6/08/22

# UNIT 1
## INTRODUCTION TO MACHINE LEARNING.

Q1) What is Machine Learning?

→ • Machine learning is a form of technology whereby the machine itself has a complex range of knowledge that allows it to take certain data inputs & use complex statistical analysis strategies to create output values that fall within the specific range of knowledge, data or information.

• "Machine learning focuses on development of computer programs which can access data & use it to learn for themselves".

• It essentially takes data, looks for patterns & trends & other specified information to create predictions or recommendations.

• The goal is for computers to learn how to use the data & information to be able to learn automatically, rather than requiring humans to intervene or assist with the learning process.

• An ML process begins with feeding the machine lots of data, by using this data the machine is trained to detect hidden traits, insights & trends.

• These insights are then used to build a Machine Learning Model by using an algorithm in order to solve a problem.

Data → Training the machine → Building a model → Predicting outcome.

Q2) What is the importance of Machine learning?

→ • With the availability of a large amount of data, it is

finally possible to build predictive models which study and analyze complex data to find useful insights & trends.

- Machine learning & data mining are crucial tools to get insights from massive datasets held by companies and researchers today.
- Reasons why Machine Learning is important:

a) Increase in Data Generation.

- Due to excessive production of data, we need a method to structure, analyze & draw useful insights from data.
- Machine Learning uses data to solve problems & find solutions to the most complex tasks faced by organizations.

b) Improve decision-making

By making use of various algorithms, Machine Learning can be used to make better business decisions.

c) Uncovers patterns & trends in data:

- This is the most essential part of machine learning.
- By building predictive models & using statistical techniques, ML allows you to dig beneath the surface & explore the data at a minute scale.
- Machine learning algorithms perform tasks such as understanding data & patterns in less than a second.

d) Solve complex problems.

From detecting the genes linked to the deadly ALS disease to building self driving cars, ML can be used to solve the most complex problems.

Q3. Define the following:

1) Algorithm:

- An ML algorithm is a set of rules & statistical techniques used to learn patterns from data & draw insights from it.

- It is the logic behind a machine learning model.
- Eg. Regression algorithm.

2) Model:
- A model is the main component of Machine learning.
- A model is trained by using a Machine learning algorithm.
- An algorithm maps all the decisions that a model ~~has~~ is supposed to take based on the given input, in order to get the correct output.

3) Predictor variable.
- A feature of data that can be used to predict the output.

4) Response variable
- A feature / output variable which needs to be predicted by using the predictor variables.

5) Training Data.
- The ML model is built using the training data.
- The training data helps the model to identify key trends & patterns essential to predict the output.

6) Testing Data
- After the model is trained, it ~~is~~ must be tested to evaluate how accurately it can predict an outcome which is done by testing data.

Q4. What are the various steps in machine learning process?

A

→ The Machine Learning process involves building a Predictive model that can be used to find a solution for a Problem

Statement.

i) Step 1: Define the objective of the Problem Statement.

- At this step, we must understand what exactly needs to be predicted.
- It is also essential to make a few mental notes on what kind of data can be used to solve this problem, or type of approach you must follow to get to the solution.

ii) Step 2: Data Gathering.

- At this stage, questions such as: What kind of data is needed to solve the problem? Is the data available? How can I get the data?
- Once we know the type of data required, we must understand how we can derive the data.
- Data can be collected manually or by web scraping.

iii) Data Preparation.

- The data collected is raw data.
- There are a lot of inconsistencies in the data such as missing values, redundant variables, duplicate values, etc.
- Removing the inconsistencies is necessary as they might lead to wrongful computations & predictions.
- The inconsistencies are fixed in this stage

iv) Exploratory Data Analysis.

- Data Exploration involves understanding the patterns and trends in the data.
- At this stage, all useful insights are drawn & correlations between variables are understood.

v) Building a Machine Learning Model.

- All the insights & patterns derived are used to build the Machine learning model.
- This stage begins with splitting the data into two parts: training data & testing data.

- The training data is used to build & analyze the model.
- The logic of the model is based on Machine Learning Algorithm which is implemented.

vi) Model Evaluation & Optimization
- Testing data is used to check the efficiency of how accurately it can predict the outcome.
- Once accuracy is calculated, any further improvements in the model can be implemented at this stage.
- Methods like parameter ~~tuning~~ tuning & cross-validation can be used to improve the performances of the model.

vii) Predictions
- Once the model is evaluated & improved, it is finally used to make predictions.
- The final output can be a categorical variable or it can be a continuous quantity.

Q5. State various applications of Machine Learning.

→ 1) Image recognition
- It is used to identify objects, persons, places, digital images etc.
- It is used in face detection.

2) Speech recognition.
- Speech recognition is the process of converting voice instructions into text.
- It is also known as "speech to text".
- Google Assistant, Siri, cortana & alexa use speech recognition.

3) Traffic Prediction.
- Google Maps predict the traffic conditions such as whether traffic is cleared, slow-moving or heavily congested with the help of Real time location of vehicle & sensors the
- Avg time taken on past days at same time

4) Product recommendation.
- ML is widely used by various e-commerce & entertainment companies for product recommendation to the user.
- Google understands the user interest using various ML algos & suggests products as per customer interest.
- Similarly, Netflix recommends movies & shows based on the previously watched series or movies.

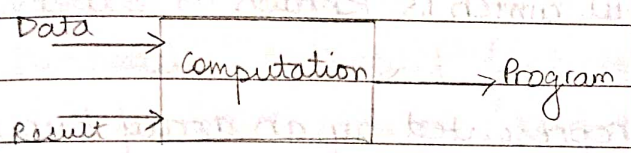
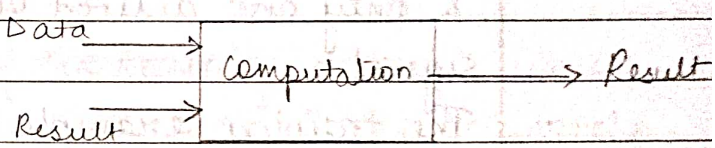
5) Email Spam & Malware detection.
- ML is used to filter mails.
- Whenever we recieve a ~~mail~~ new email, it is fitered automatically as important, normal & spam.
- Important mails are recieved in the inbox & spam emails in the spam box.

6) Online Fraud Detection.
- Machine Learning is making our online transactions safe & secure by detecting fraud transactions.
- Feed Forward Neural networks help us by checking whether it is a genuine transaction or fraud transaction.

7) Stock Market Trading
- ML is widely used in stock market trading.
- Machine Learning's short term memory neural network is used for the prediction of stock market trends.

8) Automatic Language Translation
- Google's GNMT (Google Neural Machine Translation) provide the feature, in which there consists a Neural Machine Learning that converts text into familiar language & it is called automatic translation.

Q6. Differentiate between Machine Learning and traditional programming.

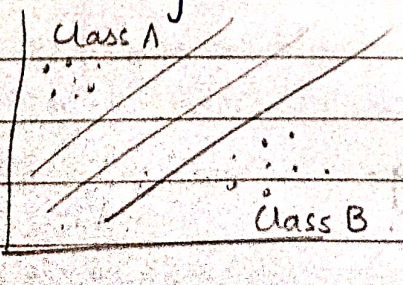
| Machine Learning | Traditional Programming |
|---|---|
| 1) Machine Learning is not a manual process. | Traditional programming is a manual process. |
| 2) The algorithm automatically formulates the rules from data. | A person/programmer creates the program. |
| 3) Machine learning approach:   | Traditional Programming:   |
| 4) In Machine learning we show the examples & machine figures out how to solve the problem by itself. | In Traditional Programming, we write down the exact steps reqd. to solve the problem. |
| 5) An ML algo. takes an input & output & gives out some logic which can be used to work with new input to give an output. | A traditional algorithm takes some input & logic in the form of code & gives output. |

Q7. What are the types of Learning?

→1) Supervised Learning
2) Unsupervised Learning
3) Reinforcement Learning.

88) What is Supervised Learning? How does it work?

→ • Learning that takes place based on a class of examples is known as supervised learning.
• It is based on labelled data.
• It comprises of a series of algorithms that build mathematical models of certain data sets that are capable of containing both inputs & the desired outputs for that particular machine.
• The data being inputted is called training data which consists of training examples which contain one or more inputs & only one desired output which is known as supervisory signal.
• The training example is represented by an array known as vector or feature vector.
• The training data is represented by a matrix.
• Ideally, if the supervised learning algorithm is working properly, the machine will be able to correctly determine the output for the inputs that were not a part of training data.
• Supervised learning uses classification uses classification & regression techniques to develop predictive models.
• Predictive classification techniques predict categorical responses. Regression techniques predict continuous responses.
• Let us take an example of classification of documents.
• In this case, learner learns based on the available docs & their classes.
• Program which maps the input documents to appropriate classes is called a classifier, as it assigns class (document type) to an object (document).
•



Q9) Q8. How does supervised learning work?

→ . In supervised learning, models are trained using labelled dataset, where model learns about each type of data.
. Once the model is tested on the basis of test data & it predicts the output.
. Following are the steps involved in supervised learning:
1) Determine type of training set.
2) Collect the labelled training data
3) Split the training dataset into data set, test dataset & validation dataset.
4) Determine the input features of the training dataset, which should have enough knowledge so that model can accurately predict the output.
5) Determine the suitable algorithm for model, such as support vector machine, decision tree etc.
6) Execute the algorithm on dataset.
7) Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, it means our model is accurate.
. Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle & Polygon.
. Now the first step is to train the model for each shape:
- If the given shape has 4 sides, & all sides are equal, it is a labelled as a square
- If the given shape has 3 sides, it is labelled as triangle.
- If the given shape has 6 equal sides, it is labelled as hexagon.
. Now, after training, we test the model using test set & the task of the model is to identify the shape.
. The machine is already trained on all types of shapes, & when it finds a new shape, it classifies the shape on the basis

of a no. of sides & predicts the output.

Q10. What are the advantages & disadvantages of Supervised Learning

→ Advantages

i) The model can predict the output on the basis of prior experiences.
ii) We can have an exact idea about the classes of objects.
iii) It helps to solve various real-world problems such as fraud detection, spam filtering etc.

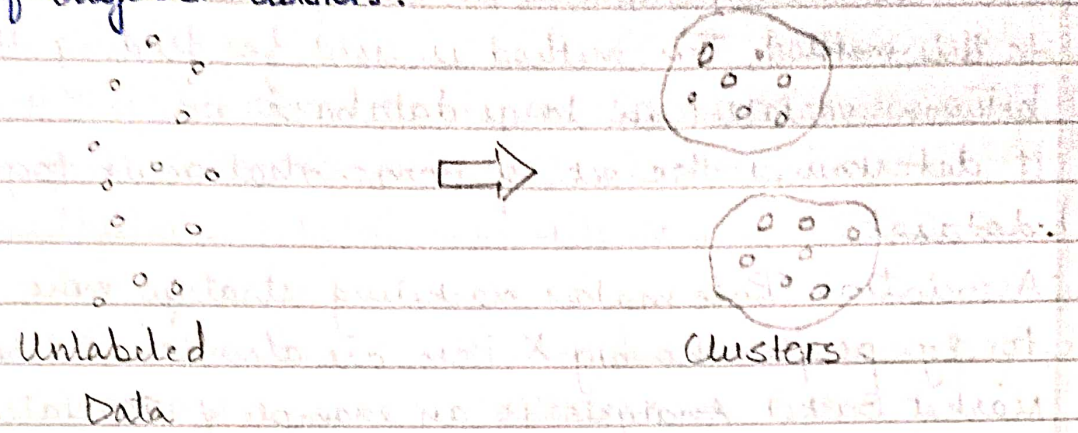
Disadvantages

i) They are not suitable for handling the complex task.
ii) They cannot predict the correct output if ~~data~~ test data is different from the training dataset.
iii) Training requires lots of computation times.
iv) In supervised learning, we need enough knowledge about the classes of object.

Q11. What is unsupervised learning?

→ • Unsupervised learning refers to learning from unlabelled data.
• Learning is based more on similarities & differences that are visible.
• In this type of learning, all similar items are clustered together in a particular class where label of the class is not known.
• The differences & similarities are mathematically represented in * unsupervised learning.
• The criterion in initial stages is the most visible aspects of those objects.
• Many unsupervised learning algorithms create heirarchial arrangements based on similarity-based mappings.

- The task of heirarchial clustering is to arrange a set of objects into a heirarchy such that similar objects are grouped together.
- Non-heirarchial clustering seeks to partition the data into some number of disjoint clusters.




- A learner is fed with a set of scattered points, & it generates two clusters with representative centroids after learning.
- Clusters show that points with similar properties & closeness grouped together.
- Unsupervised learning is a set of algorithms where the only information is being uploaded is inputs.
- The device is responsible for grouping together & creating ideal OPs based on the data it discovers.
- The information in the algorithms being run by unsupervised learning methods is not labelled or classified by humans.
- Instead, the unsupervised algorithm rejects responding to feedback in favor of identifying commonalities in the data.

Q12. What are the types of unsupervised learning?

→ The unsupervised learning algorithm can be categorized into two parts:

i) Clustering

- Clustering is a method ~~of~~ of grouping the objects into clusters such that objects with most similarities remains into a group & has less or no similarities with objects of another group.

- Cluster analysis finds commonalities between data objects & categorizes them as per the presence & absence of commonalities

ii) Association

- An association rule is an unsupervised learning method.
- In this method, This method is used for finding the relationships between variables in large database.
- It determines the set of items that occur together in the dataset.
- Association Rule makes marketing strategy more effective.
- For Eg. people who buy X item are also tend to buy Y item.
- Market Basket Analysis is an example of association rule.

Q13. What are the advantages & disadvantages of Unsupervised Learning?

→ * Advantages..

a) Unsupervised learning is used more for complex tasks in supervised learning as in unsupervised learning we don't have labeled input data

b) Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

* Disadvantages..

a) Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output

b) The result of unsupervised learning may be less accurate as input data is not labeled, & the algo does not know the exact output in advance.

Q14. What is the difference between Supervised Learning and Unsupervised Learning?

| | Supervised learning | Unsupervised learning. |
|---|---|---|
| 1) | Supervised learning algos are trained using labelled data. | Unsupervised learning algos are trained using unlabeled data. |
| 2) | It takes direct feedback from to check if it is predicting correct output. | It does not take any feedback. |
| 3) | Supervised learning model predicts the output. | Unsupervised learning finds hidden patterns in data |
| 4) | Both input & output data is provided to the model. | Only input data is provided to the model |
| 5) | The goal is to train the model to predict output when it is given new data. | The goal is to find hidden patterns & insights from unknown dataset |
| 6) | It is categorized into Classification & Regression prob. | It is categorized into Clustering & Association problems. |
| 7) | Can be used in cases where we know the input as well as output. | Can be used in cases where we know the input only. |
| 8) | It produces accurate result. | It may give less accurate result. |
| 9) | Includes algorithms such as Linear Regression, ~~& KNN~~, Decision Tree etc. | Includes algos such as Clustering, KNN & Apriori algorithm. |

Reinforcement

Q15. What is Bayesian learning ? Explain with an example.

→ • Reinforcement learning is a type of machine learning method where an intelligent agent interacts with the environment & learns to act within that.
• Agent learns to behave in an environment by performing the actions & seeing the results of the actions.
• If the agent gets a positive feedback, it is a good action & if the agent gets negative feedback or penalty it is a bad action.
• The agent learns automatically using feedbacks without any labelled data unlike supervised learning.
• RL solves a specific type of problem where decision making is sequential, & goal is long-term.
• The primary goal of the agent is to gain maximum positive points & improving performance.
• Eg - Suppose there is an AI agent present in a maze environment & its goal is to find the diamond.
The agent interacts with its environment by performing some actions & based on those actions, the state of agent gets changed & it also recieves a reward or penalty as feedback.
The agent continues doing these things - take action, change state / remain in same state & take action.
The agent learns what leads to positive & negative feedback.
As positive feedback → positive point
negative feedback → penalty / negative point.

Q16. What are the key features of Reinforcement learning?

→ a) In RL, the agent is not instructed about the environment & what actions need to be taken.

b) It is based on hit & trial process.

c) The agent takes the next action states according to the feedback of the previous action.

d) The agent may get a delayed reward.

e) The environment is stochastic, & the agent needs to explore it to reach to get the maximum positive rewards.

Q17) What are the approaches for Reinforcement Learning?

→ There are 3 ways to implement RL in ML:

a) Value-based

- The value-based approach is about to find optimal value function, which is the maximum value at a state under any policy.
- Therefore, the agent expects the long-term return at any state under policy π.

b) Policy-based

- Policy-based approach is to find the optimal policy for the maximum future rewards without using the value function.
- In this approach, the agent tries to apply such a policy that the action performed in each step helps to maximize the future reward
- It has mainly two types of policy:

1) Deterministic: same action is produced by the policy(π) at any state

2) Stochastic: Probability determines the produced action.

c) Model-based

- In model-based approach, a virtual model is created for the environment & agent explores that environment to learn it.
- There is no particular solution for algorithm for this approach as the model's representation is different for each environment.
- Important characteristics of reinforcement learning:

- There is no supervisor, only a real number or reward signal.
- Sequential decision making.
- Time plays a crucial role in Reinforcement problems.
- Feedback is always delayed
- Agent's actions determine the subsequent data it recieves.

Q18) What is the difference between Reinforcement learning and Supervised Learning?

| | Reinforcement Learning | Supervised Learning |
|---|---|---|
| 1) | RL helps you take your decisions sequentially. | A decision is made on the input given in the beginning. |
| 2) | Works on interacting with the environment | Works on the given sample data. |
| 3) | In RL, learning decision is dependent. ∴ You should give labels to all dependent decisions. | In supervised learning the decisio which are independent of each other, so labels are given for each decision. |
| 4) | Eg- Chess | Eg. Object recognition. |

Q19) What is meant by Probabilistic Model?

→ • In contrast to deterministic models where the relationship between quantities is already known, probabilistic models are based on the assumption of relationship bet$^n$ quantities which is reasonably accurate but other components are also

taken into consideration.

- Probabilistic models are statistical models which give probability distribution to account for these components.
- Probabilistic models from machine learning, artificial intelligence etc. ~~are~~ rest on two basic rules of probability theory: the sum rule & the product rule.
- Eg - If one lives in a cold climate, one knows that traffic tends to be more difficult when snow falls & covers the roads.
- We can go further & make a hypothesis: There will be strong correlation between snowy weather & increased traffic incidents.
- Probabilistic models are used in a variety of disciplines like statistical physics, quantum mechanics etc.

Q20. What are logical models? What are its types?

→. Logical models use a logical expression to divide the instance space into segments & hence construct grouping models.

- A logical expression is an expression that returns a boolean value, i.e. True or False outcome
- Once data is grouped using a logical expression, the data is divided into homogeneous groupings for the problem we're trying to solve.
- There are two types of logical models:

1) Rule model

It consist of a collection of implications or IF-then rules. For tree-based models, 'if-parts' defined a statement & 'then' part defines behaviour of the model for this segment.
Rule model follows the same reasoning.

2) Tree Model

- They can be seen as a particular type of rule model where the if parts of the rules are organized in a tree structure.

Q21. Differentiate between Parametric & Non-Parametric Models.

| | Parametric Models | Non-Parametric Models. |
|---|---|---|
| 1) | Parametric models use fixed no. of parameters to build the model. | Non-parametric models use flexible no. of parameters to build the model. |
| 2) | Parametric analysis is for testing group means | A non-parametric analysis is for testing medians. |
| 3) | It is applicable only for variables. | It is applicable for variable & attribute |
| 4) | Always considers strong assumptions about data | Generally considers fewer assumptions & data. |
| 5) | Require lesser data | Require much more data |
| 6) | Parametric ~~models requ~~ data handles intervals data / ration data | Non-parametric model handles ordinal data. |
| 7) | Follow normal distribution | No assumed distribution in non parametric methods. |
| 8) | Output generated can be easily affected by outliers. | Output generated cannot be seriously affected by outliers. |
| 9) | Have more statistical powers. | have less statistical power. |
| 10) | Models are computationally faster. | Computationally slower |
| 11) | Eg. Logistic Regression | Eg. KNN |

Q22) What are Geometric Models?

→ • In Geometric Models, features could be described as points in 2-D (x, y) or in three-d (x, y, z).
• Even when features are not intrinsically geometric, they could be modelled in a geometric manner &
• For eg. temperature as a function of time can be modelled in two axes.
• In geometric models, there are two ways we could impose similarity:
• We could use geometric concepts like line or planes to segment the instance space, these are called linear models.
• Alternatively, we can use geometric notion of distance to represent similarity.
• In this case, if two points are close together, they have similar values for features & thus can be classed as similar., they are known as Distance-based models.

Q23) What are groping models?

→ • Tree models are repeatedly split the instance space into smaller subsets
• Trees are usually of limited depth & don't contain all information.
• Subsets are the leaves of Trees, partition the instance space with finite resolution.
• Instances filtered into the same leaf of the tree are treated the same, regardless of any features the the tree that might be able to distinguish them.

Q24) What are Grading models?

→ • Grading models don't use the the notion of segment.
• Forms one global model over instance space.

- Grading models are usually able to distinguish between arbitrary instances, no matter how similar they are.
- Resolution is in theory infinite, particularly when working in a cartesian space.
- Examples: Support vector machines.
- They work in cartesian instance space
- They exploit the minutest difference between ~~st~~ instances.

# UNIT : 2

## Feature Engineering.

Q1) Explain the concept of feature.

→ . Features are individual independent variables that act like input in your system.
. Feature is an attribute of a data set & used in machine learning process.
. The features in dataset also known as dimensions.
. Data having 'n' features is called n-dimensional dataset.
. A good feature representation is central to achieving high performance in any machine learning task.
. Two features are redundant if they are highly correlated regardless of whether they are correlated with the task or not.
. Feature engineering is the process of creating features that do not exist in the dataset.

Q2) Define feature engineering. Explain four processes in feature engineering.

→ . Feature engineering is the pre-processing step of ML machine learning, which extracts features from raw data.
. It is the process of creating features that don't exist in the dataset.
. Feature engineering refers to the process of translating a data into features such that these features are able to represent the data more effectively & result in better learning performance.
. The four processes in feature engineering are as follows:

1) Feature creation

Feature creation is finding the most useful variables to be

used in a predictive model.

- The process is subjective & requires human creativity & intuition.
- New features are created by mixing existing features using addition, subtraction, & ration, & these new features have great flexibility.

2) Transformations

- It involves adjusting the predictor variable to improve accuracy & performance of the model.
- It ensures that the model is flexible to take input of variety of data.
- It ensures that all variables are on the same scale, making the model easier to understand.
- It improves accuracy of the model.

3) Feature Extraction

- Feature extraction generates new variables by extracting them from the raw data.
- It reduces the volume of data so that it can be easily used and managed for data modelling.
- Feature extraction methods include cluster analysis, text analysis, edge detection algorithms & PCA.

4) Feature selection

- Feature selection is the way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant or noisy features.

Q3) Define data preprocessing. Explain steps involved in data preprocessing

→ • Data pre-processing is the process of preparing raw data to be used on a machine learning model.
- It is the first & very important step in developing an ML model.

- Real-world data typically contains noise, missing values, and may be in unusable format that cannot be used directly for machine learning models.
- Data preprocessing is a necessary task for cleaning & preparing the data for machine learning model.
- Data preprocessing involves the following steps:

i) Get the dataset.
- The first thing we need is to create a machine learning model is a dataset.
- We usually save the dataset in CSV format while using it in our code.

ii) Importing libraries.
- To perform data preprocessing with Python, we must first import some predefined Python libraries.
- These libraries are used to carry out specific tasks.
- For data preprocessing we use Numpy, Pandas & Matplotlib.

iii) Importing the datasets.
- We must now import the datasets we have gathered for the project.
- We use pandas library's read_csv() function which reads a csv file & performs various operations on it.
- It is essential to distinguish feature matrix from the dataset.

iv) Handling the missing data
- This step deals with missing data in the dataset.
- There are two ways to deal with missing data:
  a) Removing the specific row, which is slightly inefficient as it involves in loss of data.
  b) By getting the mean of column/row which contains the missing value.

v) Encoding categorical data.
- Categorical data is data which has some categories.

- As machine learning models are based entirely on maths & nos, having a categorical variable in the dataset may cause problems when building the model.
- As a result, these categorical variables must be encoded into numbers
- We can use OneHotCoding or Label Encoding Technique.

vi) Splitting the Dataset into Training Set & Test Set.
- In ML data preprocessing, the data is split into training set & test ~~step~~ set.
- This is an important step as it allows us to improve the performance of our machine learning model.
- Assume we trained our model on one dataset & then tested it on another.
- It will then be difficult for our model to understand the correlation between the models.

viii) Training set : subset of dataset to train machine model.
Test set : subset of dataset to test the model.

vii) Feature scaling:
- Final step
- It is method for standardizing the independent variables of a dataset within a given range.
- In feature scaling we place our variables in the same range & scale so that no variable dominates the other.

Q4) Write a short note on normalization & scaling.

→ i). Normalization & scaling are so similar that they are used interchangeably, but they have different effects on the data.
- In both normalization & scaling we are transforming the value of numeric values so that the transformed data points have specific helpful properties.
- In scaling, we change the range of distribution of data.

- In normalization, we change the shape of distribution of data.
- In scaling we are transforming the data so that it fits within specific scale like 0-100 or 0-1.
- Normalization is more radical transformation.
- The point of normalizations is to ~~check~~ change your observations so that they can be described as normal dist.
- Types of Scaling:

1) Simple feature scaling.
- This method divides each value by max. value for that feature
- Resultant values range betⁿ 0-1.

$$X_{new} = \frac{X_{old}}{X_{max}}$$

2) Min-Max Scaling.
- This scaler takes each value, subtracts the minimum & divides it by range.
- Resultant values range betⁿ 0-1.

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}}$$

- Types of Normalization.

1) Z-Score / Standard score.
- In this technique, values are normalized based on mean and standard deviation of data A.
- Formula used:

$$X_{new} = \frac{X_{old} - \mu_A}{\sigma_A} \qquad \sigma_A = S.D \qquad \mu_A = mean$$

2) Box-cox Transformation
- It is a transformation of non-normal dependent variable into a normal shape
- At the heart of the box-cox is an exponent $\lambda$ whose value varies from -5 to 5.

All values of $\lambda$ are considered & optimal value for your data is selected.

$$W_t = \begin{cases} \log(y_t) & ; \text{if } \lambda = 0 \\ \frac{y_{t-1}}{\lambda} & ; \text{otherwise.} \end{cases}$$

(Q5) Explain Standardization.

→
- Standardization entails ~~fitting~~ scaling of data to fit a standard normal distribution.
- A standard normal distribution is defined a distribution with a mean of 0 & S.D. of 1.
- When your data has variable dimensions, standardization is useful.
- It is used for feature scaling when your data follows Gaussian distribution.
- It is most useful for:
  1) Optimizing algorithms such as gradient descent.
  2) Clustering models.
- Data standardization is the process of placing dissimilar features on the same scale.
- It can be defined as rescaling the attributes in such a way that their mean is 0 and s.d. becomes 1.
- Let $x$ be an individual feature value & $\min(x)$, ~~min~~ $\max(x)$ be minimum & maximum values of this feature over ~~the~~ the entire dataset.
- Min max squeezes all feature values to be in between $[0,1]$.
- Feature standardization,

$$\bar{x} = \frac{X - \text{Mean}(x)}{\sqrt{\text{Var}(x)}}.$$

- Z score is the most popular methods to standardize data.

$$z = \frac{\text{value} - \text{mean}}{\text{Standard dev.}}$$

Q6. Explain how missing values are handled in data preprocessing.

→ i) Ignore the tuple.
- When class label is missing, this technique is used.
- However, unless the tuple contains numerous attributes with missing values, this approach is not particularly useful.

ii) Fill in the missing value manually.
This approach is effective on small datasets with some missing values.

iii) Use a global constant to fill in the missing value.
You can replace all missing attribute value with a global constant like "Unknown" or $-\infty$.

iv) Use measure of central tendency for attribute to fill in the missing value.

v) Use the attribute mean or median for all samples belonging to the same class as given tuple.
- For eg. if you are classifying customers according to their credit-score, then you can replace the missing values with mean income value for customers in the credit-score category as that of the given tuple.
- If the data distribution is skewed, use median value.

vi) Use the most probable value to fill in the missing value.
This can be determined using regression, Bayesian classifier or decision tree.

Q7. Explain how PCA helps in dimensionality reduction.

→.
- PCA - stands for Principal Component Analysis
- It is an unsupervised learning algorithm that is used for the dimensionality reduction in ML.
- It converts the observations of correlated features into a set of

linearly uncorrelated features with the help of orthogonal transformation.

- It is a technique to draw strong patterns from the given dataset by reducing the variances.
- Some real-world applications of PCA are image-processing, movie recommendation system, optimizing the power allocation in various communication channels.

* Steps.

1) Standardize the dataset.
   - We standardize the dataset, by calculating the mean and standard deviation for each feature.
   - z-score is used.
2) Calculate the covariance matrix for the given dataset will be calculated as

   For population

$$Cov(x,y) = \frac{\Sigma(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

   For sample

$$Cov(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

   For 3 variables x, y, z

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(yz) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

3) Calculate the Eigen values & Eigen Vectors.
4) Sort the eigen vectors ~~for~~ from the highest eigen value to the lowest.
5) Select the the no. of principal components.
6) Transform the original matrix

   feature matrix * top k eigen vectors = Transformed Data.

Q8. What is feature selection? Explain different selection algorithms.

→ • Feature selection is the method of reducing the input variable to your model by using only relevant data & getting rid of noise in data.

• It is the process of automatically choosing relevant feature for your ML model based on the type of problem you are trying to solve.

• We do this by including important features without changing them

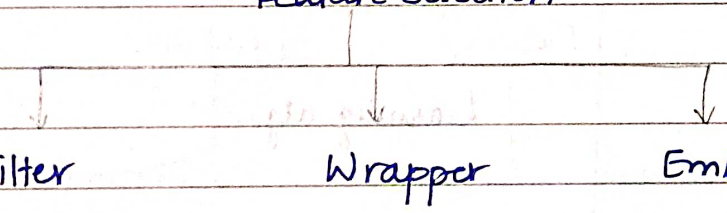
• There are 3 types of selection algorithms:-

a) Filter methods

• The role of feature selection in ML is:

1. to reduce the dimensionality of feature space.
2. to speed up a learning algorithm.
3. to improve the predictive accuracy of classification algorithm
4. to improve the comprehensibility of the learning results.

• There are 3 types of approach for feature selection:

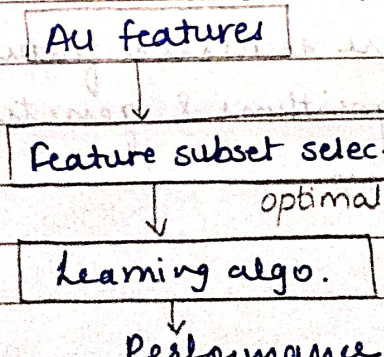



1) Filter method.

• Filter methods are generally used as preprocessing step.

• The selection of features is independent to of any ML algorithms.

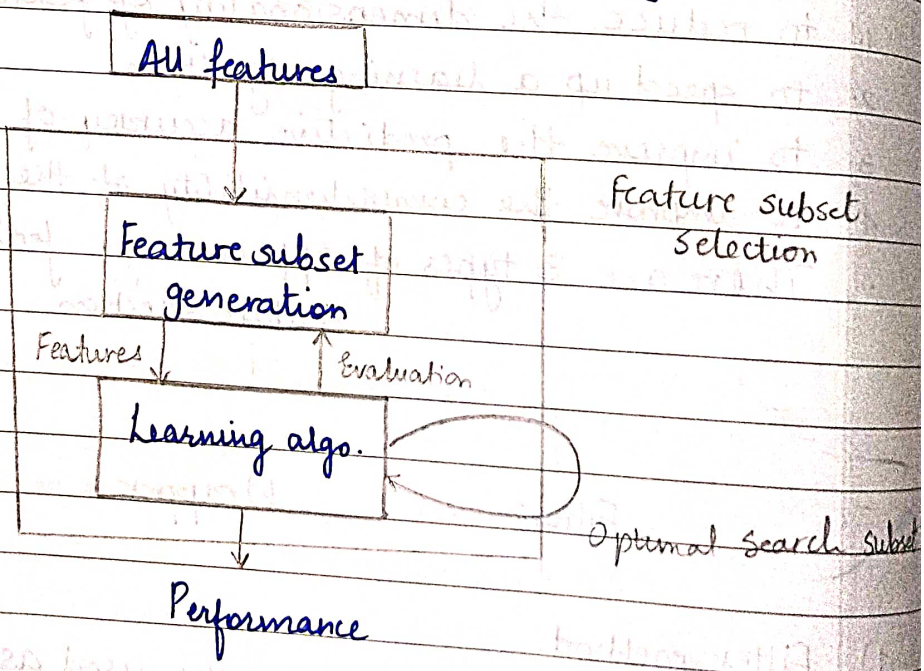
• Features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.



- The filter feature method make use of statistical techniqu predict the relationship between each independent input var & output variable which assigns scores for each feature.
- Correlation based feature selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlat based heuristic evaluation function.
- Eg- co-orelation, chi-square test, ANOVA, info-gain etc.

2) Wrapper Method

- In wrapper method, the learner is considered a blackbox.
- The interface of the blackbox is used to score subsets of variable according to predictive power of learner when using subsets.




- The feature selection algorithm searches for a good feature subset using the induction algorithm itself as a part of the evaluat function.
- Wrapper methods are recursive feature elimination, sequenti feature selection algorithms & genetic algorithms.

3) Embedded methods.

- Embedded methods are similar to wrapper method as they are also used to optimize the objective function or performance of a learning algorithm or model.
- It's implemented by algorithms that have their own feature selection methods in them.
- A learning algorithm takes advantage of its own variable selection process & performs feature selection & classification at the same time.
- Random forest, LASSO are most commonly used embedded techniques.

x

## Statistical Feature Engineering.

1) **Mean**

The mean of a dataset is the average of all values.

2) **Median**

The median of a dataset is the value in the middle when dataset items are arranged in ascending order.

Odd → middle term

Even → avg. of middle two values.

3) **Mode**

The mode of dataset is the value that occurs with greatest frequency.