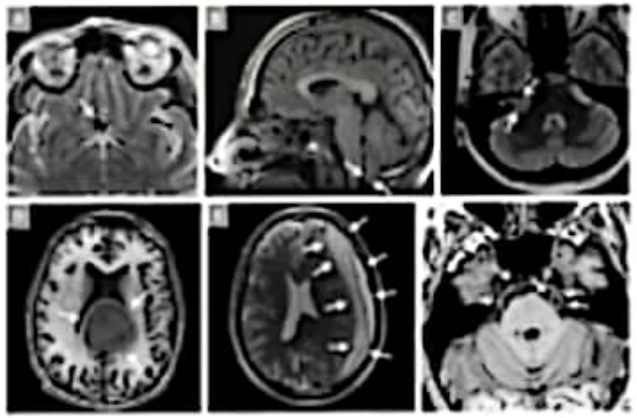
Unit-1

- **Introduction:** What is Human Learning? Types of Human Learning, what is Machine Learning? Types of Machine Learning, Problems Not to Be Solved Using Machine Learning, Applications of Machine Learning, State-of-The-Art Languages/Tools in Machine Learning, Issues in Machine Learning
- **Preparing to Model:** Introduction, Machine Learning Activities, Basic Types of Data in Machine Learning, Exploring Structure of Data, Data Quality and Remediation, Data Pre-Processing

Introduction to Machine Learning

- The machine learning is a mature technology area finding its application in almost every where in life.
- It predicts the future market to help unprofessional traders compete with seasoned stock traders.
- It helps an oncologist find whether a tumor is mild or serious .
- It helps in optimizing energy consumption thus helping the cause of Green Earth.
- **Google** has become one of the front-runners focusing a lot of its research on machine learning and artificial intelligence –
- Google self-driving car and Google Brain being two most ambitious projects of Google in its journey of innovation in the field of machine learning.



Machine Learning

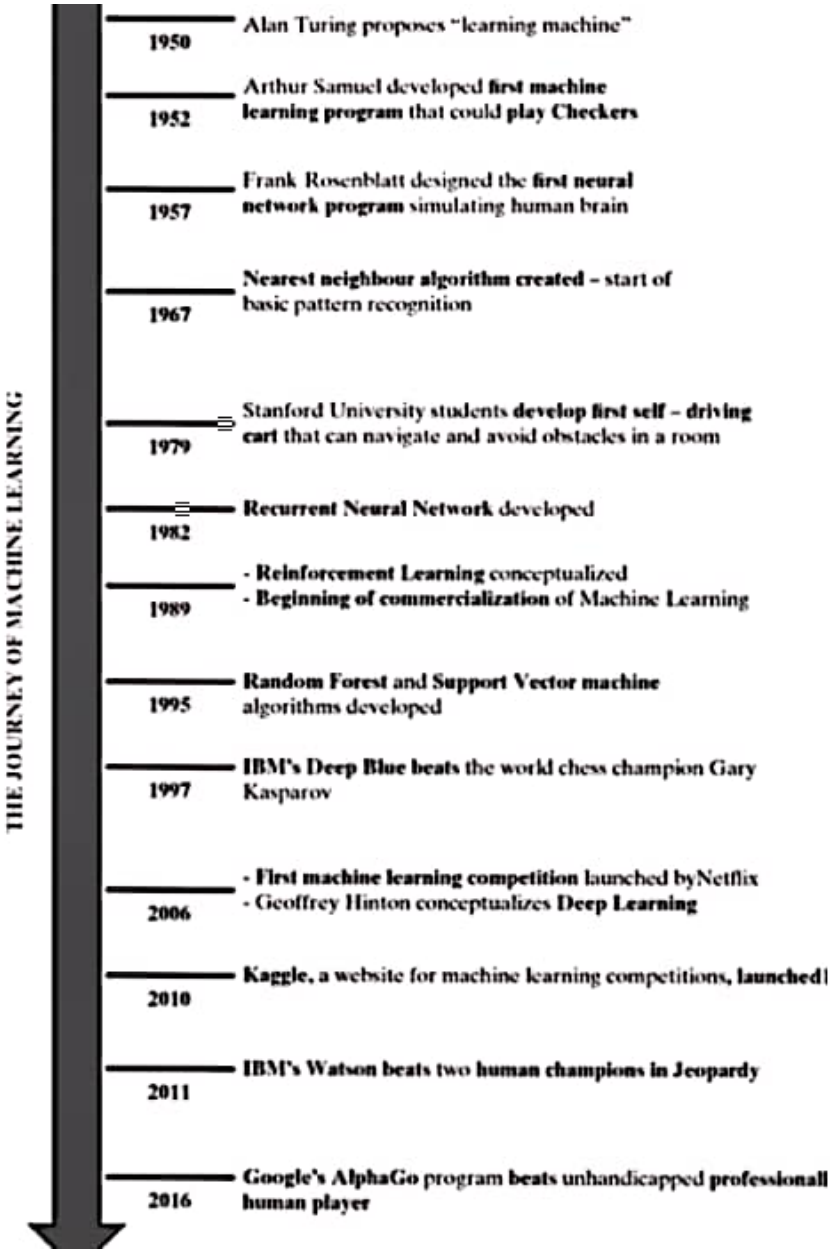
- Machine learning is a branch of Artificial Intelligence (AI) And Computer Science which focuses on the use of data and algorithms to imitate the way that **humans learn**, gradually improving its accuracy.

by IBM

Types of Machine Learning

- Machine learning algorithms are classified by four basic approaches, based on how an algorithm learns to become more accurate in its predictions.
- Supervised Learning,
- Unsupervised Learning,
- Semi-supervised Learning and
- Reinforcement Learning.

The evolution of machine learning



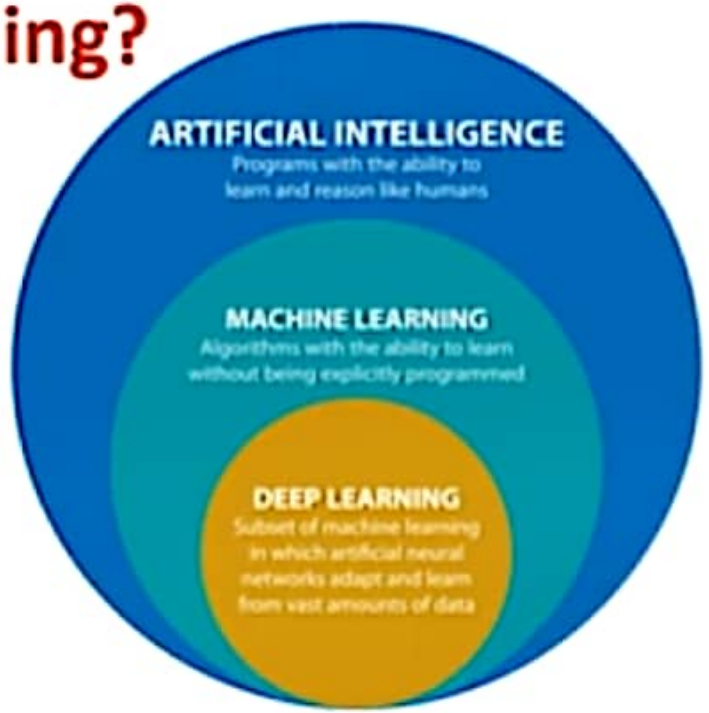
Machine Learning

Subject Code: 20A05602T

UNIT I – Introduction to Machine Learning & Preparing to Model

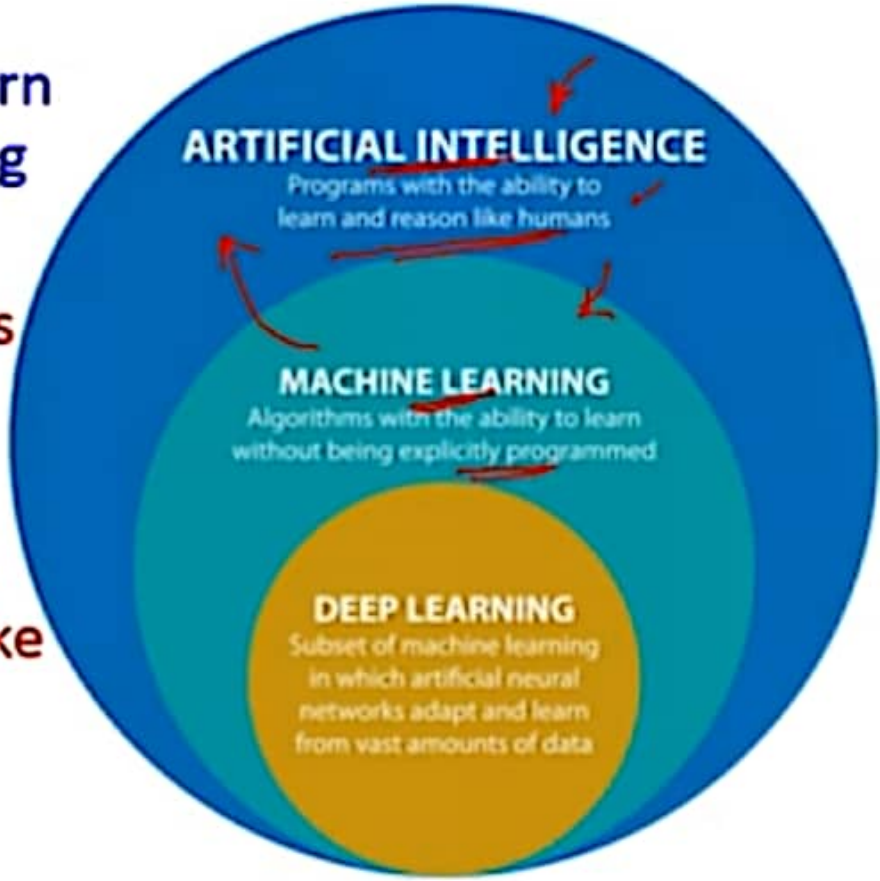
What is Machine Learning?

- How do machines learn?
- 1. Data input
- 2. Abstraction
- 3. Generalization



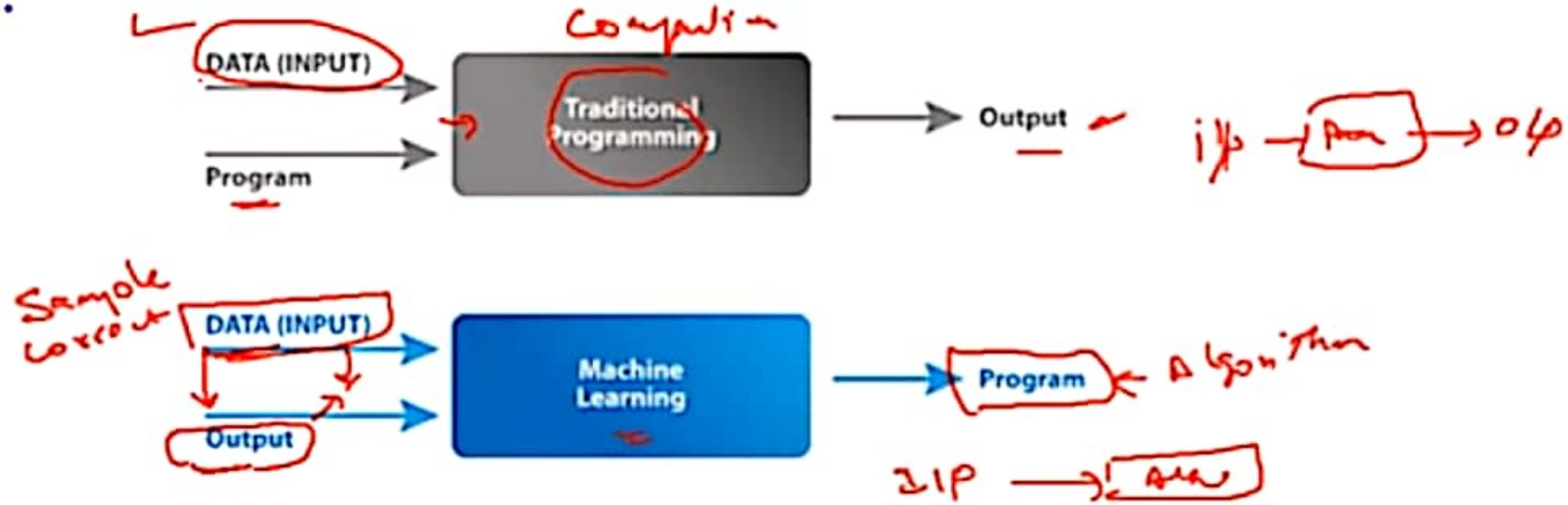
What is Machine learning?

- Machine learning is a subfield of artificial intelligence, which enables machines to learn from past data or experiences without being explicitly programmed.
- Machine learning uses statistical techniques to enable computers to learn and make decisions.
- It is predicated on the idea that computers can learn from data, spot patterns, and make judgments with little assistance from humans.



Difference between Traditional and Machine Learning programming

- In traditional programming, we would feed the input data and a well-written and tested program into a machine to generate output.
- In machine learning, input data, along with the output, is fed into the machine during the learning phase, and it works out a program for itself.



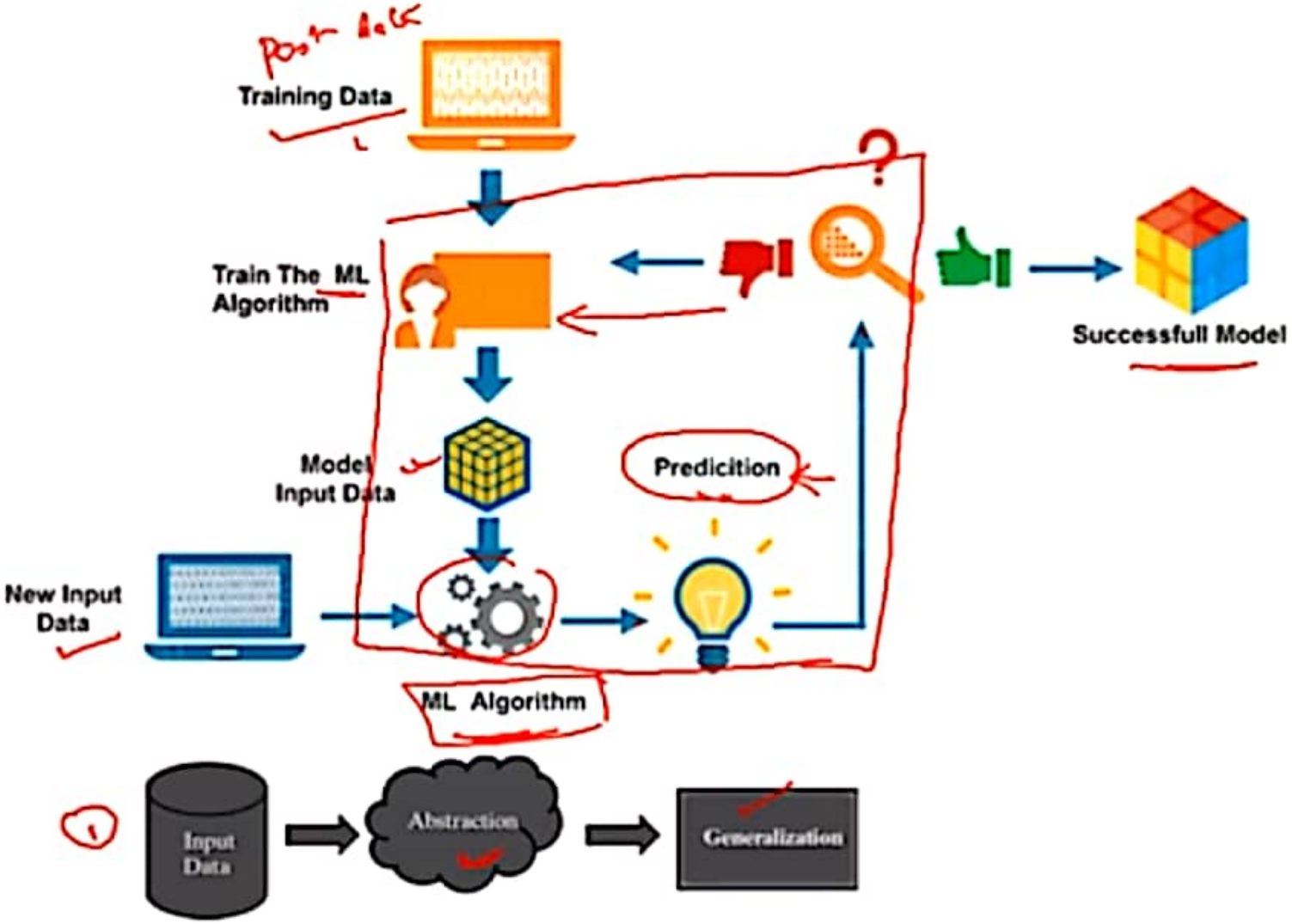
Machine Learning...

- **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.** by Tom M. Mitchell
- A machine can be considered to learn, if it is able to gather experience by doing a certain task and improve its performance in doing the similar tasks in the future.
- The past experience, it means past data related to the task, this data is an input to the machine from some source.

How do machines learn?

- The basic machine learning process can be divided into three parts.
- 1. Data Input: Past data or information is utilized as a basis for future decision-making
- 2. Abstraction: The input data is represented in a broader way through the underlying algorithm
- 3. Generalization: The abstracted representation is generalized to form a framework for making decision.

How do machines learn...



Data input

- Data is gathered from environment using sensors and/or past data taken from dataset
- A better learning strategy needs to be adopted:
- 1. to be able to deal with the vastness of the subject matter and the related issues in memorizing it
- 2. to be able to answer questions where a direct answer has not been learnt.
- Figure out the key points or ideas amongst a vast pool of knowledge.
- This helps in creating an outline of topics and a conceptual mapping of those outlined topics with the entire knowledge pool

2. Abstraction

- During the machine learning process, knowledge is fed in the form of input data.
- The data cannot be used in the original shape and form.
- Abstraction helps in deriving a conceptual map based on the input data.
- This map, or a model is known in the machine learning paradigm, is **summarized knowledge representation of the raw data.**

2. Abstraction...

- The model may be in any one of the following forms

1. Computational blocks like if/else rules
2. Mathematical equations
3. Specific data structures like trees or graphs
4. Logical groupings of similar observations

2. Abstraction...

- The choice of the model used to solve a specific learning problem
- The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:

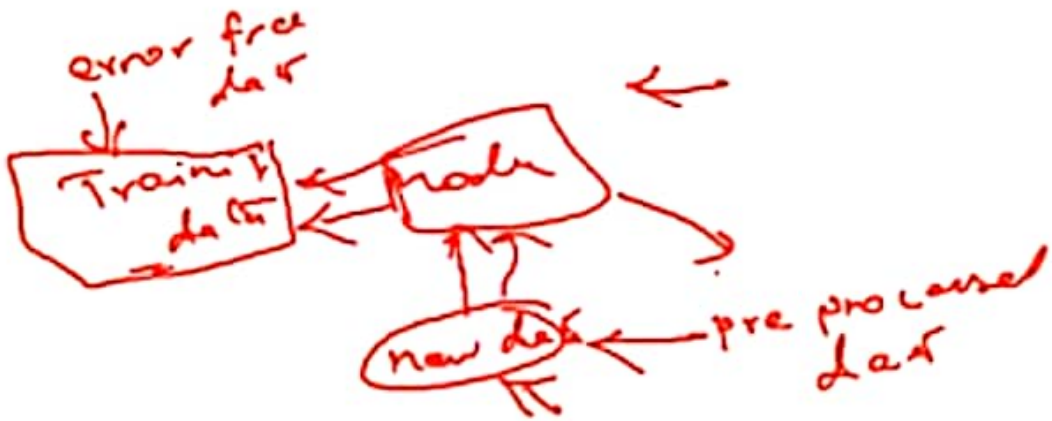
1. The type of problem to be solved:
   - Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
2. Nature of the input data:
   - How exhaustive (completeness) the input data is, and the data types, etc.
3. Domain of the problem:
   - A critical domain with a high rate of data input and need for immediate decision making.
   - e.g. fraud detection problem in banking domain.

3. Generalization

- The abstraction process, or **training the model**, used for abstract the knowledge which comes as input data in the form of a model.
- The **generalization** is, the abstracted knowledge to a form which can be used to take future decisions.
- The model is trained based on a finite set of data, which may possess a limited set of characteristics.
- Apply the model to take decision on a set of unknown data, usually termed as test data, then some problems occurs.

3. Generalization...

- then there are two problems:
- 1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
- 2. The test data have sometimes certain characteristics unknown to the training data.



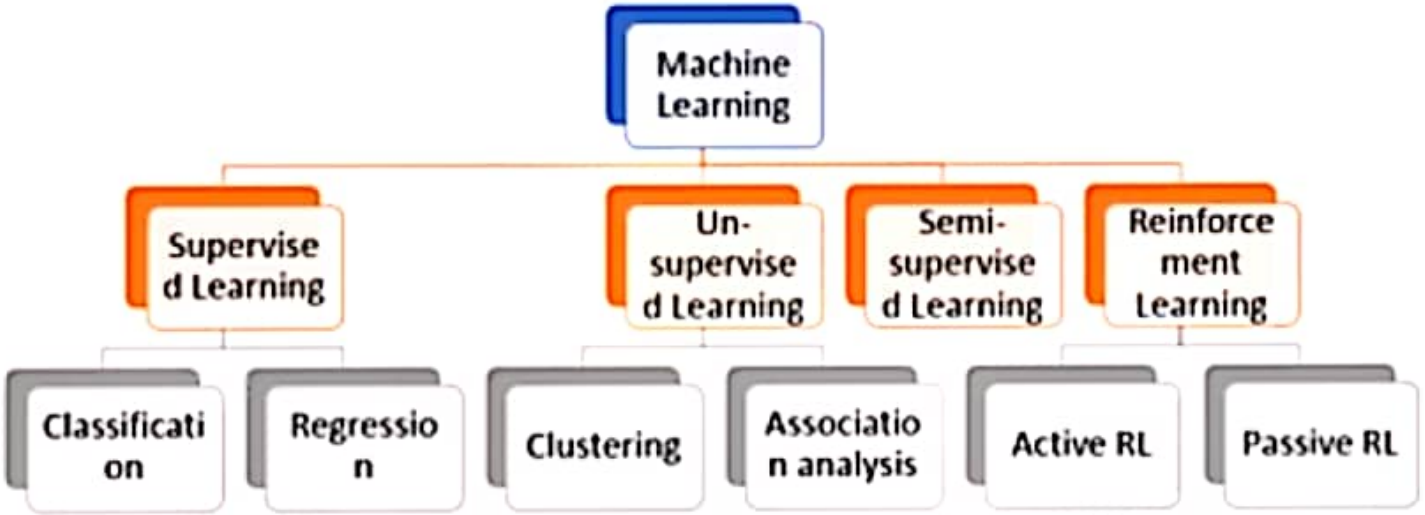
Machine Learning

Subject Code: 20A05602T

UNIT I – Introduction to Machine Learning & Preparing to Model

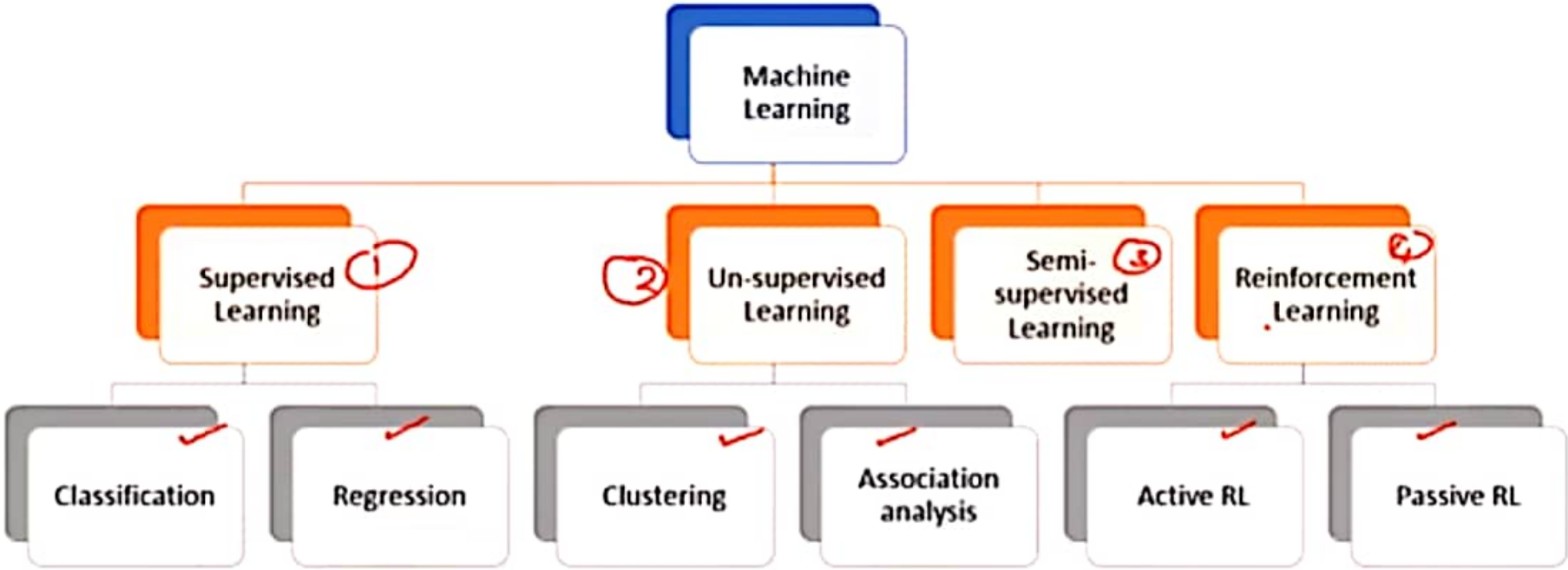
Types of Machine Learning

- Definition,
- Types,
- Applications,
- Advantages,
- Drawbacks



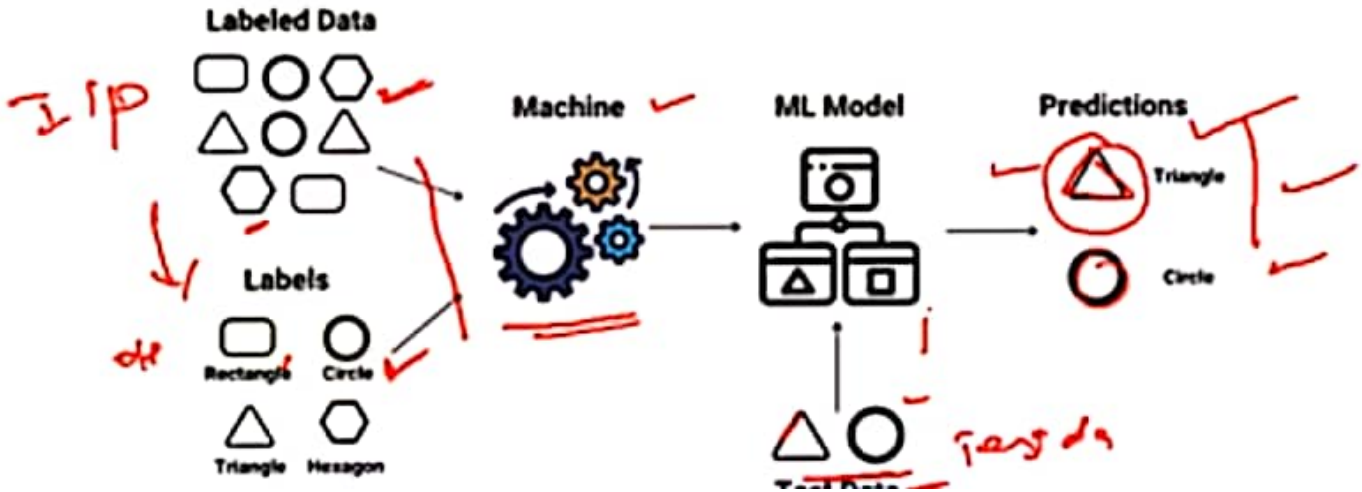
Types of Machine Learning

Based on the methods and way of learning, machine learning is divided into mainly four types



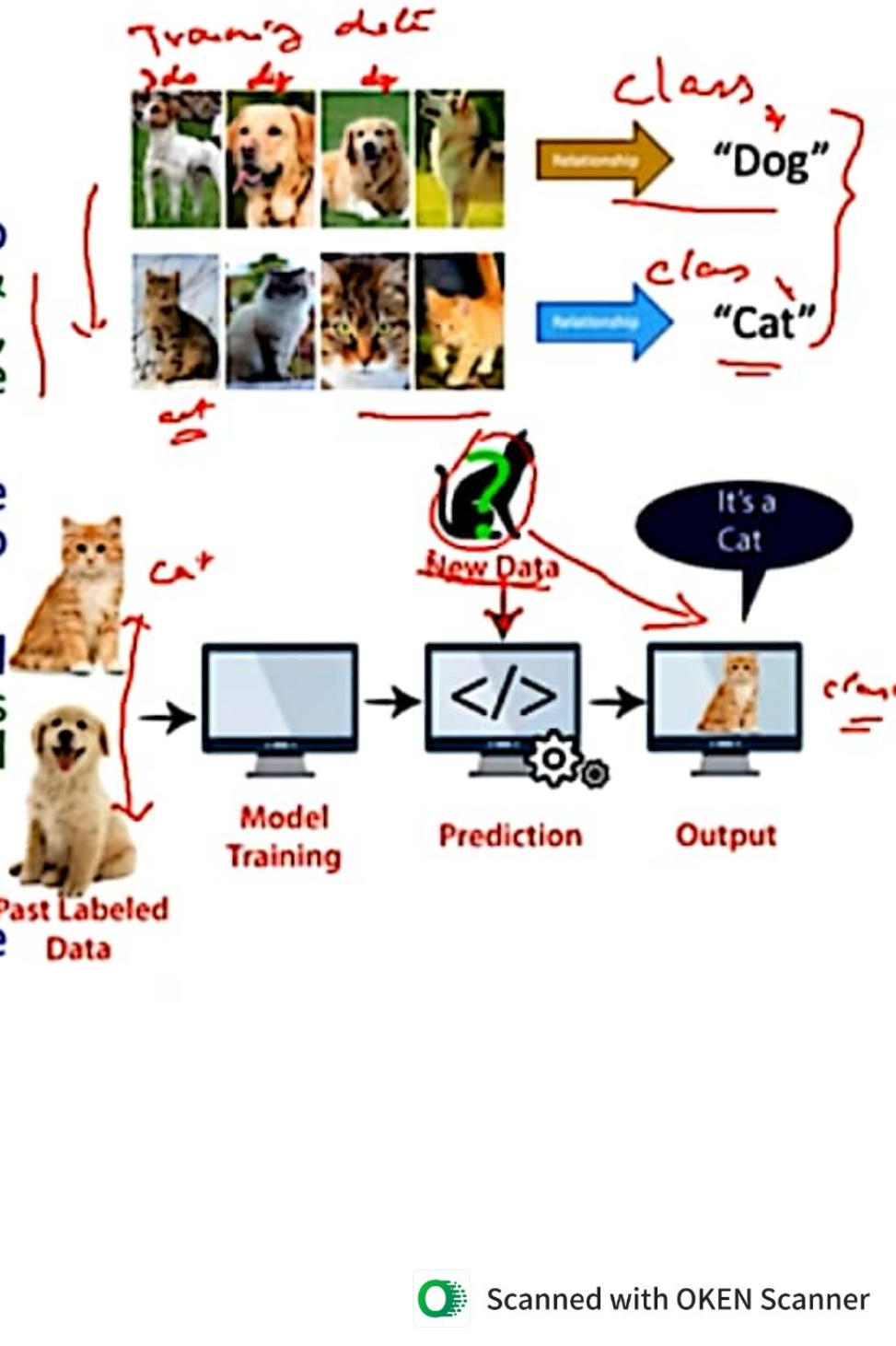
Supervised Machine Learning

- Supervised machine learning is based on supervision.
- It train the machines using the "labelled" dataset, and based on the training, the machine predicts the output.
- The labelled data specifies that some of the inputs are already mapped to the output.
- Train the machine with the input and corresponding output, and then the machine will predict the output using the test dataset.



Supervised Machine Learning...

- An input dataset of cats and dog images.
- first, provide the training to the machine to understand the images, such as the shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.
- After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output.
- Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat.
- So, it will put it in the Cat category.
- This is the process of how the machine identifies the objects in Supervised Learning.



Applications of Supervised Machine Learning

- Some real-world applications of supervised learning are
  - Fraud Detection,
  - Risk Assessment,
  - Spam filtering, etc.

Types of Supervised Machine Learning

- Supervised machine learning can be classified into two types of problems,
  - Classification
  - Regression

Classification

- Classification algorithms are used to solve the classification problems in which the output variable is categorical,
- such as "Yes" or No, Male or Female, Red or Blue, etc.
- The classification algorithms predict the categories present in the dataset.
- Some applications of classification algorithms are
  - Spam Detection, Email filtering, etc.
- Some popular classification algorithms are given below:
  - Random Forest Algorithm
  - Decision Tree Algorithm
  - Logistic Regression Algorithm
  - Support Vector Machine Algorithm

SVM

Regression

- Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables.
- These are used to predict continuous output variables, such as market trends, weather prediction, mark of student etc.
- Some popular Regression algorithms are given below:
  - Simple Linear Regression Algorithm
  - Multivariate Regression Algorithm
  - Decision Tree Algorithm
  - Lasso Regression

Advantages and Disadvantages of Supervised Learning

- Advantages:
- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.
- Disadvantages:
- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

- Some common applications of Supervised Learning are given below:

- Image Segmentation
- Medical Diagnosis
- Fraud Detection
- Spam detection
- Speech Recognition

Sensitive

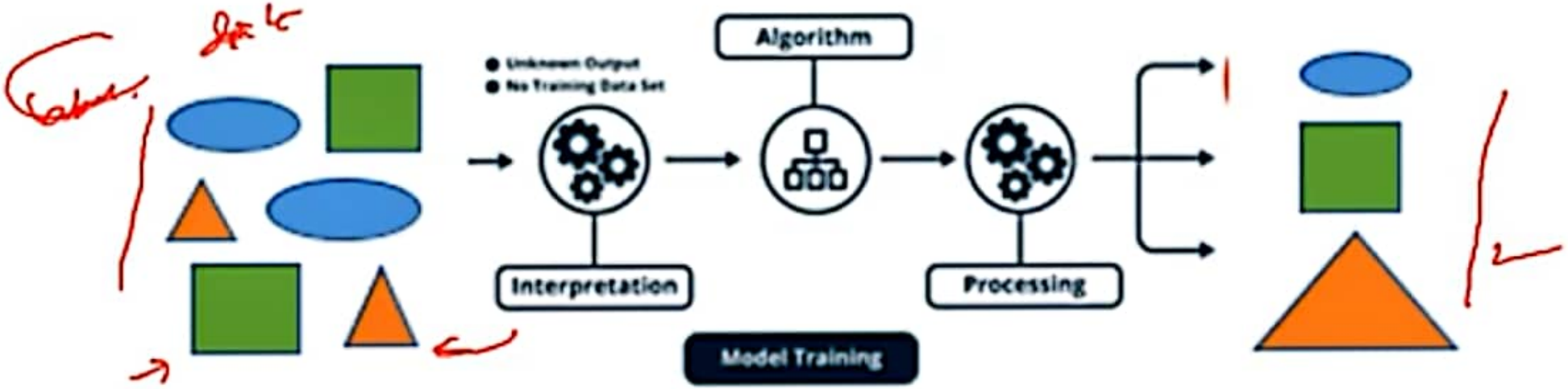
Applications of Supervised Learning

- Some common applications of Supervised Learning are given below:

- Image Segmentation
- Medical Diagnosis
- Fraud Detection
- Spam detection
- Speech Recognition

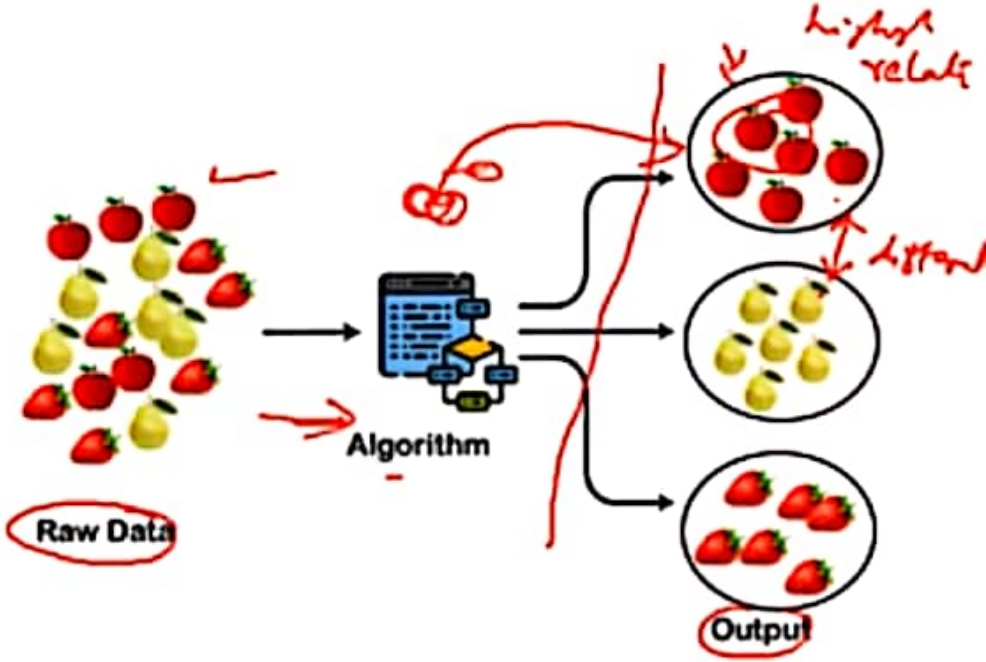
Unsupervised Machine Learning

- In unsupervised machine learning, the machine is trained using the unlabeled dataset and the machine predicts the output without any supervision.
- The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.
- Machines are instructed to find the hidden patterns from the input dataset.



Unsupervised Machine Learning

- Input it into the machine learning model is a basket of fruit images
- The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects.
- The machine will discover its patterns and differences, such as colour difference, shape difference.
- Then it should predict the output when it is tested with the test dataset.

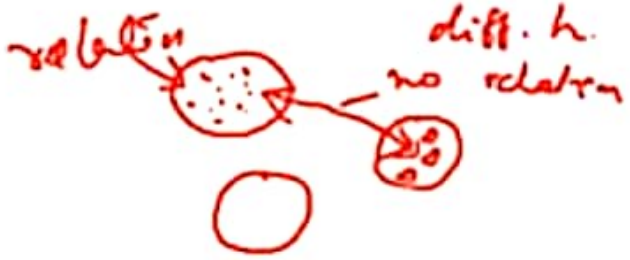


Categories of Unsupervised Machine Learning

- Unsupervised Learning can be further classified into two types, which are given below:

1. Clustering
2. Association

Clustering



- The clustering is used to find the inherent groups from the data.
- the objects in the group are most similar to each other and no similarities with the objects of other groups.
- An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

The Popular Clustering Algorithms

- Some of the popular clustering algorithms are given below:
  - K-Means Clustering algorithm
  - Mean-shift algorithm
  - DBSCAN Algorithm
  - Principal Component Analysis
  - Independent Component Analysis

Association

- Association finds interesting relations among variables within a large dataset.
- It is used to find the dependency of one data item on another data item
- Based of dependency it maps those variables so that it can generate maximum.

Association...

- This algorithm is mainly applied in
  - Market Basket analysis,
  - Web usage mining,
  - continuous production, etc.
- Some popular algorithms of Association rule learning are
  - Apriori Algorithm,
  - Eclat,
  - FP-growth algorithm, etc

Advantages and Disadvantages of Unsupervised Learning Algorithm

- Advantages:
- These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
- Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.
- Disadvantages:
- The output of an unsupervised algorithm can be less accurate as the dataset is not labelled,
- The algorithms are not trained with the exact output in prior.
- It is more difficult as it works with the unlabeled dataset that does not map with the output.

Applications of Unsupervised Learning

- Network Analysis
- Recommendation Systems
- Anomaly Detection
- Singular Value Decomposition

Semi-Supervised Learning

- **To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced.**
- It is the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms
- Hence, it uses the combination of labelled and unlabeled datasets during the training period.
- Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data.
- It is because labelled data is a comparatively more expensive than unlabeled data.

Advantages and disadvantages of Semi-supervised Learning

- Advantages:
- It is simple and easy to understand the algorithm.
- It is highly efficient.
- It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.
- Disadvantages:
- Iterations results may not be stable.
- We cannot apply these algorithms to network-level data.
- Accuracy is low.

4. Reinforcement Learning

- Reinforcement learning works on a feedback-based process, learning from experiences, and improve its performance.
- Agent gets rewarded for each good action and get punished for each bad action;
- hence the goal of reinforcement learning agent is to maximize the rewards.
- A reinforcement learning problem can be formalized using Markov Decision Process (MDP)
- In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.
- In RL, there is no labelled data like supervised learning, and agents learn from their experiences only.

Categories of Reinforcement Learning

- Reinforcement learning is categorized mainly into two types of methods/algorithms:
- **Passive Reinforcement Learning:**
  - the agent's policy (sequence of actions) is fixed which means that it is ***told what to do***.
  - Therefore, the goal of a passive RL agent is to execute a fixed policy and evaluate it.
- **Active Reinforcement Learning:**
  - An agent ***needs to decide what to do*** as there's no fixed policy that it can act on.
  - An active RL agent is to act and learn an optimal policy.

Applications of Reinforcement Learning

- Video Games
- Resource Management
- Robotics
- Text Mining and etc.

Advantages and Disadvantages of Reinforcement Learning

- **Advantages**
- It helps in solving complex real-world problems which are difficult to be solved by general techniques.
- The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
- Helps in achieving long term results.
- **Disadvantage**
- RL algorithms are not preferred for simple problems.
- RL algorithms require huge data and computations.

Thank You

- **UNIT I – Introduction to Machine Learning &**
- **Preparing to Model**
- **Types of Machine Learning**
- Definition,
- Types,
- Applications,
- Advantages,
- Drawbacks

Machine Learning

Subject Code: 20A05602T

UNIT I – Introduction to Machine Learning & Preparing to Model

Problems not to be Solved Using Machine Learning

- The few problems which ML fails to answer.
- 1. Reasoning Power
- 2. Contextual Limitation
- 3. Scalability
- 4. Regulatory Restriction For Data In ML
- 5. Internal Working Of Deep Learning

Problems not to be Solved Using Machine Learning

- Machine learning should not be applied to tasks in which humans are very effective or frequent human intervention is needed.
- For example, air traffic control is a very complex task needing intense human involvement.
- At the same time, for very simple tasks which can be implemented using traditional programming paradigms, there is no sense of using machine learning.
- For example, simple rule-driven or formula-based applications like price calculator engine, dispute tracking application, etc. do not need machine learning technique

ML

Reasoning Power

- The ML has not mastered successfully is reasoning power,
- Algorithms available today are mainly focused on specific use-cases and are narrowed down to an application.
- They cannot think as to why a particular method is happening that way.
- For example, an image recognition algorithm identifies apples and oranges in a given scenario.
- But, it cannot say if the apple (or orange) is good or bad, or why is that fruit an apple or orange.
- Mathematically, all of this learning process can be explained by us

2. Contextual Limitation.

- If we consider the area of natural language processing (NLP), text and speech information are there to understand languages by NLP algorithms.
- They may learn letters, words, sentences or even the syntax, but, the algorithms do not understand the context of the language used.
- ML does not have an overall idea of the situation.
- It is limited by mnemonic interpretations rather than thinking to see what is actually going on.

3. Scalability.

- ML algorithms are depends on data as well as its scalability.
- Data is growing at an enormous rate, and has many forms which largely affects the scalability of an ML project.
- Algorithms cannot do much about this, unless they are updated constantly for new changes to handle data.
- This is where ML regularly requires human intervention in terms of scalability and remains unsolved mostly.
- In addition, growing data has to be distributed the right way, if shared on an ML platform, which again needs examination through knowledge and lacked by current ML.

4. Regulatory Restriction for Data in ML

- ML usually need considerable amounts (in fact, massive) of data in stages such as training, cross-validation etc.
- Sometimes, data includes private (sensitive) as well as general information.
- Collecting and maintaining these types of data are complicated
- Because, the risk of the wrong usage of data, especially in critical areas such as medical research, health insurance etc.,
- Hence, this is the reason regulatory rules are imposed heavily when it comes to using private data.

5. Internal Working of Deep Learning

- Deep Learning (DL) now powers applications such as voice recognition, image recognition and etc., through artificial neural networks.
- But, the internal working of DL is still unknown and yet to be solved.
- Millions of neurons that form the neural networks in DL increase abstraction at every level, which cannot be understood at all.
- Advanced DL algorithms still confuse the researchers in terms of its working and efficiency.
- This is why deep learning is dubbed a 'black box' since its internal agenda is unknown.

Machine Learning

Subject Code: 20A05602T

UNIT I – Introduction to Machine Learning & Preparing to Model

Applications of Machine Learning

- Banking and finance
- Insurance
- Healthcare

Banking and Finance

- In the banking industry, fraudulent transactions, especially the ones related to credit cards, are extremely harmful.
- Since the volumes and velocity of the transactions are extremely high,
- So, a high performance machine learning solutions are implemented by almost all leading banks across the globe.
- The models work on a real-time basis,
- i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence.

Insurance

- Insurance industry is extremely data intensive.
- Two major areas in the insurance industry where machine learning is used are risk prediction during new customer onboarding and claims management.
- Based on the quantum of risk predicted, the quote is generated for the potential customer.
- When a customer claim comes for settlement, past information related to historic claims, along with the adjustor notes are considered to predict.
- The information related to similar customers, i.e. customer belonging to the same geographical location, age group, ethnic group, etc., are also considered to formulate the model.

Healthcare

- ML algorithms are used to predict the health conditions of the person in real time.
- In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action.
- In case of some extreme problem, doctors or healthcare providers in the locality of the person can be alerted.
- **Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.**

Healthcare...

- Suppose an elderly person goes for a morning walk in a park close to his house.
- Suddenly, while walking, his blood pressure shoots up beyond a certain limit, which is tracked by the wearable.
- The wearable data is sent to a remote server and a machine learning algorithm is constantly analyzing the streaming data.
- It also has the history of the elderly person and persons of similar age group.
- If the model predicts some casualty, then immediate action is taken.
- Alert can be sent to the person to immediately stop walking and take rest.
- Also, doctors and healthcare providers can be alerted to be on standby.

Other Applications of Machine Learning

- We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc.

1. Image Recognition

- It is used to identify objects, persons, places, digital images, etc.
- The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion.**
- Facebook provides us a feature of auto friend tagging suggestion.
- Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name.
- The technology behind this is **face detection** and **recognition algorithm.**
- The Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

2. Speech Recognition

- While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.
- Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition.**"
- At present, machine learning algorithms are widely used by various applications of speech recognition.
- **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. Traffic prediction.

- If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.
- It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:
  - **Real Time location** of the vehicle form Google Map app and sensors
  - **Average time has taken** on past days at the same time.
- Everyone who is using Google Map is helping this app to make it better.
- It takes information from the user and sends back to its database to improve the performance.

4. Product Recommendations

- Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon, Netflix, YouTube**, etc., for product recommendation to the user.
- Whenever we search for some product on Amazon, then we started getting an advertisement for the same product, while internet surfing on the same browser and this is because of machine learning.
- Google understands the user interest using various machine learning algorithms, and suggests the product as per customer interest.
- As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars

- One of the most exciting applications of machine learning is self-driving cars.
- Machine learning plays a significant role in self-driving cars.
- Tesla, the most popular car manufacturing company is working on self-driving car.
- Already Google's Car is very popular.
- It is using unsupervised learning method to train the car models to detect people and objects while driving

6. Email Spam and Malware Filtering

- Whenever we receive a new email, it is filtered automatically as important, normal, and spam.
- We always receive an important mail in our inbox with the important symbol and
- spam emails in our spam box, and the technology behind this is Machine learning.
- Below are some spam filters used by Gmail:
  - Content Filter
  - Header filter
  - General blacklists filter

7. Virtual Personal Assistant - VPA

- We have various virtual personal assistants such as **Google assistant, Alexa, Cortana, Siri**
- As the name suggests, they help us in finding the information using our voice instruction
- These assistants can help us in various ways just by our voice instructions such as
  - Play music,
  - call someone,
  - Open an email,
  - Scheduling an appointment, etc.

8. Online Fraud Detection

- Machine learning is making our online transaction safe and secure by detecting fraud transaction.
- Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place, such as **fake accounts**, **fake ids**, and **steal money** in the middle of a transaction.
- So to detect this, **Feed Forward Neural network** helps us by checking whether it is a genuine transaction or a fraud transaction.

9. Stock Market Trading

- Machine learning is widely used in stock market trading.
- In the stock market, there is always a risk of up and downs in shares..
- Machine learning's **long short term memory neural network** is used for the prediction of stock market trends.

10. Medical Diagnosis

- In medical science, machine learning is used for diseases diagnoses.
- With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of wounds in the brain.
- It helps in finding brain tumors and other brain-related diseases easily.

Thank You

- UNIT I – Introduction to Machine Learning &
- Preparing to Model
- Applications of Machine Learning
  - Banking and finance
  - Insurance
  - Healthcare

Machine Learning

Subject Code: 20A05602T

UNIT I – Introduction to Machine Learning & Preparing to Model

State-of-art Languages / Tools in Machine Learning

python™ sas

R SPSS STATISTICS IBM SPSS

julia

State-of-the-art Languages

- The algorithms related to different machine learning tasks are known to all and can be implemented using any language/platform.
- It can be implemented using a Java platform or C / C++ language or in .NET.
- However, there are certain languages and tools which have been developed with a focus for implementing machine learning.
  - Python
  - R Programming
  - MATLAB - matrix laboratory
  - SAS - Statistical Analysis System
  - SPSS - Statistical Package for the Social Sciences

Python

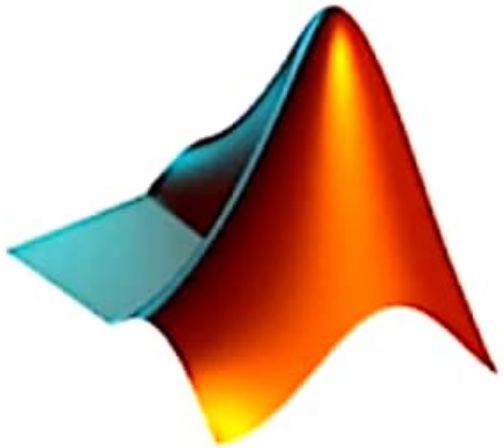
=

- Python is one of the most popular, open source programming language widely adopted by machine learning community.
- Python has very strong libraries for
  - advanced mathematical functionalities (NumPy),
  - algorithms and mathematical tools (SciPy) and
  - numerical plotting (matplotlib).
- Built on these libraries, there is a machine learning library named **scikit-learn**, which has various classification, regression, and clustering algorithms embedded in it.

R Programming

- R is a language for statistical computing and data analysis.
- It is an open source language, extremely popular in the academic community – especially among statisticians and data miners.
- R is a very simple programming language with a huge set of libraries available for different stages of machine learning.
- Some of the libraries standing out in terms of popularity are
  - plyr/dplyr (for data transformation),
  - caret ('Classification and Regression Training' for classification),
  - RJava (to facilitate integration with Java),
  - tm (for text mining),
  - ggplot2 (for data visualization).
- Other than the libraries, the packages like Shiny and R Markdown are used to develop interactive web applications, documents, dashboards and etc. on R, without much effort.

MATLAB



- MATLAB (matrix laboratory) is a licensed commercial software with a robust support for a wide range of numerical computing.
- MATLAB provides extensive support of statistical functions and has a huge number of **machine learning algorithms** in-built.
- It also has the ability to scale up for large datasets by parallel processing on clusters and cloud.

SAS

- SAS ('Statistical Analysis System') is another licensed commercial software which provides strong support for machine learning functionalities.
- SAS is a software suite comprising different components.
- The basic data management functionalities are embedded in the Base SAS component.
- The other components like SAS/INSIGHT, Enterprise Miner, SAS/STAT, etc. help in specialized functions related to data mining and statistical analysis.

SPSS - Statistical Package for the Social Sciences

- IBM, **SPSS** (Statistical Package for the Social Sciences) is a popular package supporting specialized data mining and statistical analysis.
- It is a predictive analysis software (PASW)
- SPSS is used in commercial, government, academic, and other organizations around the world to solve business and research problems.
  - Descriptive statistics (Mean, Median, Mode, Standard deviation, Range)
  - Discrete probability distributions (Binomial, Poisson, Geometric, Hyper geometric)
  - Continuous probability distributions (Normal, T, Chi Square, F)
  - Correlation (Rank correlation, Pearson's correlation)
  - Linear regression (Simple and Multiple linear regression)
  - Logistic regression and Market research. Etc.

Julia

- **Julia** is programming language used to solve many problems in machine learning numerical analysis and computational science and etc.
- It inherits all good things of MATLAB, Python, R, and other programming languages used to solve complicated machine learning applications.
- Julia has ability to implement high-performance mathematical and scientific machine learning algorithms.

Popular Tools in Machine Learning

- There are different tools, software, and platform available for machine learning, and also new software and tools are evolving day by day.
- some popular and commonly used Machine learning tools and their features.
  - TensorFlow
  - Pytorch
  - Google Cloud ML Engine
  - Amazon ML
  - Accord.NET & Azure ML
  - Apache Mahout
  - Jupyter,
  - Rapidminer
  - Accord.Net
  - Knime
  - Weka
  - Apache Spark MLlib
  - Keras
  - Shogun

1. TensorFlow

- TensorFlow is one of the most popular open-source libraries used to train and build both machine learning and deep learning models.
- It provides a JS library and was developed by **Google Brain Team**, to build ML applications.
- It offers a powerful library, tools, and resources for numerical computation, specifically for large scale projects.
- For training and building the ML models, TensorFlow provides a high-level Keras API, which lets users easily start with TensorFlow and machine learning.

2. PyTorch

- PyTorch is an open-source machine learning framework, which is based on the Torch library, developed by FAIR (Facebook's AI Research lab).
- It is one of the popular ML frameworks, which can be used for various applications, including computer vision and natural language processing.
- PyTorch has Python and C++ interfaces;
- Different deep learning software is made up on top of PyTorch, such as PyTorch Lightning, Hugging Face's Transformers, Tesla autopilot, etc.
- A Tensor class containing an n-dimensional array that can perform tensor computations along with GPU support.

3. Google Cloud ML Engine

- The Google Cloud ML Engine is the best choice for machine learning or deep learning projects, which requires millions or billions of training datasets, or the algorithms taking a long time for execution.
- It provides a managed service that allows developers to easily create ML models with any type of data and of any size.
- Provides machine learning model training, building, deep learning and predictive modelling.
- The two services - prediction and training, can be used individually or combinedly.

4. Amazon Machine Learning (AML)

- It is a cloud-based and robust machine learning software application, which is widely used for building machine learning models and making predictions.
- Enables the users to identify the patterns, build mathematical models, and make predictions.
- It provides support for three types of models, which are multi-class classification, binary classification, and regression.
- It permits users to import the model into or export the model out from Amazon Machine Learning.
- It also provides core concepts of machine learning, including ML models, Data sources, Evaluations, Real-time predictions and Batch predictions.

5. Accord.Net

- Accord.Net is .Net based Machine Learning framework, which is used for scientific computing.
- It is combined with audio and image processing libraries that are written in C#, it contains 38+ kernel Functions.
- This framework provides different libraries for various applications in ML, such as **Pattern Recognition, linear algebra, Statistical Data processing.**
- 40 non-parametric and parametric estimation of statistical distributions.
- Used for creating production-grade computer audition, computer vision, signal processing, and statistics apps.
- One popular package of the Accord.Net framework is
  - **Accord.Statistics,**
  - **Accord.Math, and**
  - **Accord.MachineLearning.**

6. Apache Mahout

- Apache Mahout is an open-source project of Apache Software Foundation, which is used for developing machine learning applications
- It is a distributed linear algebra framework and mathematically expressive Scala DSL, to develop own algorithms.
- It enables developers to implement machine learning techniques, including recommendation, clustering, and classification.
- It is an efficient framework for implementing scalable algorithms.
- It consists of matrix and vector libraries.
- It provides support for multiple distributed backends
- It runs on top of Apache Hadoop using the MapReduce paradigm.

7. Jupyter

- Jupyter is the latest web-based interactive development environment for notebooks, code, and data.
- Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning.

8. Shogun

将軍

- Shogun is a free and open-source machine learning software library.
- This software library is written in C++ and supports interfaces for different languages such as Python, R, Scala, C#, Ruby, etc., using **SWIG** (Simplified Wrapper and Interface Generator).
- It supports different kernel-based algorithms such as Support Vector Machine (SVM), K-Means Clustering, etc., for regression and classification problems.
- It also provides the complete implementation of Hidden Markov Models.

9. Oryx2

- Oryx2 is a realization of the lambda architecture and built on **Apache Kafka** and **Apache Spark**.
- It is widely used for real-time large-scale machine learning projects.
- It is a framework for building apps, including end-to-end applications for filtering, packaged, regression, classification, and clustering.
- It is written in Java languages, including Apache Spark, Hadoop, Tomcat, Kafka, etc.
- The latest version of Oryx2 is Oryx 2.8.0

10. Weka

- Weka is open source software, and has collection of machine learning algorithms for data mining tasks.
- It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.
- Weka support to solve much complicated deep learning problems.
- It can Run any workload on any cloud, even the "impossible" ones.
- It has High performance, low latency storage for I/O-intensive workloads like AI and machine learning.
- Used to Build Cloud Native Apps and Deploy Anywhere .

Thank You

- **UNIT I – Introduction to Machine Learning & Preparing to Model**
- **State-of-art Languages / Tools in Machine Learning**
  - Python
  - R
  - MATLAB
  - SAS
  - SPSS
  - **TensorFlow**
  - **Pytorch**
  - **Google Cloud ML Engine**
  - **Amazon ML**
  - **Accord.NET & Azure ML**
  - **Apache Mahout**
  - **Jupyter,**
  - **Rapidminer**
  - **Knime**
  - **Weka**
  - **Apache Spark MLlib**
  - **Keras**
  - **Shogun**

Machine Learning

Subject Code: 20A05602T

UNIT I – Introduction to Machine Learning & Preparing to Model

Issues in Machine Learning

- Inadequate Training Data
- Poor quality of data
- Non-representative training data
- Overfitting and Underfitting
- Monitoring and maintenance
- Getting bad recommendations
- Lack of skilled resources
- Customer Segmentation
- Process Complexity of Machine Learning
- Data Bias

Issues in Machine Learning

- The machine learning is being used in every industry and helps organizations make more informed and data-driven choices that are more effective than classical methodologies,
- it still has so many problems that cannot be ignored.
- Here are some common issues in Machine Learning that professionals face to train ML skills and create an application from scratch.

1. Inadequate Training Data

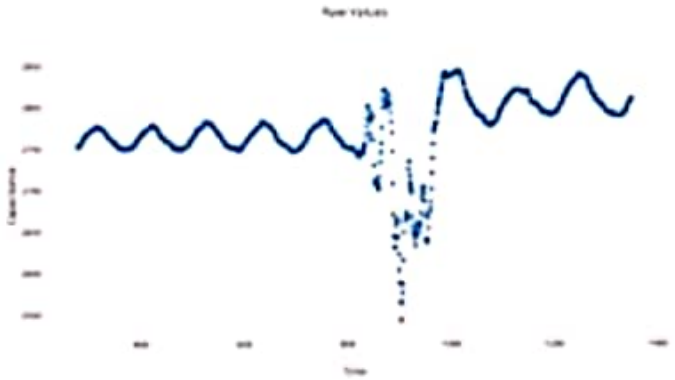
- The major issue is the lack of quality as well as quantity of data.
- The data plays a vital role in the processing of machine learning algorithms, but the inadequate data, noisy data, and unclean data are very big issue in modeling machine learning algorithms.
- For example, a simple task requires thousands of sample data, and an advanced task such as speech or image recognition needs millions of sample data examples.

2. Poor Quality of Data

- **The data quality** is also important, otherwise it can be affected by some factors as follows:
- **Noisy Data-** It is responsible for an inaccurate prediction that affects the decision as well as accuracy in classification tasks.
- **Incorrect data-** It is also responsible for faulty programming and results obtained in machine learning models. Hence, incorrect data may affect the accuracy of the results also.
- **Generalizing of output data-** generalizing output data becomes very complex, which results in comparatively poor future actions

2. Poor Quality of Data...

- Noisy data, incomplete data, inaccurate data, and unclean data lead to less accuracy in classification and low-quality results.
- Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.

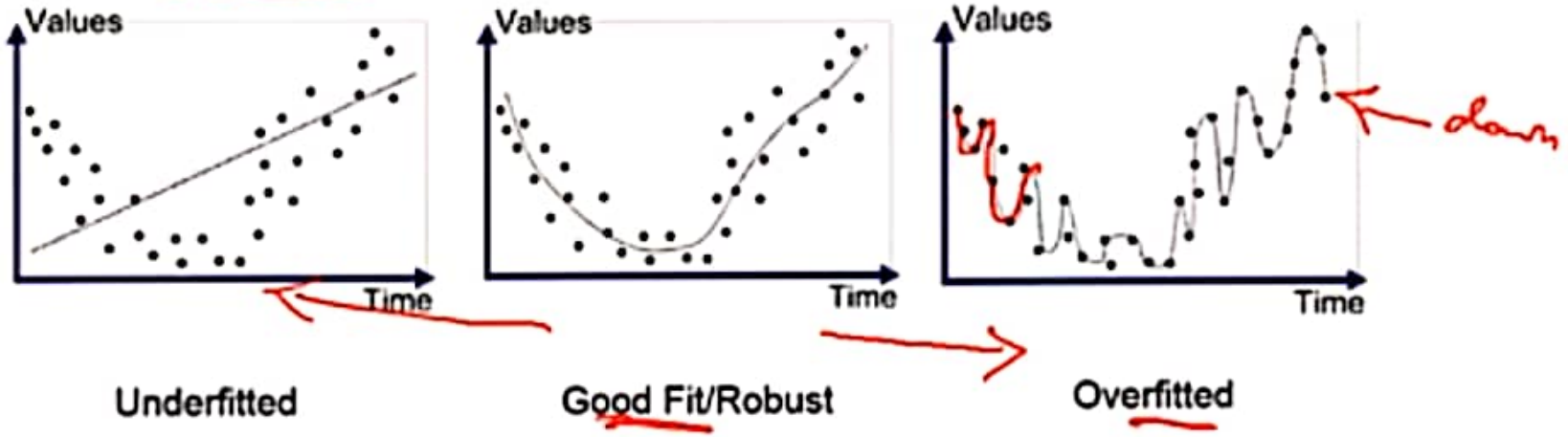


3. Non-representative Training Data

- Ensure that sample training data must be representative of new cases that we need to generalize.
- The training data must cover all cases or all classes of data
- **If there is less training data, then there will be a sampling noise in the model, called the non-representative training set.**
- If we are using non-representative training data in the model, it results in less accurate predictions.

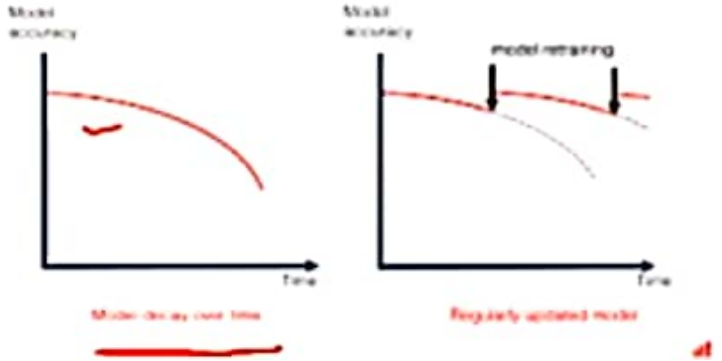
4. Overfitting and Underfitting

- Overfitting
- Refers to a model that models the training data too well.
- Overfitting happens when a model learns the detail and noise in the training data set to the extent that it negatively impacts the performance of the model.
- Underfitting
- When the model does not learn enough from the data, hence the machine cannot capture the underlying trend of the data.



6. Getting bad recommendations

- A machine learning model operates under a specific context, which results in bad recommendations.
- The recommendation model should understand the interest of users on specific time, because customers requirement changes over time,
- but still machine learning model showing same recommendations to the customer, while customer expectation has been changed.
- This is called a Data Drift.



7. Lack of skilled resources

- Although Machine Learning and Artificial Intelligence are continuously growing in the market, still these industries are fresher in comparison to others.
- The absence of skilled resources in the form of manpower is also an issue.
- Hence, we need manpower having in-depth knowledge of mathematics, science, and technologies for developing and managing scientific substances for machine learning

8. Customer Segmentation

- Customer segmentation is also an important issue while developing a machine learning algorithm.
- To identify the customers who paid for the recommendations shown by the model and who don't even check them.
- Hence, an algorithm is necessary to recognize the customer behavior and trigger a relevant recommendation for the user based on past experience.

9. Process Complexity of Machine Learning

- The machine learning process is very complex, which is also another major issue faced by machine learning engineers and data scientists.
- However, Machine Learning and Artificial Intelligence are very new technologies but are still in an experimental phase, and continuously being changing over time.
- There is the majority of hits and trial experiments; hence the probability of error is higher than expected.
- Further, it also includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, etc., making the procedure more complicated.

10. Data Bias

- The data bias exist when certain elements of the dataset are heavily weighted or need more importance than others.
- Biased data leads to inaccurate results, skewed outcomes, and other analytical errors.

Machine Learning

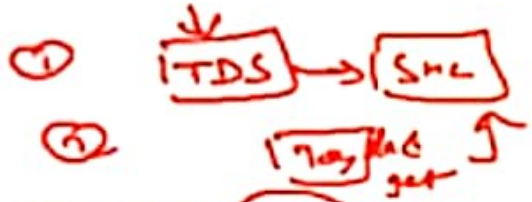
Subject Code: 20A05602T

UNIT I –Preparing to Model

Machine Learning Activities

- Preparation activities
- Detailed process of Machine Learning
- Summary of Steps and Activities Involved

Machine Learning Activities



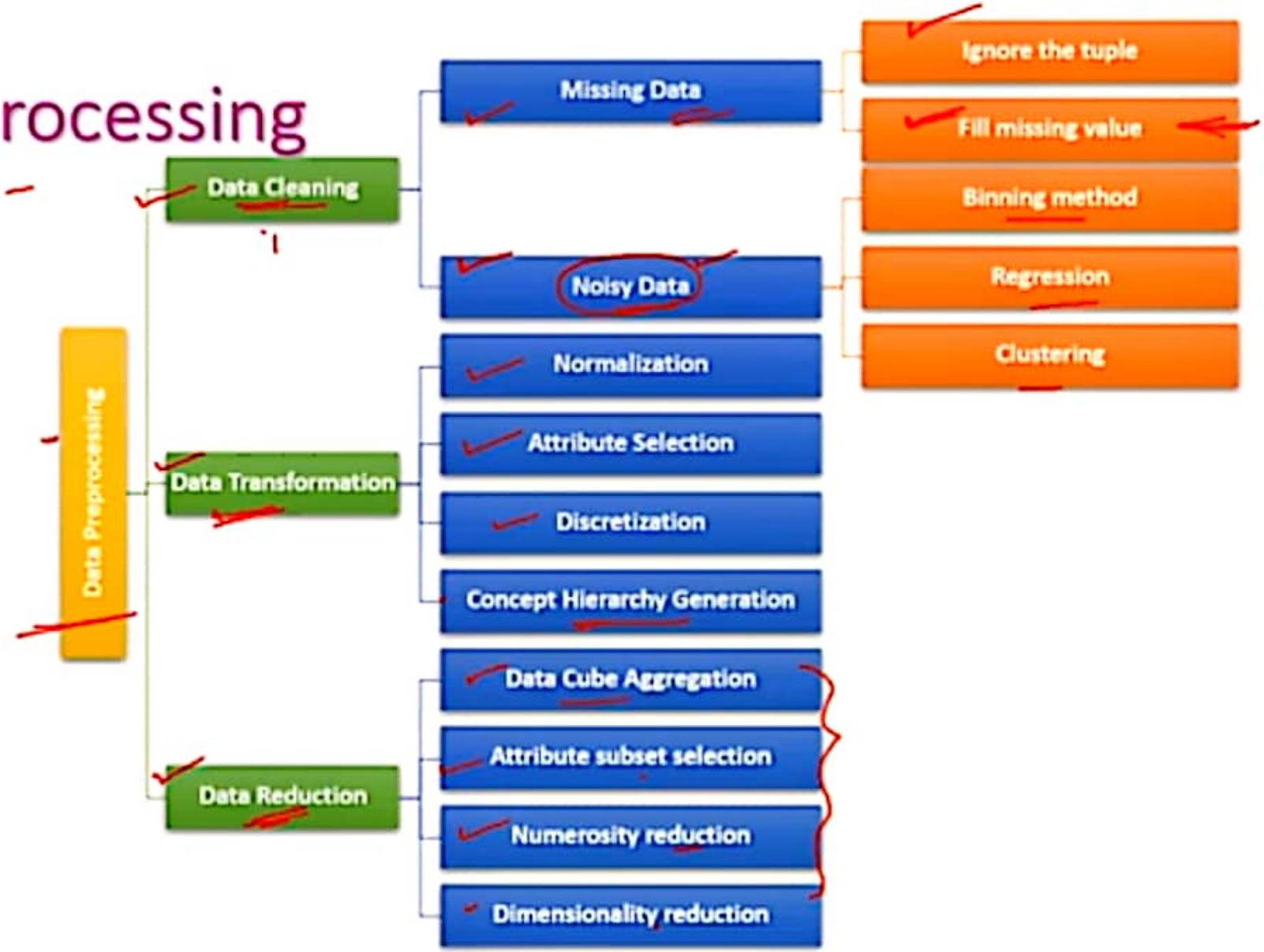
- The first step in machine learning activity starts with data.
- Supervised Machine Learning, has a labelled training data set followed by test data which is not labelled.
- Unsupervised Machine Learning, there is no labelled data (no training) but to find patterns in the input data.
- A thorough review and exploration of the data is needed to understand
  - the type of the data,
  - the quality of the data and
  - relationship between the different data elements.
- Based on that, multiple pre-processing activities to be done on the input data before the machine learning activities.

Machine Learning Activities – Preparation activities

- Following are the typical **preparation** activities done once the input data comes into the machine learning system:

1. Understand the type of data in the given input data set
2. Explore the data to understand the nature and quality.
3. Explore the relationships amongst the data elements, e.g. inter-feature relationship.
4. Find potential issues in data.
5. Do the necessary remediation, e.g. include missing data values, etc., if needed.
6. Apply pre-processing steps, as necessary.

Data Preprocessing

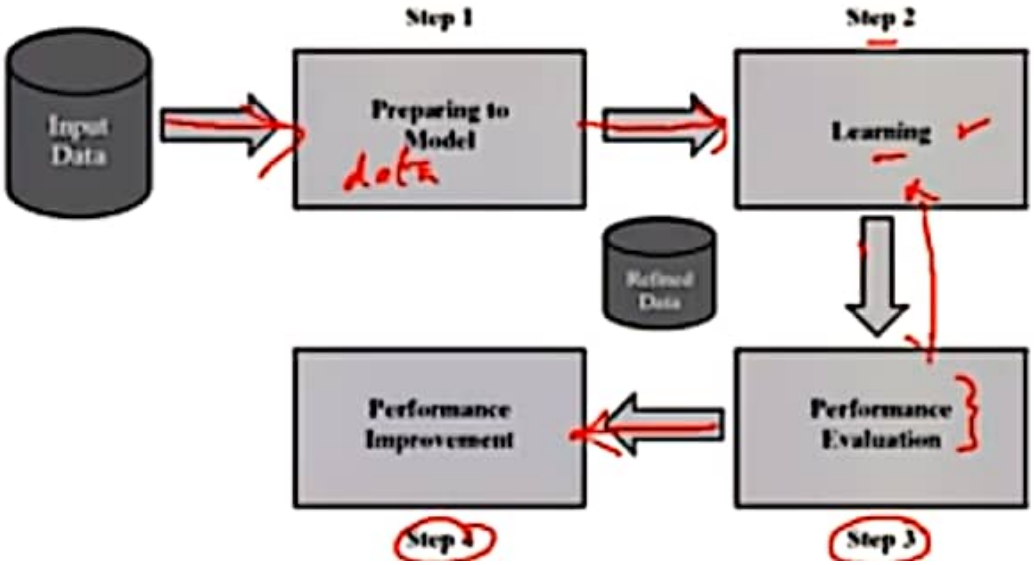


Machine Learning Activities – Preparation activities...

- **Once the data is prepared for modelling, then the learning tasks starts with the following activities.**
  1. The input data is first divided into two parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
  2. Consider different models or learning algorithms for selection.
  3. Train the model based on the training data for supervised learning problem and apply to unknown data.
  4. Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.

Detailed process of Machine Learning

- After the model is selected,
- trained (for supervised learning), and applied on input data,
- the performance of the model is evaluated.
- Based on options available, specific actions can be taken to improve the performance of the model, if possible.



Summary of Steps and Activities Involved

Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

Machine Learning

Subject Code: 20A05602T

UNIT I –Preparing to Model

Basic Types Of Data In Machine Learning

- Data Set
- Qualitative data
- 1. Nominal data
- 2. Ordinal data
- Quantitative data
- 1. Interval data
- 2. Ratio data

Data Set

- A data set is a collection of related information or records.
- The information may be on some entity or some subject area.
- Example-
- 1. A data set on students in which each record consists of information about a specific student.

Student data set:

	Roll Number	Name	Gender	Age
1	129/011	Mihir Karmarkar	M	14
2	129/012	Geeta Iyer	F	15
3	129/013	Chanda Bose	F	14
4	129/014	Sreenu Subramanian	M	14
5	129/015	Pallav Gupta	M	16
6	129/016	Gajanan Sharma	M	15

Data Set...

- 2. Another data set on student performance which has records providing performance, i.e. marks on the individual subjects.

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

Data Set...

- Each row of a data set is called a record.
- Each data set also has multiple attributes, each of which gives information on a specific characteristic.
- For example, in the data set on students, there are four attributes namely Roll Number, Name, Gender, and Age,
- Attributes can also be termed as feature, variable, dimension or field.
- Both the data sets, Student and Student Performance, are having four features or dimensions;
- hence they are told to have four dimensional data space.

Student data set:

Roll Number	Name	Gender	Age
129011	Mihir Karmarkar	M	14
129012	Geeta Iyer	F	15
129013	Chanda Bose	F	14
129014	Sreenu Subramanian	M	14
129015	Pallav Gupta	M	16
129016	Gajanan Sharma	M	15

Student performance data set:

Roll Number	Maths	Science	Percentage
129011	89	45	89.33%
129012	89	47	90.67%
129013	68	29	64.67%
129014	83	38	80.67%
129015	57	23	53.33%
129016	78	35	75.33%

Data Set...

- each row has specific values for each of the four attributes or features.
- Value of an attribute, vary from record to record.
- For example, if we refer to the first two records in the Student data set, the value of attributes Name, Gender, and Age are different.

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15

Types of Data

- Data can broadly be divided into following two types:
- 1. Qualitative data
- 2. Quantitative data

Qualitative data

- **Qualitative data** provides information about the quality of an object or information which cannot be measured.
- For example, if we consider the quality of performance of students in terms of 'Good', 'Average', and 'Poor',
- it falls under the category of qualitative data.
- Also, name or roll number of students are information that cannot be measured using some scale of measurement.
- Qualitative data is also called **categorical data**.
- Qualitative data - two types:
- 1. Nominal data
- 2. Ordinal data

Qualitative data - Nominal data

- **Nominal data** is one which has no numeric value, but a named value.
- It is used for assigning named values to attributes.
- Nominal values cannot be quantified.
- Examples of nominal data are
- 1. Blood group: A, B, O, AB, etc.
- 2. Nationality: Indian, American, British, etc.
- 3. Gender: Male, Female, Other

Qualitative data - Nominal data

- mathematical operations such as addition, subtraction, multiplication, etc. and statistical functions such as mean, variance, etc. cannot be performed on nominal data.
- a basic count is possible
- The mode is possible, i.e. most frequently occurring value, can be identified for nominal data.

Qualitative data - Ordinal data,

- **Ordinal data**, naturally ordered.
- This means ordinal data assigns named values to attributes
- They can be arranged in a sequence of increasing or decreasing value
- Hence comparison is possible here.
  - 1. Customer satisfaction: 'Very Happy', 'Happy', 'Unhappy', etc.
  - 2. Grades: A, B, C, etc.
  - 3. Hardness of Metal: 'Very Hard', 'Hard', 'Soft', etc.
- Like nominal data, basic counting is possible for ordinal data.
- Hence, the mode and median can be identified.
- But Mean can not be calculated.

Quantitative data

- **Quantitative data** relates to information about the quantity of an object – hence it can be measured.
- For example, if we consider the attribute 'marks', it can be measured using a scale of measurement.
- Quantitative data is also termed as numeric data.
- There are two types of quantitative data:
- 1. Interval data
- 2. Ratio data

Quantitative data - Interval data

- **Interval data** is numeric data – identify the order and difference between values
- Example - Celsius temperature - the difference between each value remains the same.
- For example, the difference between 12°C and 18°C degrees is measurable and is 6°C
- Other examples include date, time, etc.
- For interval data, mathematical operations such as addition and subtraction are possible.
- For that reason, for interval data, the central tendency can be measured by mean, median, or mode.
- Standard deviation can also be calculated.

Quantitative data - Interval data

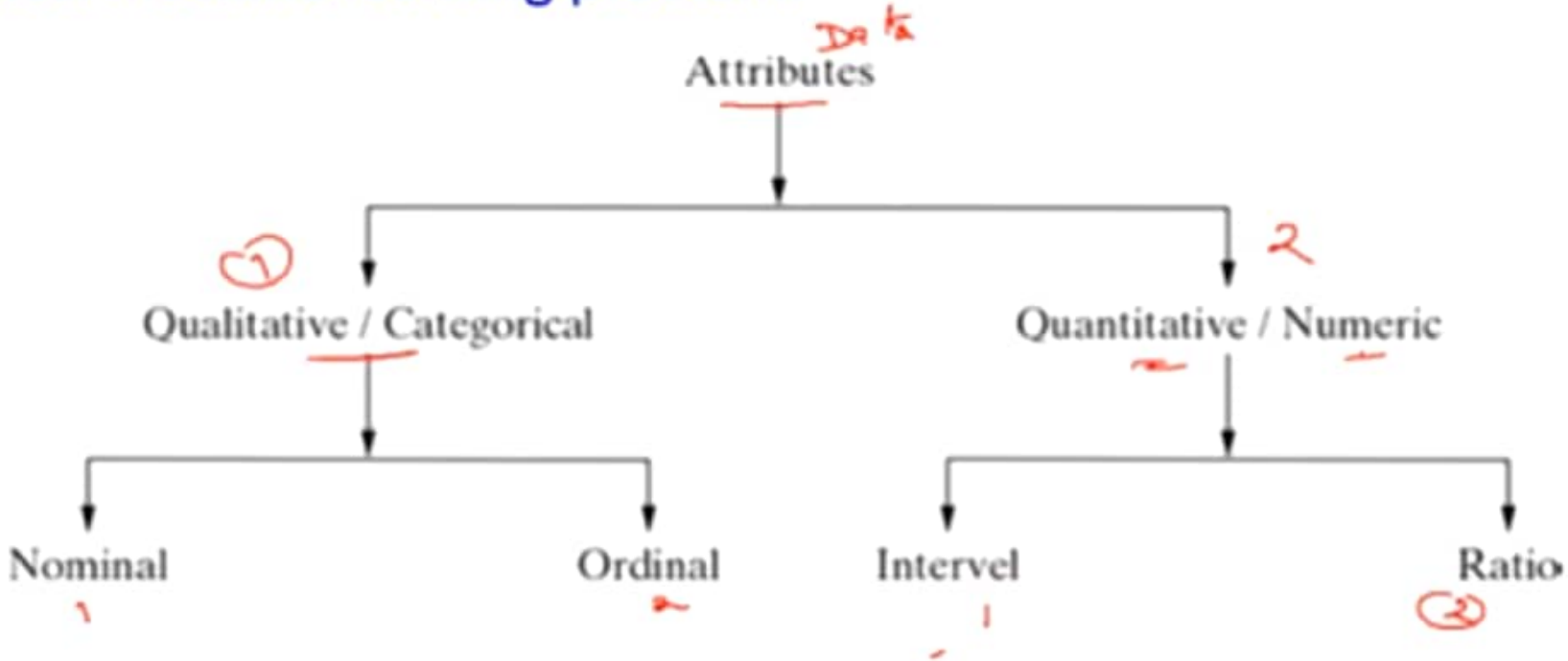
- Interval data do not have a 'true zero' value.
- For example, there is nothing called '0 temperature' or 'no temperature'.
- Hence, only addition and subtraction applies for interval data.
- The ratio cannot be applied.
- This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C.
- However, we cannot say the temperature of 40°C means it is twice as hot as in temperature of 20°C.

Quantitative data - Ratio data

- **Ratio data** represents numeric data for which exact value can be measured.
- Absolute zero is available for ratio data.
- Also, these variables can be added, subtracted, multiplied, or divided.
- The central tendency can be measured by mean, median, or mode and also standard deviation.
- Examples of ratio data include height, weight, age, salary, etc.

Data types based on Attributes.

- A summarized view of different types of data that we may find in a typical machine learning problem.



Data Attributes

- Attributes can also be categorized into types based on a number of values that can be assigned.
- The types of attributes based on factor.

1. Discrete Attributes
2. Nominal attributes
3. Numeric Attributes
4. Binary Attributes
5. Continuous Attributes

Data Attributes...

- Discrete attributes can assume a finite or countably infinite number of values
- Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values.
- Numeric attributes such as count, rank of students, etc. can have countably infinite values.
- binary attribute, a special type of discrete attribute which can assume two values only is called .
- Examples of binary attribute include male/ female, positive/negative, yes/no, etc.
- Continuous attributes can assume any possible value which is a real number.
- Examples of continuous attribute include length, height, weight, price, etc.

Machine Learning

Subject Code: 20A05602T

UNIT I –Preparing to Model

EXPLORING STRUCTURE OF DATA – Intro.

- Basic Data Types
- Data Sets - Data dictionary
- Auto MPG Data set from UCI Machine Learning Repository

mpg	cylinder	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
18	8	307	130	3504	12	70	1	Chevrolet chev-elle malibu
15	8	350	165	3693	11.5	70	1	Buick skylark 320
18	8	318	150	3436	11	70	1	Plymouth satellite
16	8	304	150	3433	12	70	1	Amc rebel sst
17	8	302	140	3449	10.5	70	1	Ford torino
15	8	429	198	4341	10	70	1	Ford galaxie 500

EXPLORING STRUCTURE OF DATA

- Two basic data types –
  - numeric ← marks, Age
  - categorical. ← grade, gender
- The attributes of data are numeric or categorical in nature.
- Exploring numeric data and categorical data are different.

Data set - Data dictionary

- A standard data set, may have the data dictionary, for reference.
- Data dictionary is a metadata repository, i.e. the repository of all information related to the structure of each data element contained in the data set.
- The data dictionary gives detailed information on each of the attributes – the description as well as the data type and other relevant details.
- If data dictionary is not available, use standard library function of the machine learning tool.
- Eg. UCI machine learning repository.
- The data set that we take as a reference is the **Auto MPG** data set available in the UCI repository.

google.com/search?q=uci+repository&domains=ics.uci.edu&sxsrf=AJOqlzX0wYwuDv8266v9jfT3HiazPrVChQ%3A1677471269404&ei=JS78Y_OwGLukz7sPsPiooAc&oq=...

To get future Google Chrome updates, you'll need Windows 10 or later. This computer is using Windows 7. Learn more

Google uci repository

News Images Dataset download Full form Videos CSV Bank dataset Dat All filters Tools

About 40,70,000 results (0.35 seconds)

https://archive.ics.uci.edu

UCI Machine Learning Repository

Welcome to the new **Repository** admins Dheeru Dua and Efi Karra Taniskidou 04-04-2013
Welcome to the new **Repository** admins Kevin Bache and Moshe Lichman! 03
You've visited this page 2 times. Last visit: 24/2/23

Data Sets

Undo - Regression - Abalone dataset - Adult Data - Business

The UC Irvine Machine ...

View Datasets - See More New Datasets - Iris - Heart Disease

About

The UCI Machine Learning Repository is a collection of

Undo

3W dataset - Basketball dataset - Acute Inflammations - Audit Data

More results from uci.edu »

UC Irvine Machine Learning Repository

The UCI Machine Learning Repository is **a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms**. 09-Sept-2021

https://www.re3data.org › repository

UCI Machine Learning Repository | re3data.org

Feedback

archive.ics.uci.edu/ml/datasets.php

To get future Google Chrome updates, you'll need Windows 10 or later. This computer is using Windows 7. Learn more

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

View ALL Data Sets

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

Browse Through: 622 Data Sets

Table View List View

Default Task
- Classification (486)
- Regression (151)
- Clustering (121)
- Other (56)

Attribute Type
- Categorical (38)
- Numerical (422)
- Mixed (55)

Data Type
- Multivariate (480)
- Univariate (30)
- Sequential (59)
- Time-Series (126)
- Text (69)
- Domain-Theory (23)
- Other (21)

Area

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
			Categorical			

archive.ics.uci.edu/ml/datasets/auto+mpg

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

View ALL Data Sets

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

Auto MPG Data Set

Download Data Folder Data Set Description

Abstract: Revised from CMU StatLib library, data concerns city-cycle fuel consumption

Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	636415

Source:

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

Data Set Information:

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg". 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original"

archive.ics.uci.edu/ml/datasets/auto+mpg

Data Set Information:

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original".

"The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)

Attribute Information:

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Relevant Papers:

Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann. [Web Link]

Papers That Cite This Data Set[1]:

Rexa.info
• Research • People • Connections

Dan Pelleg. Scalable and Practical Probability Density Estimators for Scientific Anomaly Detection. School of Computer Science Carnegie Mellon University. 2004. [View Context]

Qingping Tao Ph. D. MAKING EFFICIENT LEARNING ALGORITHMS WITH EXPONENTIALLY MANY FEATURES. Qingping Tao A DISSERTATION Faculty of The Graduate College University of Nebraska In Partial Fulfillment of Requirements. 2004. [View Context]

UCI machine learning repository – Auto MPG data set

- From the data, the attributes
- ‘mpg’, ‘cylinders’, ‘displacement’, ‘horsepower’, ‘weight’, ‘acceleration’, ‘model year’, and ‘origin’ are all numeric.
- ‘cylinders’, ‘model year’, and ‘origin’ are discrete -only finite number of values
- ‘mpg’, ‘displacement’, ‘horsepower’, ‘weight’, and ‘acceleration’ are real value.

The UCI repository.

- First few lines of
- Auto MPG data set

mpg	cylinder	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
18	8	307	130	3504	12	70	1	Chevrolet chev-elle malibu
15	8	350	165	3693	11.5	70	1	Buick skylark 320
18	8	318	150	3436	11	70	1	Plymouth satellite
16	8	304	150	3433	12	70	1	Amc rebel sst
17	8	302	140	3449	10.5	70	1	Ford torino
15	8	429	198	4341	10	70	1	Ford galaxie 500
14	8	454	220	4354	9	70	1	Chevrolet impala
14	8	440	215	4312	8.5	70	1	Plymouth fury iii
14	8	455	225	4425	10	70	1	Pontiac catalina
15	8	390	190	3850	8.5	70	1	Amc acbassador dpl
15	8	383	170	3563	10	70	1	Dodge challenger se
14	8	340	160	3609	8	70	1	Plymouth ' cuda 340
15	8	400	150	3761	9.5	70	1	Chevrolet monte carlo
14	8	455	225	3086	10	70	1	Buick estate wagon (sw)
24	4	113	95	2372	15	70	3	Toyota corona mark ii
22	6	198	95	2933	15.5	70	1	Plymouth duster
18	6	199	97	2774	15.5	70	1	Amc hornet

Auto MPG data set...

- Hence, these attributes are continuous in nature.
- The only remaining attribute 'car name' is of type categorical, or more specifically nominal.
- This data set is regarding prediction of fuel consumption in miles per gallon, i.e. the numeric attribute 'mpg' is the target attribute.
- With this understanding of the data set attributes, we can start exploring the numeric and categorical attributes separately.

EXPLORING STRUCTURE OF DATA

- Exploring numerical data
- Plotting and exploring numerical data
  - *Box plots*
  - *Histogram*
- Exploring categorical data
- Exploring relationship between variables

Machine Learning

Subject Code: 20A05602T

UNIT I –Preparing to Model

EXPLORING STRUCTURE OF DATA – Intro.

- **Exploring Numerical Data – Part-1**
- *Understanding central tendency of Numerical Data*
- Importance of Mean and Median
- *Understanding data spread*
  - 1. Dispersion of data
  - 2. Position of the different data values

	mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin
Median	23	4	148.5	?	2804	15.5	76	1
Mean	23.51	5.455	193.4	?	2970	15.57	76.01	1.573
Deviation	2.17	26.67%	23.22%		5.59%	0.45%	0.01%	36.43%
	Low	High	High		Low	Low	Low	High

Understanding central tendency of Numerical Data

- In statistics, the measures of central tendency of data are mean and median – to understand the central point of a set of data.
- Mean is a sum of all data values divided by the count of data elements.
- For example, marks of 5 students in a class are 21, 89, 34, 67, and 96
- the mean marks, 61.4.

$$\text{Mean} = \frac{21 + 89 + 34 + 67 + 96}{5} = 61.4$$

Understanding central tendency of Numerical Data...

- Median, is the value of the element appearing in the middle of an ordered list of data elements.
- If we consider the above 5 data elements, the ordered list would be – 21, 34, 67, 89, and 96.
- the 3rd element in the ordered list is considered as the median (middle value).
- Hence, the median value of this set of data is 67.

Importance of Mean and Median

- The reason is mean and median are impacted differently by data values appearing at the beginning or at the end of the range.
- Mean is difficult, if too many data elements are having values closer to the maximum or minimum values.
- It is sensitive to outliers, i.e. the values which are unusually high or low, compared to the other values.
- In certain attributes, the deviation between values of mean and median are quite high, we should investigate those attributes further and try to find out the root cause along with the need for remediation.

Importance of Mean and Median...

- Mean vs. Median for Auto MPG

	mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin
Median	23	4	148.5	?	2804	15.5	76	1
Mean	23.51	5.455	193.4	?	2970	15.57	76.01	1.573
Deviation	2.17	26.67%	23.22%		5.59%	0.45%	0.01%	36.43%
	Low	High	High		Low	Low	Low	High

- attributes 'mpg', 'weight', 'acceleration', and 'model.year' the deviation between mean and median are low i.e. outlier values are less.
- the deviation is significant for the attributes 'cylinders', 'displacement' and 'origin'.

Understanding data spread

- Two types
- 1. Dispersion of data
- 2. Position of the different data values

Measuring data dispersion

- Consider the data values of two attributes
- 1. Attribute 1 values : 44, 46, 48, 45, and 47
- 2. Attribute 2 values : 34, 46, 59, 39, and 52
- Both the set of values have a mean and median of 46.
- The attribute 1 is more concentrated or clustered around the mean/median value.
- The attribute 2 is quite spread out or dispersed.
- To measure the extent of dispersion of a data, or to find out how much the different values of a data are spread out, the variance of the data is measured.
- The variance of a data

$$(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

Measuring data dispersion...

$$(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

- where x is the variable or attribute whose variance is to be measured
- n is the number of observations or values of variable x.

$$\text{Standard deviation } (x) = \sqrt{\text{Variance } (x)}$$

- Larger value of variance or standard deviation indicates more dispersion in the data and vice versa.

Measuring data dispersion...

- For attribute 1,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44 + 46 + 48 + 45 + 47}{5}\right)^2$$

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

- attribute 1 values are quite concentrated around the mean

Measuring data dispersion…

- For attribute 2,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

46.

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

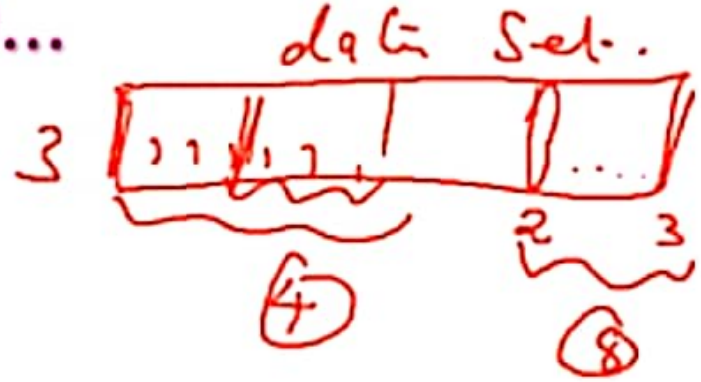
$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

- attribute 2 values are extremely spread out.

Measuring data value position

- When the data values of an attribute are arranged in an increasing order, median gives the central data value, which divides the entire data set into two halves.
- Similarly, if the first half of the data is divided into two halves so that each half consists of one quarter of the data set, then that median of the first half is known as first quartile or Q1.
- In the same way, if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q3 .
- **The overall median is also known as second quartile or Q2 .**
- So, any data set has five values -minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.
- (Example : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15, in this min=1,Q1=4, Q2=8, Q3=12 and max=15)

Measuring data value position...



- the attribute 'displacement',
  - difference between minimum value and Q1 is 36.2
  - difference between Q1 and median is 44.3.
  - the difference between median and Q3 is 113.5
  - Q3 and the maximum value is 193.
- the larger values are more spread out than the smaller ones.
- This helps in understanding why the value of mean is much higher than the median for the attribute 'displacement'.

	cylinders	displacement	origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

Measuring data value position…

- attribute 'cylinders',
  - the difference between minimum value and median is 1
  - the difference between median and the maximum value is 4.
- attribute 'origin',
  - the difference between minimum value and median is 0
  - the difference between median and the maximum value is 2.

	cylinders	displacement	origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

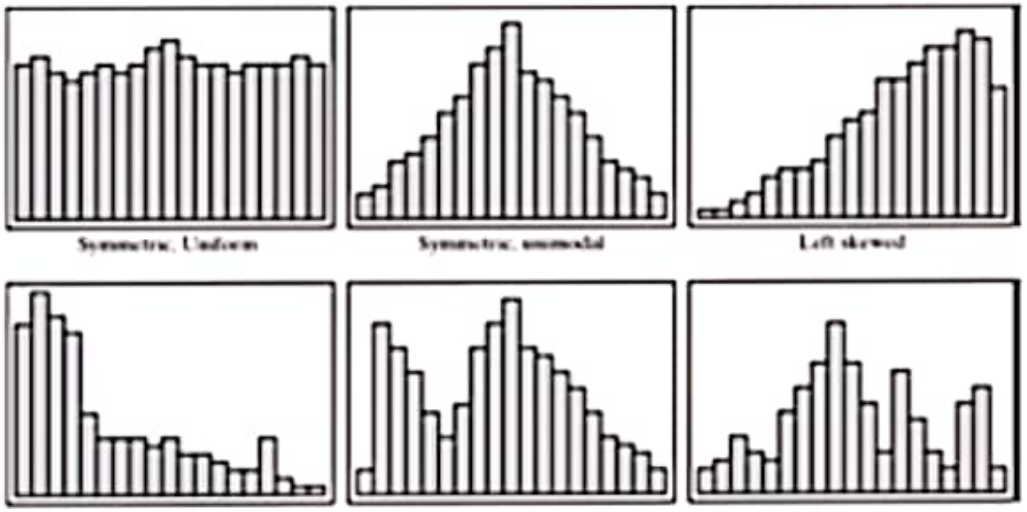
Machine Learning

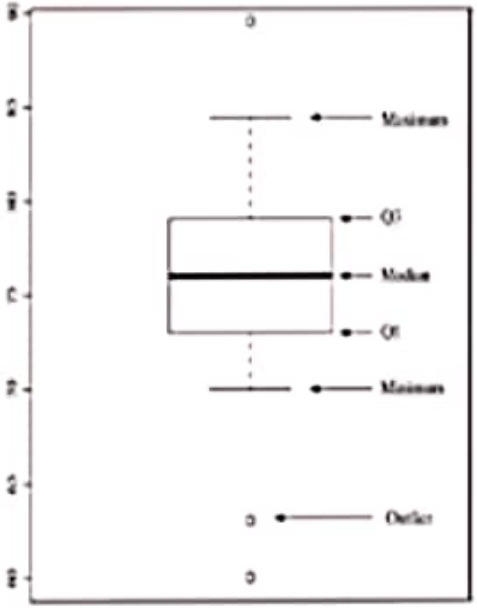
Subject Code: 20A05602T

UNIT I –Preparing to Model

EXPLORING STRUCTURE OF DATA – Intro.

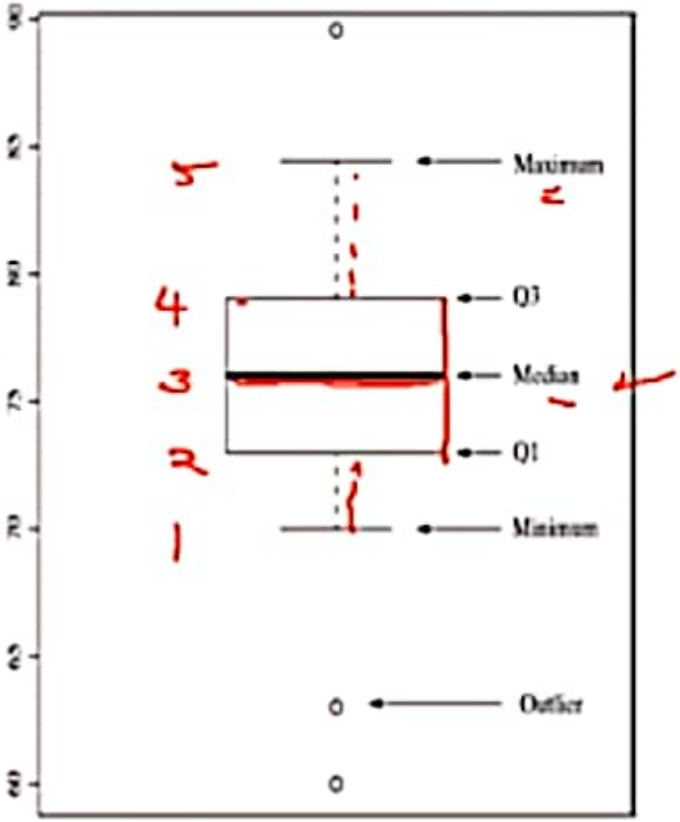
- **Exploring Numerical Data – Part-2**
- **Mathematical plots to explore numerical data**
- **Box Plot**
- **Histogram**





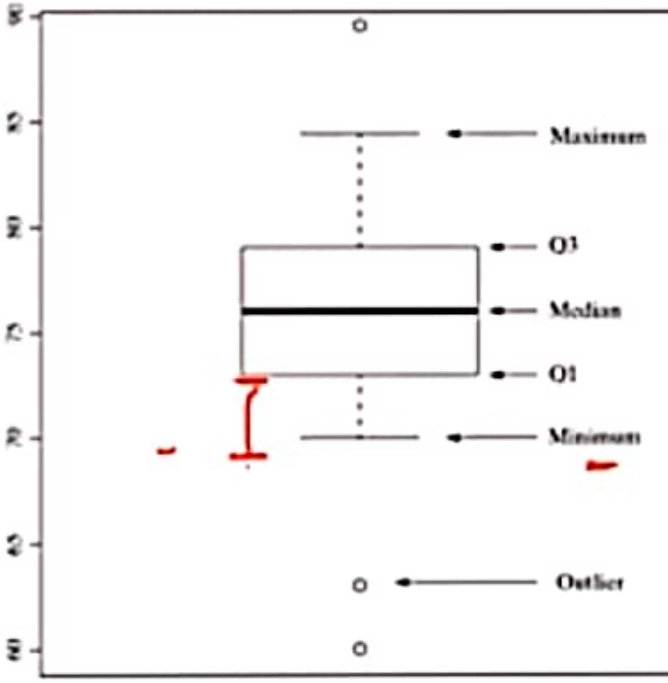
Exploring Numerical Data

- There are two most effective mathematical plots to explore numerical data – box plot and histogram.
- Box plot is an excellent visualization medium for numeric data and easy to identify if there is any outlier present in the data.
- A whisker plot—also called a box plot—displays the five-number summary of a set of data.
- The five-number summary is the minimum, first quartile, median, third quartile, and maximum.
- In a box plot, we draw a box from the first quartile to the third quartile.
- The whiskers go from each quartile to the minimum or maximum.

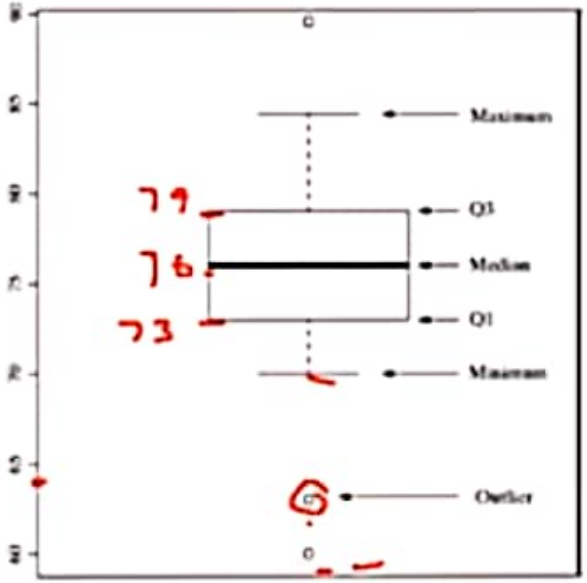


Box Plot...

- The lower whisker extends up to 1.5 times of the inter-quartile range (or IQR) from the bottom of the box, i.e. the first quartile or Q1.
- the actual length of the lower whisker depends on the lowest data value that falls within
- (Q1 – 1.5 times of IQR).



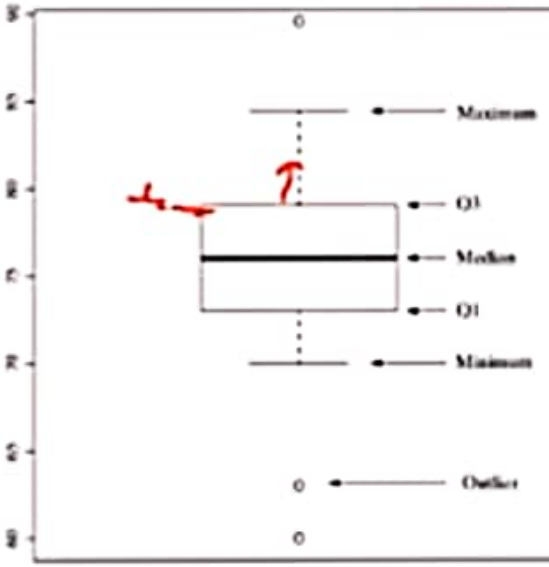
Box Plot...



- Example.
- In a set of data, Q1 = 73, median = 76 and Q3 = 79.
- Hence, IQR will be 6 (i.e. Q3 – Q1).
- So, lower whisker can extend maximum till
- (Q1 – 1.5 × IQR) = 73 – 1.5 × 6 = 64.
- there are lower range data values such as 70, 63, and 60.
- So, the lower whisker will come at 70 as this is the lowest data value larger than 64.

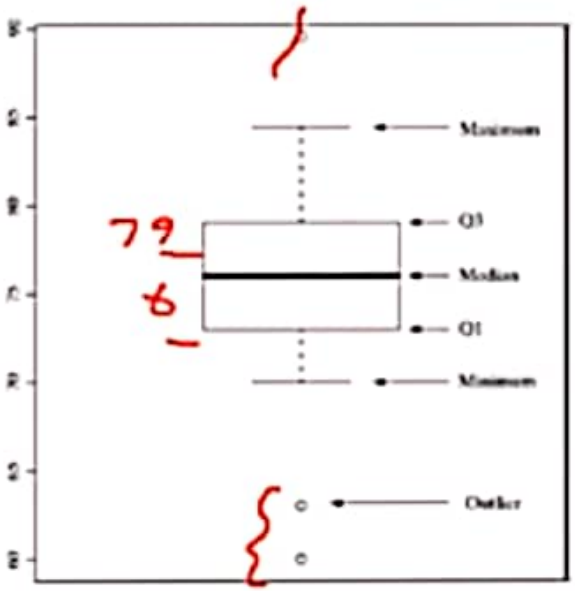
Box Plot...

- The upper whisker extends up to 1.5 as times of the inter-quartile range (or IQR) from the top of the box, i.e. the third quartile or Q3.
- the actual length of the upper whisker will also depend on the highest data value that falls within
- (Q3 + 1.5 times of IQR).



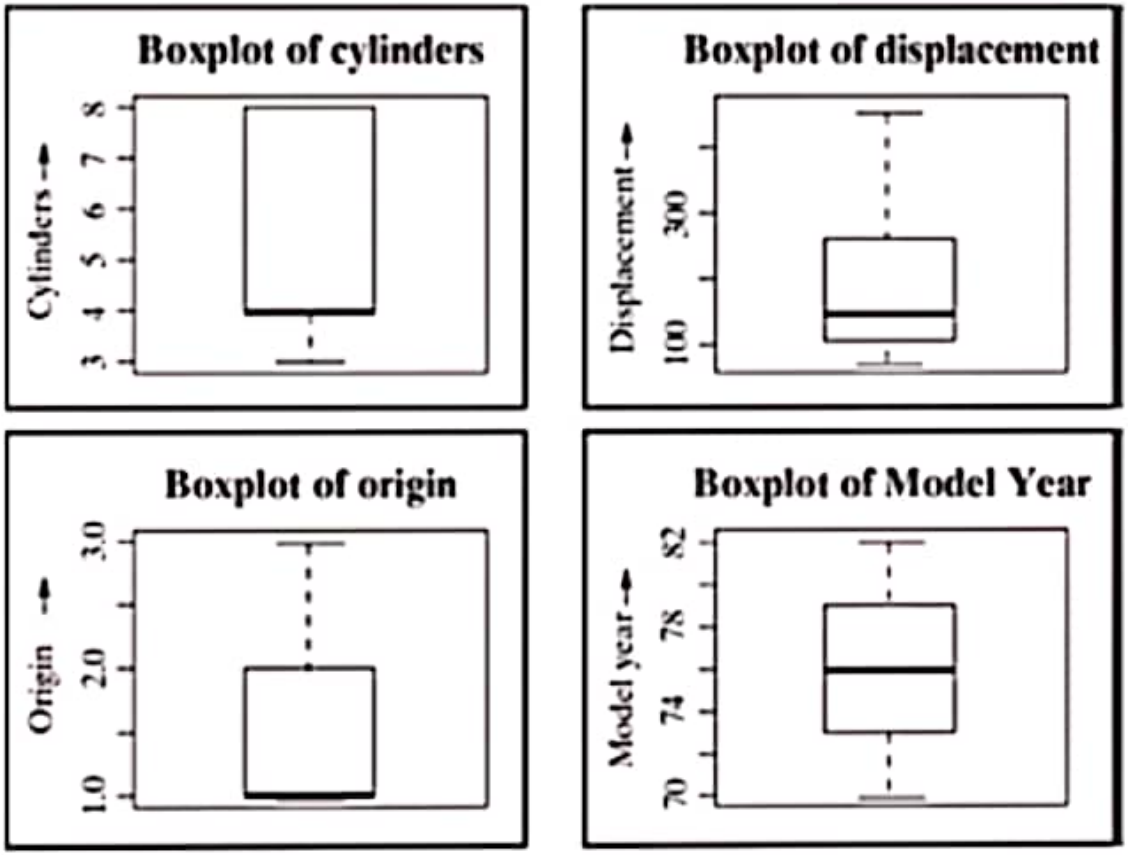
Box Plot...

- Example.
- upper whisker can extend maximum till
- (Q3 + 1.5 × IQR) = 79 + 1.5 × 6 = 88.
- If there is higher range of data values like 82, 84, and 89.
- So, the upper whisker will come at 84 as this is the highest data value lower than 88.
- The data values coming beyond the lower or upper whiskers are the ones which are of unusually low or high values respectively.
- These are **the outliers,** which may deserve special consideration.



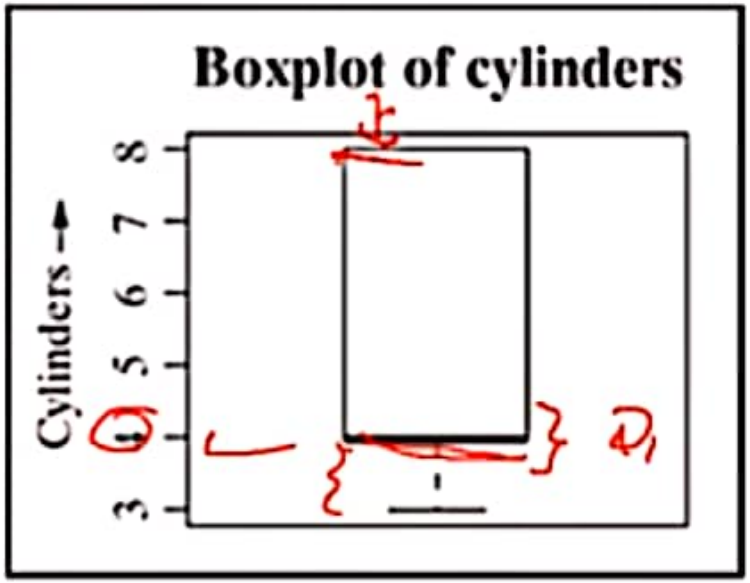
Box Plot...

- Let's visualize the box plot for the three attributes - 'cylinders', 'displacement', and 'origin'.



The box plot for attribute 'cylinders'

- The box plot for attribute 'cylinders' looks pretty weird in shape.
- The upper whisker is missing, the band for median falls at the bottom of the box, even the lower whisker is pretty small compared to the length of the box.
- The attribute 'cylinders' is **discrete** in nature having values from 3 to 8.



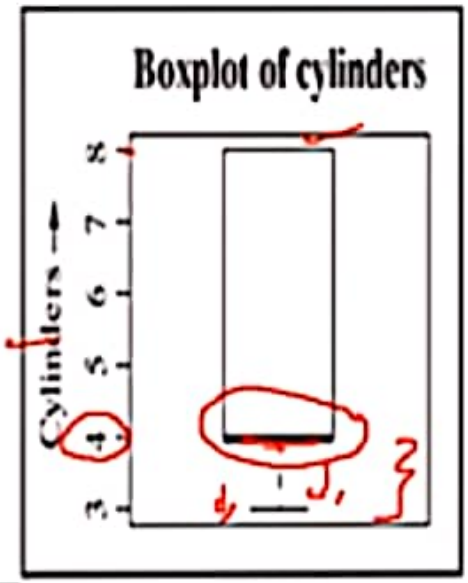
The box plot for attribute 'cylinders'

- Table captures the frequency and cumulative frequency of it.
- the frequency is extremely high for data value 4. Two other data values where the frequency is quite high are 6 and 8.
- Now find the quartiles, since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and 100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively. This way Q1 = 4, median = 4 and Q3 = 8.

Cylinders	Frequency	Cumulative Frequency
3	4	4
4	204	208 (= 4 + 204)
5	3	211 (= 208 + 3)
6	84	295 (= 211 + 84)
7	0	295 (= 295 + 0)
8	103	398 (= 295 + 103)

The box plot for attribute 'cylinders'

- Since there is no data value beyond 8, there is no upper whisker.
- Also, since both Q1 and median are 4, the band for median falls on the bottom of the box.
- Same way, though the lower whisker could have extended till -2 (Q1 – 1.5 × IQR = 4 – 1.5 × 4 = –2), in reality, there is no data value lower than 3.
- Hence, the lower whisker is also short.
- In any case, a value of cylinders less than 1 is not possible.

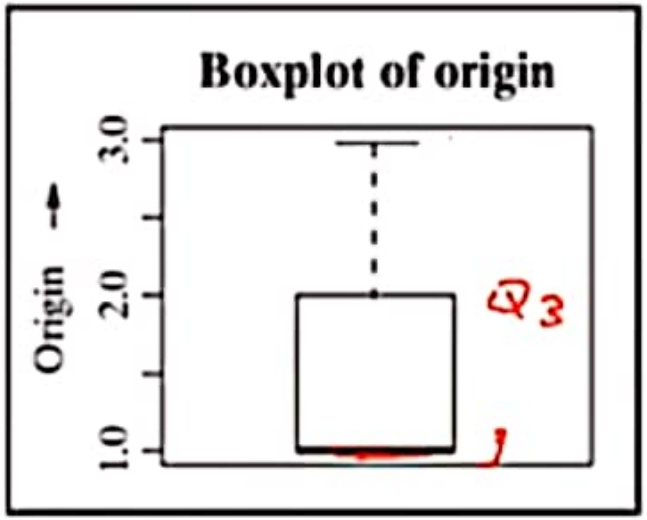


Cylinders	Frequency	Cumulative Frequency
3	4	4
4	204	208 (= 4 + 204)
5	3	211 (= 208 + 3)
6	84	295 (= 211 + 84)
7	0	295 (= 295 + 0)
8	103	398 (= 295 + 103)

Analyzing box plot for 'origin'

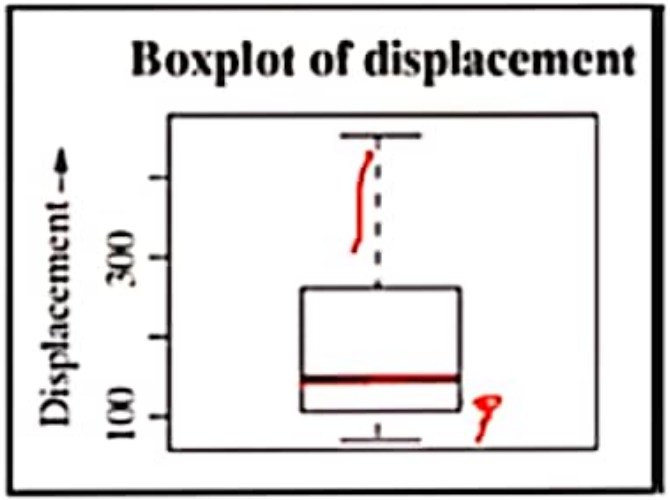
- attribute 'origin' is discrete in nature having values from 1 to 3.
- Table captures the frequency and cumulative frequency (i.e. a summation of frequencies of all previous intervals) of it.

origin	Frequency	Cumulative Frequency
1	249	249
2	70	319 (= 249 + 70)
3	79	398 (= 319 + 79)



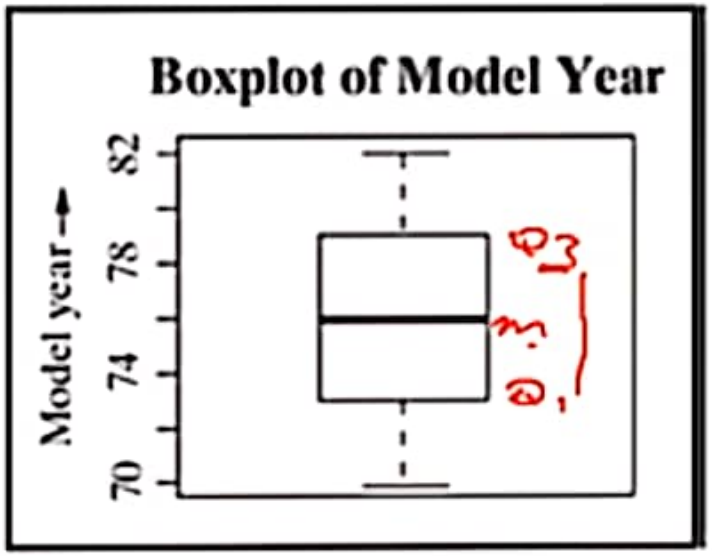
Analyzing box plot for 'displacement'

- The box plot for the attribute 'displacement' looks better than the previous box plots.
- But few small abnormalities, which needs to be reviewed.
- The lower whisker is much smaller than an upper whisker.
- Also, the band for median is closer to the bottom of the box.



Analyzing box plot for 'model Year'

- The box plot for the attribute 'model. year' looks perfect.
- First quartile, Q1 = 73
- Median, Q2 = 76
- Third quartile, Q3 = 79
- So, the difference between median and Q1 is exactly equal to Q3 and median (both are 3).
- the median is exactly equidistant from the bottom and top of the box.

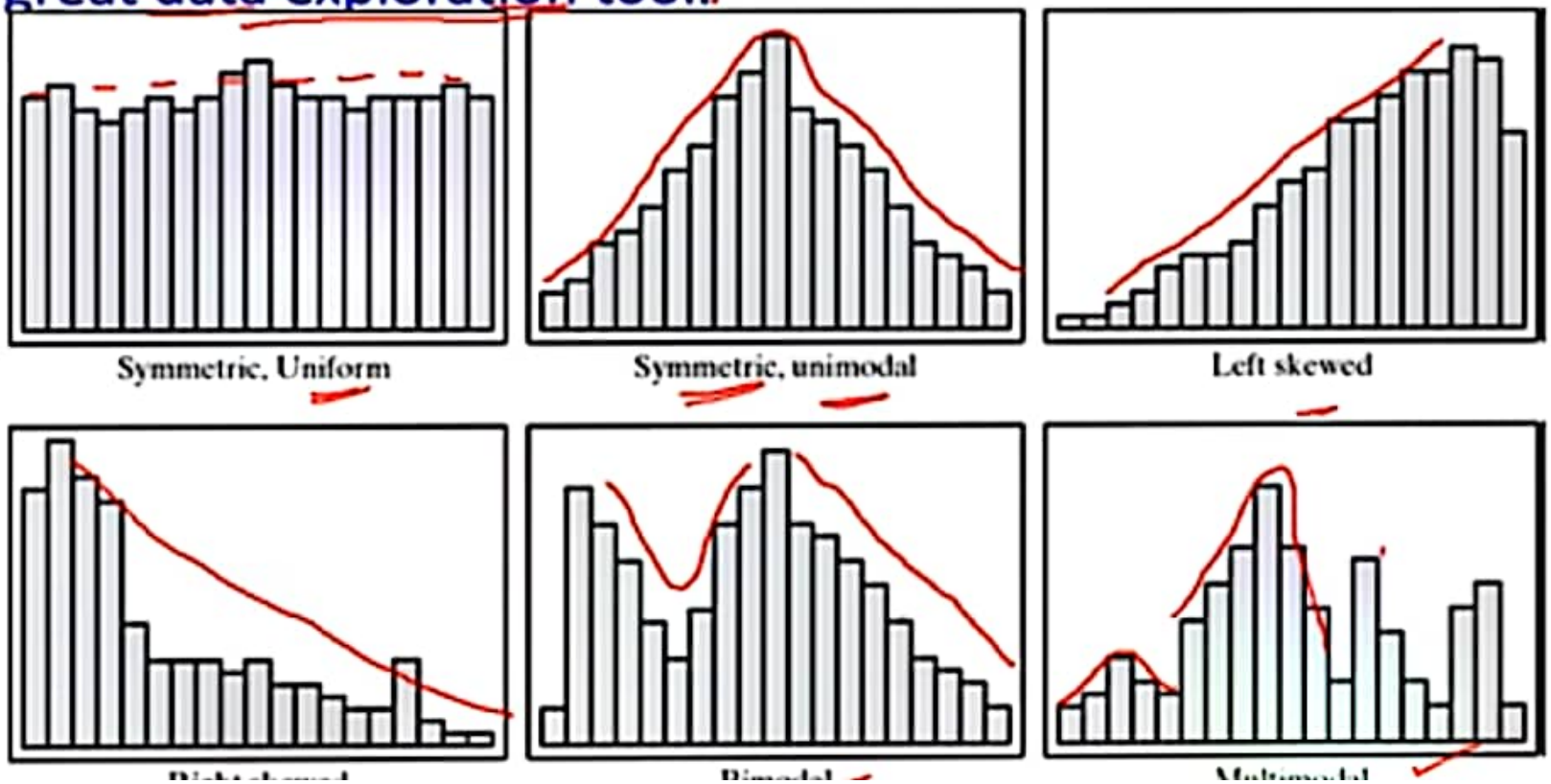


Plotting and exploring numerical data - *Histogram*

- It helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'.
- The histogram is composed of a number of bars, one bar appearing for each of the 'bins'.
- The height of the bar reflects the total count of data elements whose value falls within the specific bin value, or the frequency.
- Here, the data are visualized like bar chart.

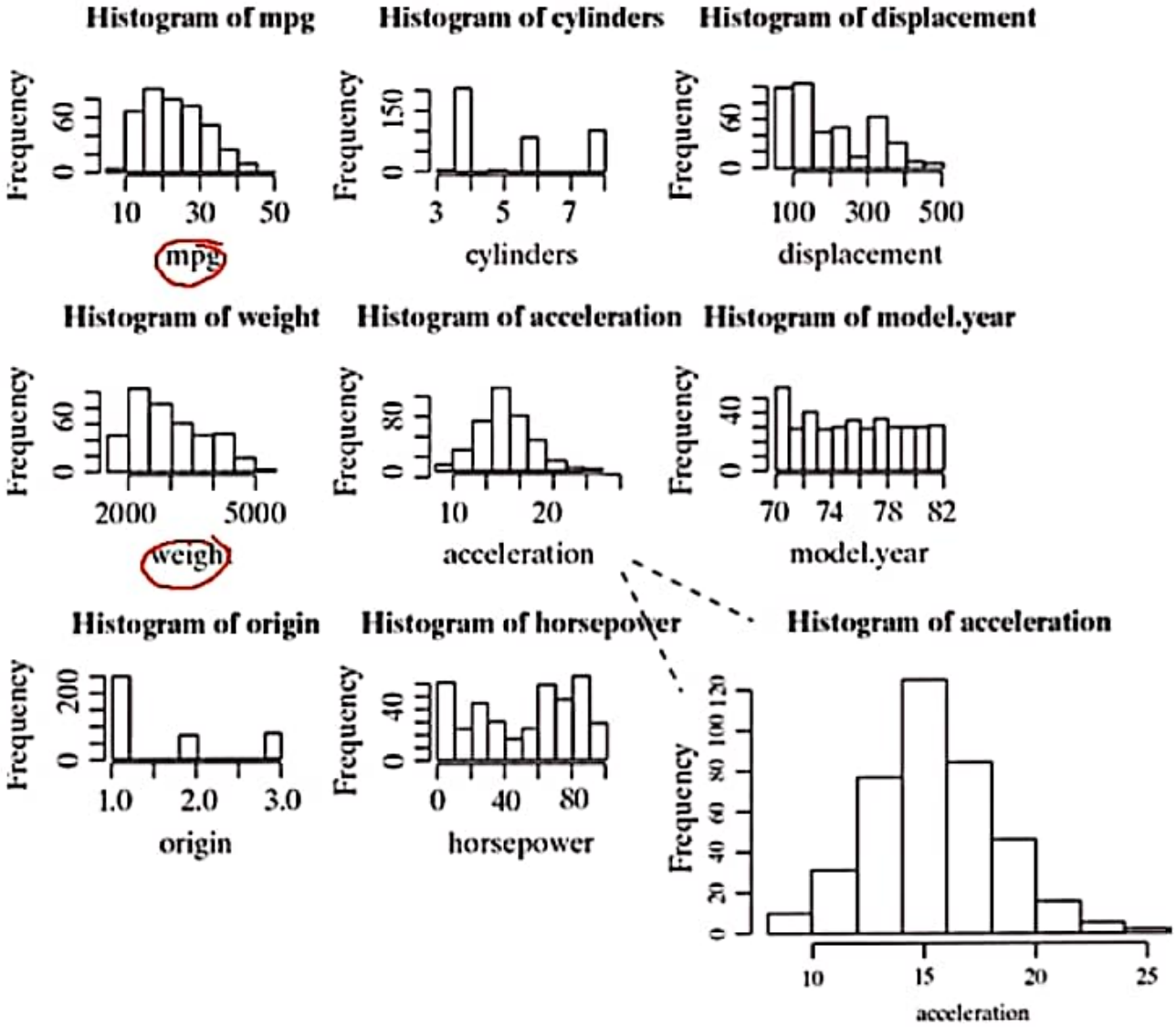
Histograms

- Histograms might be of different shapes depending on the nature of the data, e.g. skewness.
- These patterns give us a quick understanding of the data and thus act as a great data exploration tool..



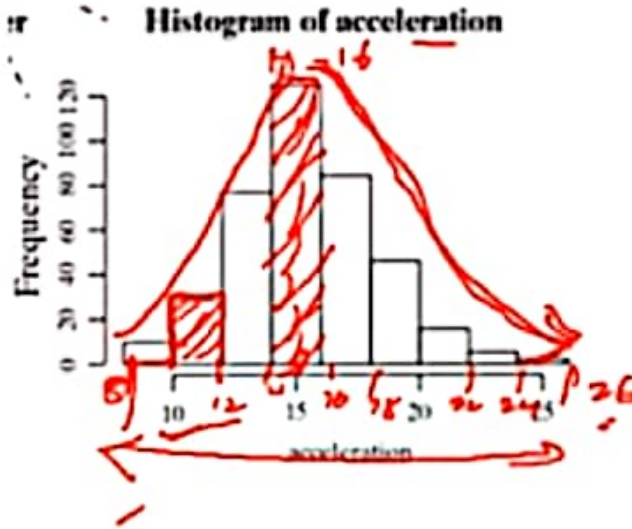
Histograms…

- The histograms for ‘mpg’ and ‘weight’ are right-skewed.
- The histogram for ‘acceleration’ is symmetric and unimodal,
- the one for ‘model.year’ is symmetric and uniform.
- For the remaining attributes, histograms are multimodal in nature.

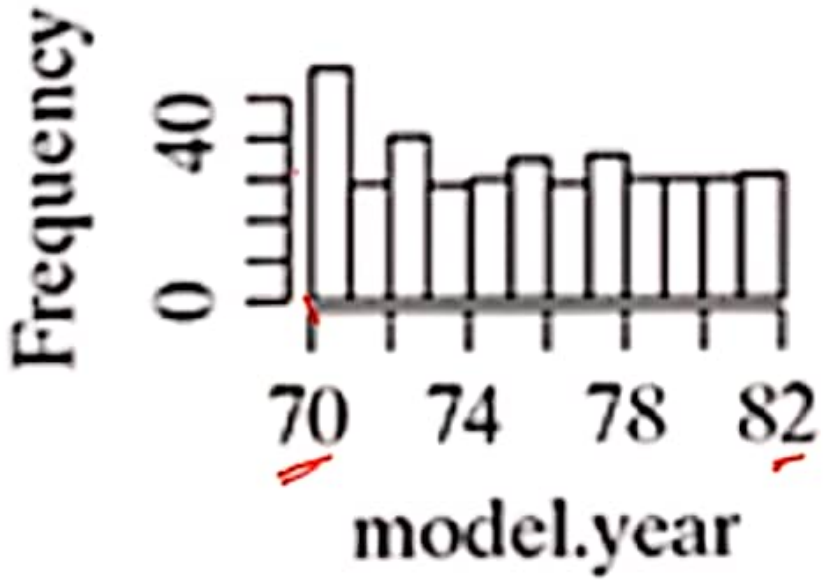


The histogram of attribute 'acceleration'.

- each 'bin' represents an acceleration value interval of 2 units.
- So the second bin, reflects acceleration value of 10 to 12 units.
- The corresponding bar chart height reflects the count of all data elements whose value lies between 10 and 12 units.
- Also, it is evident from the histogram that it spans over the acceleration value of 8 to 26 units.
- The frequency of data elements corresponding to the bins first keep on increasing, till it reaches the bin of range 14 to 16 units.
- At this range, the bar is tallest in size. So we can conclude that a maximum number of data elements fall within this range.
- After this range, the bar size starts decreasing till the end of the whole range at the acceleration value of 26 units.



- attribute ‘model. year’, it gives a hint that all values are equally likely to occur. (Uniform)



Thank You

- UNIT I –Preparing to Model
- EXPLORING STRUCTURE OF DATA – Intro.
  - Exploring Numerical Data – Part-2
- Mathematical plots to explore numerical data
- Box Plot
- Histogram

Machine Learning

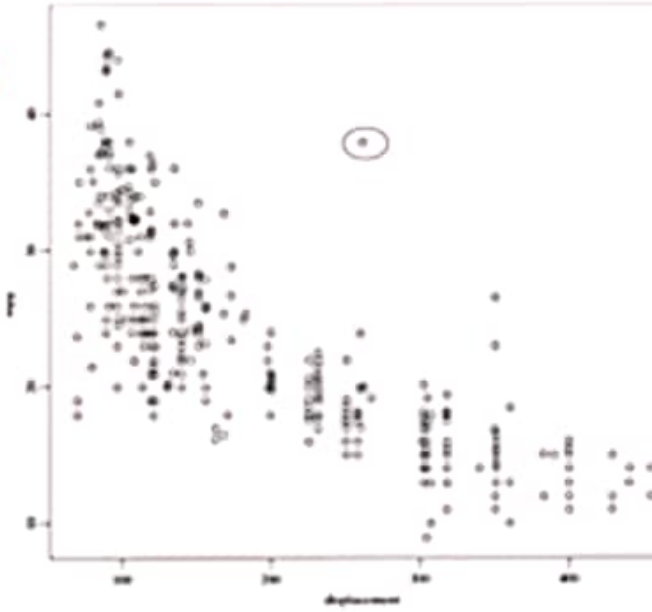
Subject Code: 20A05602T

UNIT I –Preparing to Model

EXPLORING STRUCTURE OF DATA – Intro.

- **Exploring Categorical Data**
- **Exploring relationship between variables**
  - *Scatter plot*
  - *Two-way cross-tabulations*

Cylinders \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
3	0	0	1	1	0	0	0	1	0	0	1	0	0
4	7	13	14	11	15	12	15	14	17	12	25	21	28
5	0	0	0	0	0	0	0	0	1	1	1	0	0
6	4	8	0	8	7	12	10	5	12	6	2	7	3
8	18	7	13	20	5	6	9	8	6	10	0	1	0



Categorical data

- Categorical data is used to group the information with similar characteristics. Example : Gender.
- Categorical variables can be divided into two categories:
- Nominal: no particular order between values
- Ordinal: there is some order between values

R S W C

S R W C

Categorical data...

- mean and median cannot be applied for categorical variables,
- **mode** is only applicable (frequency).
- An attribute may have one or more modes.
- Frequency distribution of an attribute having
  - Single mode is called 'unimodal',
  - Two modes are called 'bimodal' and
  - Multiple modes are called 'multimodal'.

Exploring categorical data

- In the Auto MPG data set, attribute 'car.name' is categorical in nature.
- For attribute 'car name'
  - 1. Chevrolet chevelle malibu
  - 2. Buick skylark 320
  - 3. Plymouth satellite
  - 4. Amc rebel sst
  - 5. Ford torino
  - 6. Ford galaxie 500
  - 7. Chevrolet impala
  - 8. Plymouth fury iii
  - 9. Pontiac catalina
  - 10. Amc ambassador dpl

Exploring categorical data…

- Also, as we discussed earlier, we may consider 'cylinders' as a categorical variable instead of a numeric variable.
- For attribute 'cylinders' the values are 8 4 6 3 5

- *Count of Categories for 'Cylinders' Attribute*

Attribute Value	3	4	5	6	8
Count	4	204	3	84	103

Exploring categorical data…

- for the attributes 'cylinders',
- the proportion of data elements belonging to the category 4 is
- 204 ÷ 398 = 0.513, i.e. 51.3%.

- *Proportion of Categories for "Cylinders" Attribute*

Attribute Value	3	4	5	6	8
Count	0.01	0.513	0.008	0.211	0.259

Exploring relationship between variables

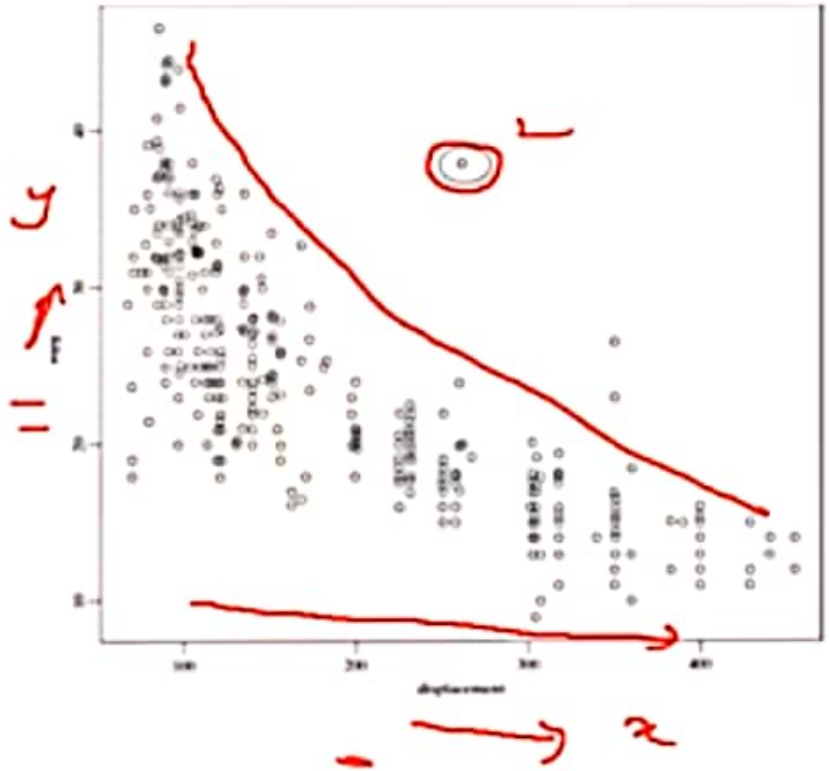
- **Basic types are**
  - *Scatter plot*
  - *Two-way cross-tabulations*

Scatter plot

- A scatter plot helps in visualizing bivariate relationships, i.e. relationship between two variables.
- It is a two dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes.
- For example, in a data set there are two attributes – attr_1 and attr_2.
- the relationship between two attributes,
- i.e. with a change in value of attr_1, then the value of the attr_2, changes.
- We can draw a scatter plot, with attr_1 mapped to *x*-axis and attr_2 mapped in *y*-axis.
- As in a two-dimensional plot, attr_1 is said to be the independent variable and attr_2 as the dependent variable.

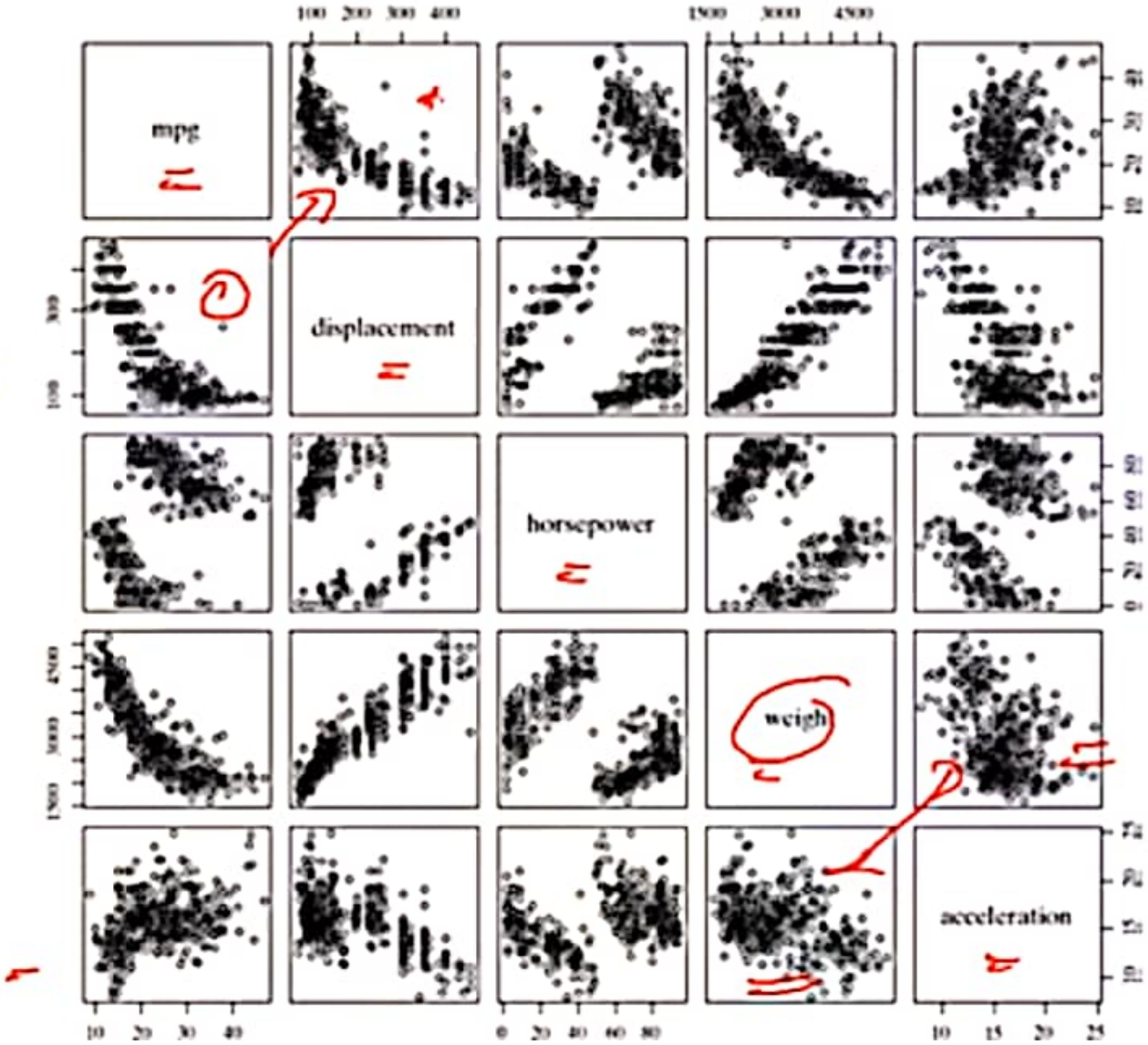
The scatter plot…

- Auto MPG the relation between
- the attributes 'displacement' and 'mpg'.
- 'displacement' as the *x*-coordinate and
- 'mpg' as the *y*-coordinate.
- The value of 'mpg' seems to steadily decrease with the increase in the value of 'displacement'.
- calculating the correlation between the variables, compute the relationship
- This gives an indication that of presence of outlier data values.



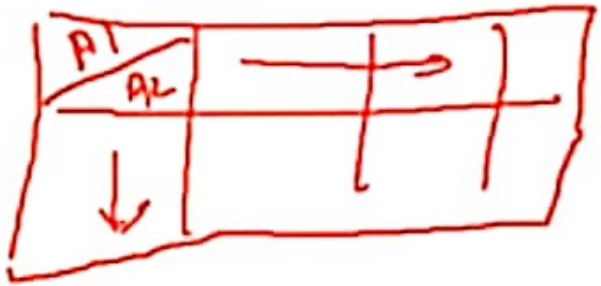
The scatter plot…

- the pair wise relationship among the features – ‘mpg’, ‘displacement’, ‘horsepower’, ‘weight’, and ‘acceleration’ have been captured.
- In most of the cases, there is a significant relationship between the attribute pairs.
- e.g. between attributes ‘weight’ and ‘acceleration’, the relationship doesn’t seem to be very strong.



Two-way cross-tabulations

- Two-way cross-tabulations (also called cross-tab or contingency table) are used to understand the relationship of two categorical attributes in a concise way.
- It has a matrix format that presents a summarized view of the bivariate frequency distribution.
- A cross-tab, helps to understand how much the data values of one attribute changes, with respect to another attribute.



Two-way cross-tabulations...

- Attribute 'model.year' captures the model year of each of the car from the year 70 to 82
- Attribute 'origin' gives the region of the car, the values for origin 1, 2, and 3 corresponding to North America, Europe, and Asia.

Origin \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
1	22	20	18	29	15	20	22	18	22	23	7	13	20
2	5	4	5	7	6	6	8	4	6	4	9	4	2
3	2	4	5	4	6	4	4	6	8	2	13	12	9

- The attributes 'cylinders', with 'origin' and 'model.year',
- **'Cylinders' vs. 'Origin'**

Cylinders \ Origin	1	2	3
3	0	0	4
4	72	63	69
5	0	3	0
6	74	4	6
8	103	0	0

- **'Cylinders' vs. 'Model year'**

Cylinders \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
3	0	0	1	1	0	0	0	1	0	0	1	0	0
4	7	13	14	11	15	12	15	14	17	12	25	21	28
5	0	0	0	0	0	0	0	0	1	1	1	0	0
6	4	8	0	8	7	12	10	5	12	6	2	7	3
8	18	7	13	20	5	6	9	8	6	10	0	1	0

Machine Learning

Subject Code: 20A05602T

UNIT I –Preparing to Model

DATA QUALITY AND REMEDIATION

- Data quality
- Data remediation
  - Handling outliers
  - Handling missing values
    - Eliminate records
    - Imputing missing values
    - Estimate missing values

mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord dl

Data Quality

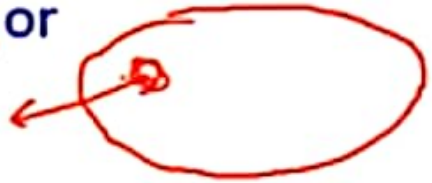
- A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning.
- It is not realistic to expect that the data will be perfect.
- two types of problems:
  - 1. Certain data elements without a value or data with a missing value.
  - 2. Data elements having value different from the other elements, which we term as outliers.
- There are multiple factors which lead to these data quality issues.
  - **Incorrect sample set selection**
  - **Errors in data collection**

Data Quality Issues - Incorrect sample set selection

- The data may not reflect normal or regular quality due to incorrect selection of sample set.

- Example 1,
  - if we are selecting a sample set of sales transactions from a festive period and trying to use that data to predict sales in future.
  - Then the prediction will be far apart from the actual scenario, just because the sample set has been selected in a wrong time.
- Example 2,

  - if we are trying to predict poll results using a training data which doesn't comprise of a right mix of voters from different segments such as **age, sex, ethnic diversities, etc.,** the prediction is bound to be a failure.
  - It may also happen due to incorrect sample size.
- a sample of small size may not be able to capture all aspects or information needed for right learning of the model.



Data Quality Issues - Errors in data collection

- **Outliers**
- In many cases, a person or group of persons are responsible for the collection of data to be used in a learning activity.
- In this manual process, there is the possibility of wrongly recording data
  - in terms of value (say 20.67 is wrongly recorded as 206.7 or 2.067)
  - in terms of a unit of measurement (say cm. is wrongly recorded as m. or mm.).
- The data elements which have abnormally high or low value from other elements are termed as *outliers*.

Data Quality Issues - Errors in data collection...

- **Missing**
- It may also happen that the data is not recorded at all.
- In case of a survey conducted to collect data, people may not response for certain question.
- So the data value for that data element in that responder's record is *missing*.

Data Remediation

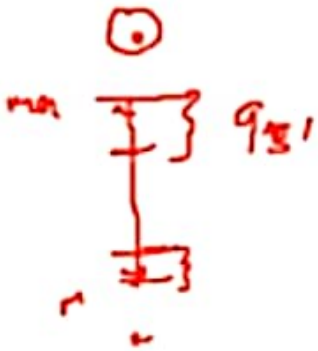
- The issues in data quality, need to be remediated,
- The right amount of efficiency to be achieved in the learning activity.
- Outliers can be remedied by proper sampling technique.
- However, human errors are bound to happen, no matter whatever checks and balances we put in.
- the missing data, a proper remedial steps need to be taken.

Data remediation -*Handling outliers*

- Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models.
- Once the outliers are identified and the decision has been taken to alter those values.
- if the outliers are natural, i.e. the value of the data element is surprisingly high or low because of a valid reason, then we should not alter it.

Outliers...

1. **Remove outliers:** If the number of records which are outliers is not many, a simple approach may be to remove them.
2. **Imputation:** assign the outlier value with mean or median or mode.
   - The value of the most similar data element may also be used for imputation.
3. **Capping:** For values that lie outside the 1.5|×| IQR limits, we can cap them by replacing those observations
   - below the lower limit with the value of 5th percentile and
   - lie above the upper limit, with the value of 95th percentile.



Outliers...

- If there is a significant number of outliers, they should be treated separately in the statistical model.
- In that case, the groups should be treated as two different groups.
- The model should be built for both groups and then the output can be combined.

Data remediation - *Handling missing values*

- In a data set, one or more data elements may have missing values in multiple records.
- It can be caused by
  - omission on part of the surveyor or
  - a person who is collecting sample data or
  - by the responder, primarily due to his/her unwillingness to respond or
  - lack of understanding needed to provide a response.
- It may happen that a specific is not applicable to a person or object with respect to which data is collected.

Data remediation - *Handling missing values...*

- There are multiple strategies to handle missing value of data elements.
  - *Eliminate records having a missing value of data elements*
  - *Imputing missing values*
  - *Estimate missing values*

Eliminate records having a missing value of data elements.

- In case the proportion of data elements having missing values is within a tolerable limit, a simple but effective approach is to remove the records having such data elements.
- This will not be possible if the proportion of records having data elements with missing value is really high.
- This will reduce the power of model because of reduction in the training data size.

Eliminate records having a missing value of data elements...

- Auto MPG data set, only in 6 out of 398 records, the value of attribute 'horsepower' is missing.
- So, we can very well eliminate the records and keep working with the remaining data set.

mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord dl

Imputing missing values

- Imputation is a method to assign a value to the data elements having missing values.
- Mean/mode/median is most frequently assigned value.
- For quantitative attributes, all missing values are imputed with the mean, median, or mode of the remaining values under the same attribute.
- For qualitative attributes, all missing values are imputed by the mode of all remaining values of the same attribute.

Imputing missing values...

- For example, the attribute 'horsepower' of the Auto MPG data set, is quantitative,
- we take a mean or median of the remaining data element values and assign that to all data elements having a missing value.
- So, we may assign the mean-104.47 to all the six data elements.
- The other approach is that we can take a similarity based mean or median.

mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord dl

Imputing missing values...

- 'cylinders' is the attribute which is logically most connected to 'horsepower'
- The increase in number of cylinders of a car, the horsepower of the car is expected to increase.

mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord dl

Imputing missing values...

- for five observations, we can use the mean of data elements of the 'horsepower' attribute having cylinders = 4; i.e. 78.28
- for one observation which has cylinders = 6, we can use a similar mean of data elements with cylinders = 6, i.e. 101.5,
- to impute value to the missing data elements.

mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	78.28	2046	19	71	1	Ford pinto
21	6	200	101.5	2875	17	74	1	Ford maverick
40.9	4	85	78.28	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	78.28	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord dl

Estimate missing values

- If there are data points similar to the ones with missing attribute values, then the attribute values from those similar data points can be planted in place of the missing value, by using the distance function.
- For example, in as student dataset, Russian student having age 12 years and height 5 ft., the weight is missing.
- Then the weight of any other Russian student having age close to 12 years and height close to 5 ft. can be assigned.

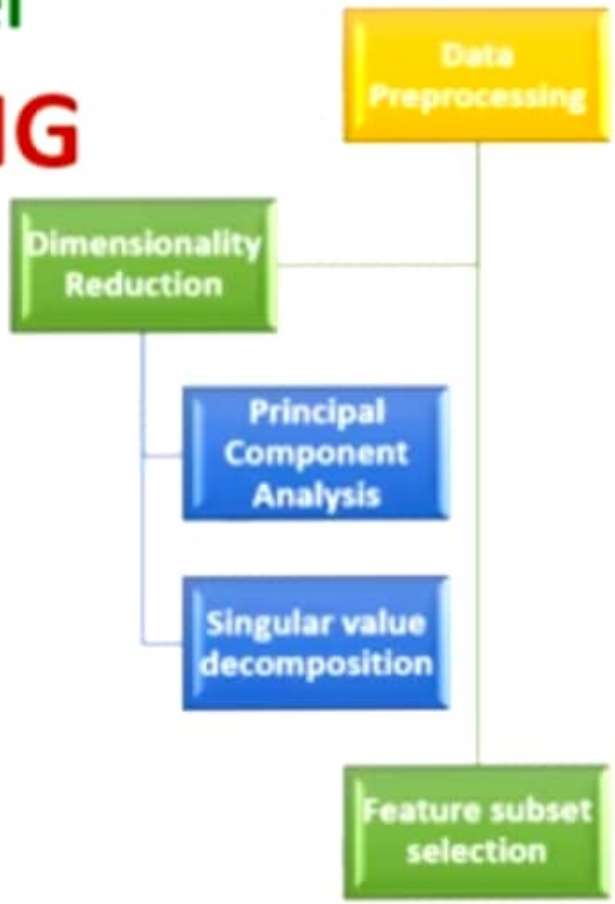
Machine Learning

Subject Code: 20A05602T

UNIT I –Preparing to Model

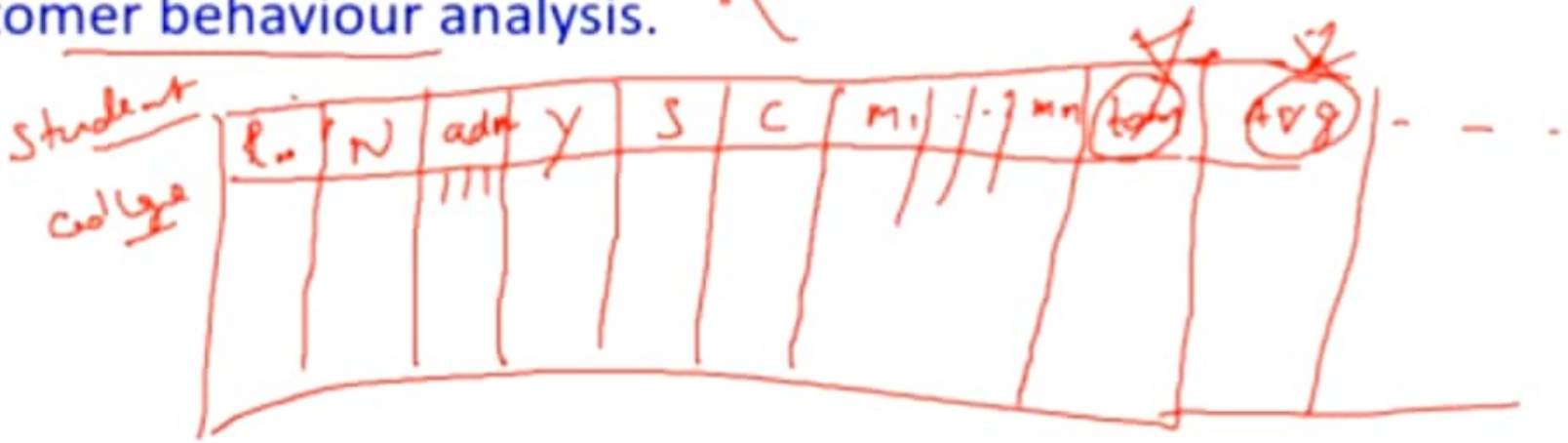
DATA PRE-PROCESSING

- Dimensionality reduction
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
- Feature subset selection



Dimensionality Reduction

- In machine learning, the number of attributes or features in data sets are in quite high.
- These projects have produced extremely high-dimensional data sets with more features being very common.
- Also, there has been a wide-spread adoption of social networking leading to a need for huge data, for example, text classification for customer behaviour analysis.



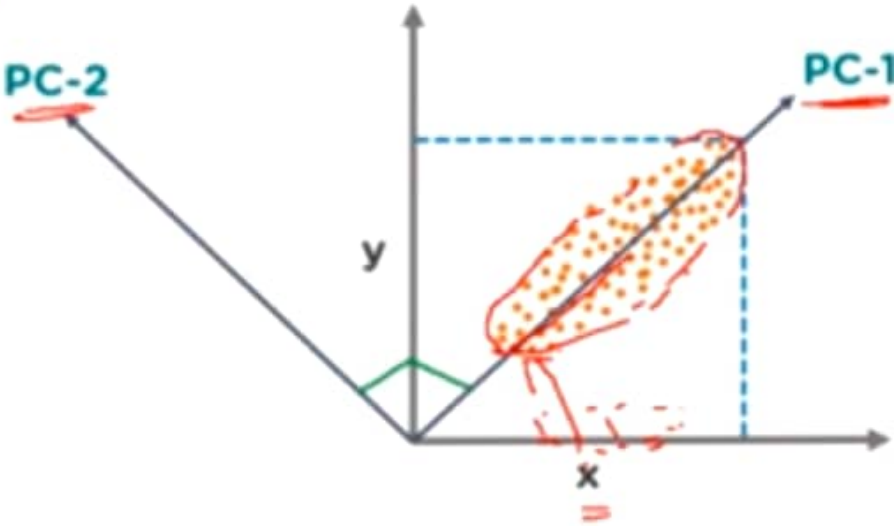
Dimensionality Reduction...

Dimensionality reduction...

- High-dimensional data sets need a high amount of computational space and time.
- In a data set, not all features are useful – they degrade the performance of machine learning algorithms.
- Most of the machine learning algorithms perform better if the dimensionality of data set,
- i.e. the number of features in the data set, is reduced.
- Dimensionality reduction helps in reducing irrelevance and redundancy in features.
- Also, it is easier to understand a model if the number of features involved in the learning activity is less.
- **Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes.**

Dimensionality reduction - Principal Component Analysis (PCA)

- PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called **principal components**.
- The principal components are a linear combination of the original variables.
- They are orthogonal to each other.
- Since principal components are uncorrelated, they capture the maximum amount of variability in the data.
- **The challenge** is that the original attributes are lost due to the transformation.



Singular Value Decomposition (SVD)

- The Singular Value Decomposition (SVD) of a matrix is, a factorization of that matrix into three matrices.
- It has some interesting algebraic properties and conveys important geometrical and theoretical insights about linear transformations.
- The SVD of m x n matrix A is given by the formula:
- $A = UWV^T$
- U: *mxn* matrix of the orthonormal eigenvectors of AA^T
- V^T: transpose of a *nxn* matrix containing the orthonormal eigenvectors of A^{T}A.
- W: a *nxn* diagonal matrix of the singular values which are the square roots of the eigenvalues of A^TA

Singular decomposition analysis(SVD)

Feature subset selection

- Feature subset selection or simply called feature selection.
- Applicable for both for supervised as well as unsupervised learning,
- Find out the optimal subset of the entire feature set which significantly reduces computational cost without any major impact on the learning accuracy.
- It may seem that a feature subset may lead to loss of useful information as certain features are going to be excluded from the final set of features used for learning.
- However, for elimination only features which are not relevant or redundant are selected.