Machine Learning -

Incorporating human like ability in the software of the machine is known as machine learning.

Type of Machine Learning Problem:-

- Supervised

- Unsupervised

- Reinforced

Aspect of learning -

**Remembering and adapting**: If a similar situation in the past resulted in a positive outcome, try doing the same thing again. If the action failed, try something different this time.

**Generalizing**: When facing a new situation that hasn't been encountered before, the machine learning system will try to find similarities with situations it has already encountered in the past and use that information to make a decision.

Types of Learning -

- Regression

- Classification

- Clustering

Give 2 points on How ML is different from Data Mining

Purpose: ML is focused on developing algorithms to learn patterns from data and make predictions, while data mining involves discovering patterns and relationships in data for decision-making.

Process: ML involves statistical and computational techniques to train and test models on data, while data mining involves exploratory analysis, preprocessing, and pattern discovery to extract insights from data.

Give 3 points on How ML is different from Deep Learning

Complexity of Models: Deep learning models are typically more complex than traditional machine learning models. Deep learning algorithms use neural networks with multiple layers to learn patterns in data, while traditional machine learning algorithms use simpler models like decision trees or logistic regression.

Data Requirements: Deep learning algorithms require a large amount of training data to perform well, while traditional machine learning algorithms can often work with smaller datasets. Deep learning models may also require more preprocessing of the data before training.

Interpretability: Traditional machine learning models are often more interpretable than deep learning models. This means that it is easier to understand how the model is making its predictions. Deep learning models can be more difficult to interpret because of their complexity and the large number of parameters involved

Supervised Learning -

Supervised learning is a type of machine learning in which an **algorithm learns** to **map input data** to a **set of output labels** or values **based on labeled training data**.

**Least Square Method : -**

1. The least-squares method is a **statistical method** that is practised to find a regression line or a best-fit ***line* for the given pattern.**
2. The method of least squares is **used in regression**.
3. In regression analysis, least square method is said to be a standard approach for estimating the value of unknown variables when, they have more equations than unknowns (overdetermined systems).
4. **In the method of least squares, we want to find the curve that best fits a set of data points. We do this by calculating the difference between the actual data points and the predicted values of the curve, squaring these differences, and adding them up. The curve that has the smallest total squared difference from the actual data points is considered to be the best fit curve. This ensures that the curve is as close to the actual data points as possible.**

**Curve fitting and Least Square Method -**

1. Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, **possibly subject to constraints.**
2. Curve fitting can **involve either interpolation**, where an exact fit to the data is required, or **smoothing**, in which a "smooth" function is constructed that approximately fits the data.
3. The least squares method is one way to compare the deviations.

**CHAPTER - 3 - Regression**

Regression – **predict value** of **response variable** from **attribute variables**.

Regression shows the change of variable in the y axis with respect to the change in explanatory variable in the x axis.(explanatory variable is the independent variable(s) and the variable along the y axis is the dependent variable.)

Regression analysis is for finding out which of the independent variable(s) have an impact on the outcome variable.

Basics of Regression Models -

**Interpolation** is the process of estimating a value of the dependent variable within the range of the independent variable based on the available data points. In other words, if we have a set of data points for a given range of the independent variable, we can use interpolation to estimate the value of the dependent variable for any value of the independent variable within that range.

**Extrapolation**, on the other hand, is the process of estimating a value of the dependent variable outside the range of the independent variable based on the available data points. In other words, if we have a set of data points for a certain range of the independent variable, we can use extrapolation to estimate the value of the dependent variable for any value of the independent variable outside that range.

Over-fitting occurs **when a model is too complex**, meaning it has too many parameters **relative to the amount of available data**. In this case, the model becomes too **closely fit** to the **training data** and starts to capture the noise or rando m fluctuations in the data, rather than the underlying patterns. **As a result, the model may perform very well on the training data, but it generalizes poorly to new, unseen data.**

Under-fitting occurs **when a model is too simple** and is **unable to capture the underlying patterns in the data**. In this case, the model may have high bias, meaning it consistently misses the correct relationship between the input features and output variable. The model may perform poorly both on the training data and new, unseen data.

Differentiate between weak-AI and Strong-AI.

**Weak AI** systems rely on supervised or unsupervised learning algorithms to solve specific problems, while **strong AI** systems are designed to learn from a wide range of inputs and **adapt to new situations** without requiring additional programming. In other words, **weak AI systems are specialized and focused**, while **strong AI systems are flexible and capable of generalization.**

Difference between ML and DL

Machine Learning:

- **ML is a subset of AI** that enables machines to learn from data, without being explicitly programmed.

- It uses a variety of algorithms to train models on historical data, which can then be used to make predictions or decisions about new data.

- ML algorithms can be supervised, unsupervised, or semi-supervised, depending on the type of data available and the problem being solved.

- ML models are typically trained on structured or semi-structured data, such as numerical or categorical data in a database.

- ML requires feature engineering, or the process of selecting, extracting, and transforming the most relevant features from the data, to make accurate predictions.

Deep Learning:

- DL is a subset of ML that uses neural networks to learn from large amounts of unstructured or raw data, such as images, audio, and text.

- It uses multiple layers of interconnected nodes, or neurons, to automatically learn complex patterns and features from the data.

- DL algorithms can be supervised, unsupervised, or reinforcement learning, and are often used in computer vision, natural language processing, and speech recognition.

- DL models can handle unstructured data, such as images, audio, and text, and do not require feature engineering.

- DL requires large amounts of data and high-performance computing resources to train models effectively.

**Explain different types of learning using suitable real-world examples-**

Supervised Learning:

Supervised learning is the most common type of learning, in which a machine learning algorithm is trained on a labeled dataset, where the desired output is already known. The algorithm learns to map inputs to outputs, and can make predictions on new, unlabeled data. Examples include:

Image classification: A supervised learning algorithm can be trained to classify images of cats and dogs based on labeled examples of each.

Fraud detection: A supervised learning algorithm can be trained on labeled data of fraudulent and non-fraudulent transactions to detect future fraudulent activity.

Sentiment analysis: A supervised learning algorithm can be trained on labeled text data to predict the sentiment of new text, such as positive or negative reviews.

Unsupervised Learning:

Unsupervised learning involves training a machine learning algorithm on an unlabeled dataset, where the desired output is unknown. The algorithm learns to find patterns and structure in the data, and can be used for clustering, dimensionality reduction, and anomaly detection. Examples include:

Customer segmentation: An unsupervised learning algorithm can be used to group customers based on their purchasing behavior, without any prior knowledge of customer segments.

Anomaly detection: An unsupervised learning algorithm can be used to detect unusual patterns or outliers in a dataset, such as fraudulent transactions or defective products.

Recommender systems: An unsupervised learning algorithm can be used to recommend products or services to users based on their preferences and behavior, without any labeled data.

**Reinforcement Learning: -**

1. Reinforcement learning involves training a machine learning algorithm to make decisions based on feedback from its environment.
2. The algorithm learns to maximize a reward function by taking actions that lead to positive outcomes and avoiding actions that lead to negative outcomes. Examples include:

Game playing: A reinforcement learning algorithm can be trained to play a game, such as chess or Go, by receiving rewards for winning and penalties for losing.

Robotics: A reinforcement learning algorithm can be used to train a robot to perform a task, such as walking or grasping objects, by receiving rewards for successful actions and penalties for unsuccessful actions.

Resource management: A reinforcement learning algorithm can be used to optimize the use of resources, such as energy or bandwidth, by receiving rewards for efficient use and penalties for waste.

Explain different data types used in modern machine learning paradigm

with examples.

What is true zero point for a numeric data type? Explain with example.

In machine learning, the concept of a "true zero point" refers to a numeric data type where the value of zero represents an absence of the quantity **being measured, rather than an arbitrary or relative value**. In other words, **a true zero point is a value that has a meaning beyond just being a numerical value.**

An example of a true zero point in machine learning is the weight of an object. If an object weighs zero, it means that the object has no weight at all. This is a true zero point because the value of zero represents an absence of weight. Another example is temperature measured in Kelvin, where zero Kelvin represents the absolute absence of heat.

Differentiate between Univariate and Multivariate data analysis.

Definition: Univariate analysis involves examining a single variable at a time, while multivariate analysis involves analyzing multiple variables simultaneously.

Goal: The goal of univariate analysis is to describe and summarize a single variable's distribution, while multivariate analysis aims to identify relationships or patterns among multiple variables.

Variables: Univariate analysis deals with only one variable, whereas multivariate analysis can deal with two or more variables.

Techniques: Univariate analysis techniques include measures of central tendency, such as mean, median, and mode, as well as measures of variability, such as range, variance, and standard deviation. Multivariate analysis techniques include regression analysis, principal component analysis, and factor analysis.

Explain at least 2 remedies for skewness in ML - Logarithmic transformation and Box-Cox transformation

Discuss at least 2 solutions for kurtosis in ML - Winsorizing and Data Transformation

**(UNREAD)**

Explain the similarity and dissimilarity between Normal distribution and

Students T-test.

Similarities:

Both are related to the concept of the mean and standard deviation.

They are both used for hypothesis testing.

Dissimilarities:

Normal distribution is a probability distribution that is symmetrical and bell-shaped, while Student's t-test is a statistical hypothesis test used to determine if two samples are significantly different from each other.

Normal distribution is characterized by two parameters: mean and standard deviation, while t-test is characterized by degrees of freedom, sample mean, and standard deviation.

Normal distribution can be used to determine probabilities of random variables, while t-test is used to determine if two groups are statistically different from each other.

Normal distribution can be used for both one-sample and two-sample tests, while t-test is primarily used for two-sample tests.

**(UNREAD)**

Explain discriminative and generative learning models with suitable

examples.

Discriminative Learning:

A discriminative learning model learns the boundary or decision boundary between different classes. It tries to learn the mapping between input features and labels directly without worrying about the underlying distribution of the data. It is mainly concerned with finding the optimal decision boundary that separates different classes. Examples of discriminative models include Support Vector Machines (SVM), Logistic Regression, and Neural Networks.

For example, in a spam email detection problem, a discriminative model learns to differentiate between spam and non-spam emails based on the features of the email such as keywords, sender, subject, etc. The model learns to predict the label (spam or not) directly based on the features without worrying about the underlying distribution of the data.

Generative Learning:

A generative learning model learns the underlying distribution of each class and then uses that distribution to classify new samples. It tries to model the joint probability distribution of the input features and labels. It is mainly concerned with finding the probability distribution of the input features given the label and then uses Bayes' rule to calculate the probability of the label given the features. Examples of generative models include Naive Bayes, Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM).

For example, in a handwriting recognition problem, a generative model learns the probability distribution of each class (digits from 0-9) and then uses that distribution to classify new images of handwritten digits. The model learns the distribution of the pixel values for each digit and then calculates the probability of each digit given the pixel values of the image.

Error in ML

Training error: The training error is the error that occurs when the model is trained on the training data. It is the difference between the predicted output and the actual output of the training data.

Validation error: The validation error is the error that occurs when the model is tested on the validation data. It is the difference between the predicted output and the actual output of the validation data.

Test error: The test error is the error that occurs when the model is tested on the test data. It is the difference between the predicted output and the actual output of the test data.

Bias error: Bias error occurs when a model is unable to represent the underlying structure of the data, resulting in an oversimplified model that performs poorly on both training and test data. It is the difference between the expected or true value and the predicted value.

Variance error: Variance error occurs when a model is too complex, resulting in a model that is highly sensitive to the noise in the training data and performs well on the training data but poorly on the test data. It is the measure of how much the predicted values of a model differ from the expected or true value.

Irreducible error: Irreducible error is the error that cannot be reduced even with the best possible model. It is the error that is due to the noise in the data and other unknown factors.

Type I and Type II error-

Type-I and Type-II errors are two different types of errors that can occur in statistical hypothesis testing.

Type-I error, also known as a false positive, occurs when we reject a null hypothesis that is actually true. In other words, it is the probability of rejecting the null hypothesis when it is actually true. The significance level, denoted by alpha, represents the probability of making a Type-I error. A lower significance level reduces the chance of making a Type-I error. For example, in a clinical trial, Type-I error may occur if a drug is deemed effective when it is actually not.

Type-II error, also known as a false negative, occurs when we fail to reject a null hypothesis that is actually false. In other words, it is the probability of failing to reject the null hypothesis when it is actually false. The power of the test, denoted by beta, represents the probability of making a Type-II error. A higher power of the test reduces the chance of making a Type-II error. For example, in a clinical trial, Type-II error may occur if a drug is deemed not effective when it is actually effective.

KNN -

Pros -

- Easy to understand

- No assumptions about data

- Can be applied to both classification and

Regression

- Works easily on multi-class problems

Cons-

Memory Intensive / Computationally expensive

Sensitive to scale of data

Struggle when high number of independent variables

K means

Pros

The K-means algorithm is simple and easy to implement. It can be applied to large datasets and can be executed quickly.

Can handle large datasets with high dimensionality.

Cons -

is sensitive to the initial conditions, which can result in different cluster assignments and different final results.

The number of clusters needs to be specified in advance. If the number of clusters is not known beforehand, then the algorithm can produce poor results.

Why do we call the Linear Regression a linear model?

We call Linear Regression a linear model **because it is based on a linear relationship between the independent variable(s) and the dependent variable.** In other words, the relationship between the independent and dependent variables can be represented by a straight line, which makes it a linear model.

In linear regression, we assume that the relationship between the input variables and the output variable is linear. The goal of linear regression is to find the best-fit line that can be used to predict the output variable based on the input variables. This best-fit line is represented by a linear equation of the form y = mx + c, where y is the output variable, x is the input variable, m is the slope of the line, and c is the y-intercept.

The linearity assumption makes it easier to estimate the parameters of the model using various statistical techniques, such as ordinary least squares. However, it is important to note that Linear Regression is a linear model only because of the linear relationship between the input and output variables. It is possible to have non-linear relationships between the variables, in which case linear regression would not be an appropriate model.

Pros and cons of Linear

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Easier to implement, interpret and efficient to train | It is often quite prone to noise and overfitting |
| It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation | Linear regression is quite sensitive to outliers |

What is regularization and it’s different types-

Regularization is a technique in machine learning that is used to **prevent overfitting** and **improve the generalization of models**. It involves **adding a penalty term** to the **loss function** of the **model during training**, which **encourages the model** to **choose simpler weights** and **avoid overemphasizing any particular feature in the data**.

In other words, regularization helps to constrain the model parameters and prevent them from becoming too complex, by adding a penalty term to the cost function that discourages overfitting. There are two main types of regularization:

L1 regularization (Lasso regression): This method adds an absolute value penalty to the sum of the weights in the model, which forces some of the coefficients to be exactly zero. **This results in a sparse model with fewer features and better interpretability.**

L2 regularization (Ridge regression): This method adds a squared penalty to the sum of the weights in the model, which shrinks all of the coefficients towards zero. This results **in a more robust model that is less sensitive to outliers and small variations in the data.**

Regularization is a powerful technique that can help to improve the performance of machine learning models, especially when dealing with high-dimensional data or datasets with a large number of features.

What does the alpha do in the regularization of models in ML

lasso = loss + α ||W||

In the regularization of models in ML, the alpha (α) parameter **controls the strength of regularization**. It is **a hyperparameter that is set before training the model**. Increasing the value of alpha will increase the strength of regularization, which shrinks the coefficients of the features towards zero and prevents overfitting. On the other hand, decreasing the value of alpha will decrease the strength of regularization, which allows the coefficients of the features to take larger values and may result in overfitting. The optimal value of alpha depends on the specific problem and dataset being used, and it is often found using techniques such as cross-validation. In the equation provided (lasso = loss + α ||W||), the ||W|| term represents the L1 norm of the weight vector W, and the value of alpha controls how much importance is given to this term in the overall objective function.

What do you mean by central tendency? Explain it with suitable example.

Central tendency is a statistical concept that describes the typical or **central value of a set of data**. It **represents a central value around which the data points tend to cluster**. There are three commonly used measures of central tendency: mean, median, and mode.

The mean is calculated by adding up all the data points and dividing by the number of data points. For example, if we have the following data on the heights (in cm) of 10 people:

165, 168, 172, 175, 177, 180, 182, 185, 188, 190

The mean height would be:

Mean = (165 + 168 + 172 + 175 + 177 + 180 + 182 + 185 + 188 + 190) / 10 = 177.2 cm

The median is the middle value when the data is arranged in order. In the example above, the median height would be:

Median = (175 + 177) / 2 = 176 cm

The mode is the value that occurs most frequently in the data. In the example above, there is no mode as no value appears more than once.

Central tendency is important in data analysis because it provides a way to summarize the data and get a sense of the typical value. However, it's important to keep in mind that the measure of central tendency that is most appropriate will depend on the nature of the data and the research question being asked.