARX), and Recurrent Neural Network (RNNs)
  6. Apps for data fitting, pattern recognition, and clustering

Python I

- Python is an interpreted high-level general-purpose programming language.
- Invented in the Netherlands, early 90s by Guido van Rossum.
- Guido Van Rossum is fan of '**Monty Python's Flying Circus** ', this is a famous TV show in Netherlands.Named after Monty Python

Python II

- Why to use Python. . . ?
  - **Python is object oriented**– Structure supports such concepts as polymorphism, operation overloading, and multiple inheritance.
  - **It's free (open source)**– Downloading and installing Python is free and easy Source code is easily accessible
  - **It is powerful**–Dynamic typing,Built-in types and tools,Library utilities,Third party utilities (e.g. Numeric, NumPy, SciPy),Automatic memory management
  - **It is portable**–Python programs will run in exactly the same manner, irrespective of platform.
- Who uses python today
  1. **Google** makes extensive use of Python in its web search system, and employs Python's creator
  2. **Intel, Cisco, Hewlett-Packard, Seagate, Qualcomm, and IBM** use Python for hardware testing
  3. **ESRI** uses Python as an end-user customization tool for its popular GIS mapping products
  4. The **YouTube** video sharing service is largely written in Python and many more.....

Machine Learning Libraries in Python I

- Python libraries that used in Machine Learning are:
  - Numpy−It is particularly useful for linear algebra, Fourier transform, and random number capabilities
  - Scipy−it contains different modules for optimization, linear algebra, integration and statistics
  - Scikit learn−Scikit learn supports most of the supervised and unsupervised learning algorithms. Scikit learn can also be used for data-mining and data analysis, which makes it a great tool who is starting out with ML.(Classification,Regression, Clustering, Dimensionality Reduction, Model Selection, Preprocessing)
  - Theano−Theano is a popular python library that is used to define, evaluate and optimize mathematical expressions involving multi dimensional arrays in an efficient manner.
  - TensorFlow−TensorFlow is widely used in the field of deep learning research and application.Handling deep neural networks,Natural Language Processing,Partial Differential Equation, Abstraction capabilities, Image, Text, and Speech recognition, Effortless collaboration of ideas and code
  - Keras−Keras makes it really for ML beginners to build and design a Neural Network.

Machine Learning Libraries in Python II

- PyTorch−supports on Computer Vision, Natural Language Processing(NLP) and many more ML programs
- Pandas−Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.
- Matplotlib−Matpoltlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning.

Where To Work For Machine Learning? I

- Integrated Development Environment(IDE)
  - Pycharm
  - Pydev
  - Spyder
  - Visual Studio
  - KDevelop
  - Geany
  - Eclipse
  - Netbeans
  - Komodo IDE
  - eric
  - wing IDE
  - Py Scripter
  - Komodo Edit
- The most recommended IDE is Thonny

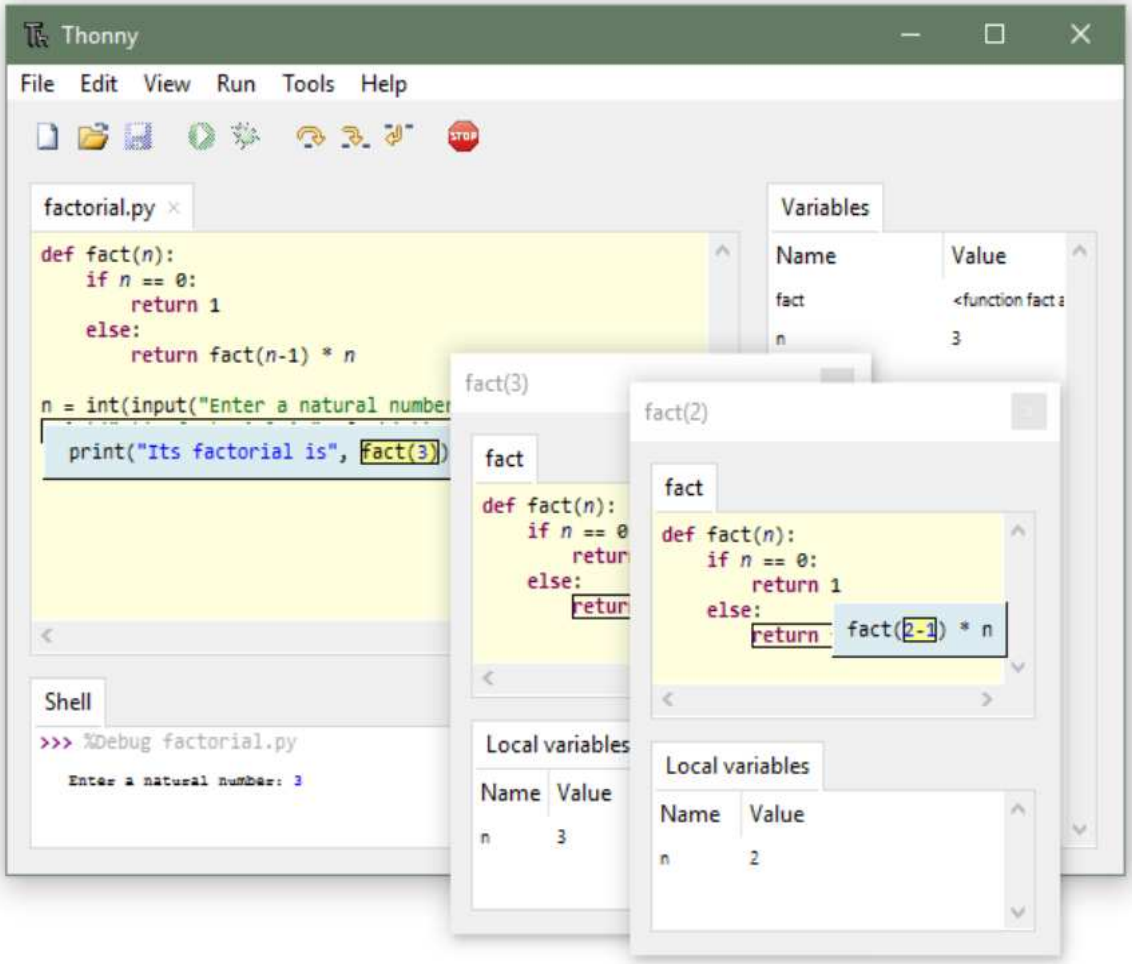
Where To Work For Machine Learning? II

Thonny

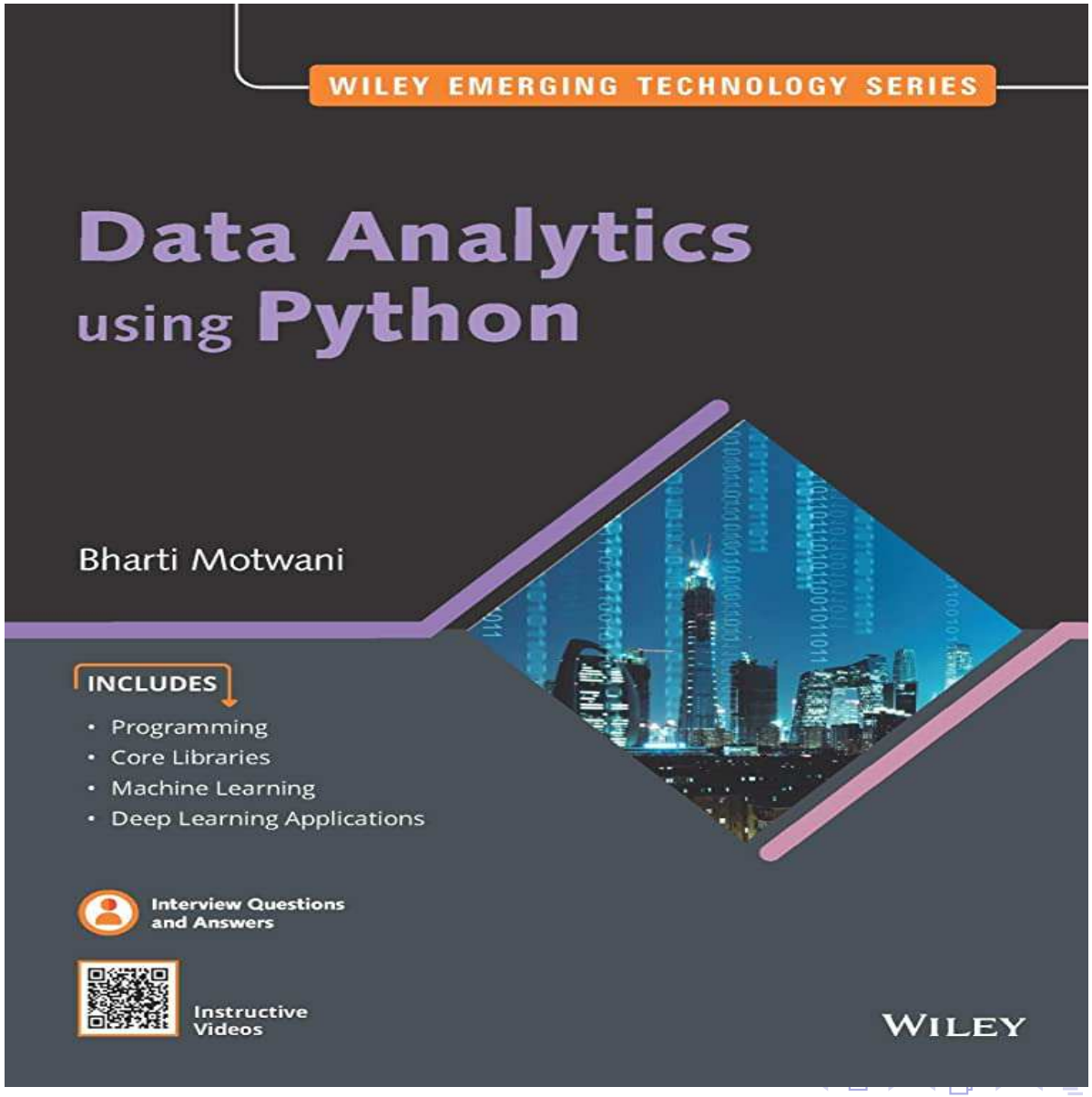
Python IDE for beginners

Download version 4.0.2 for Windows • Mac • Linux

For the curious: 4.1.0



Suggested Books



References I

1. A.C.Faul, A Concise Introduction to Machine Learning, Taylor & Francis group, 2020.
2. O.Campesato,Artificial intelligence Machine learning and deep learning, Mercury Learning and Information,New Delhi,2020.
3. Hilal M. El Misilmani and Tarek Naous, *Machine Learning in Antenna Design: An Overview on Machine Learning Concept and Algorithms*, 2019 International Conference on High Performance Computing & Simulation (HPCS),pp.600−607, 2019.
4. `https://data-flair.training/blogs/machine-learning-applications/`.

THANK YOU

THANK YOU