Practical-1

Machine learning basics:

In this lab, we will go through the basics of machine learning. The student needs to make a soft copy note on the following topics:

Topics:

1. What is Machine learning?

Machine learning is a branch of artificial intelligence that focuses on the development of algorithms and models that enable computer systems to learn from and make predictions or decisions based on data.

2. Steps in collection of data

The process of collecting data for machine learning involves several important steps to ensure that the data is relevant, accurate, and sufficient for training and evaluating the models.

Key steps in the collection of data:

- Define the Problem and Objectives
- Identify Data Sources
- Data Gathering
- Data Quality Assurance
- Data Privacy and Security
- Data Labeling (for Supervised Learning)
- Data Transformation and Feature Engineering
- Data Split
- Data Bias and Fairness
- Data Documentation
- Data Versioning
- Data Exploration and Visualization

3. Steps in importing the data in python (Through: csv, json, and other data formats)

When working with Python, there are various libraries available that facilitate importing data from different formats such as CSV, JSON, Excel, and more.

While importing external files, we need to check the following points:

- Check whether header row exists or not
- Treatment of special values as missing values
- Consistent data type in a variable (column)
- Date Type variable in consistent date format.
- No truncation of rows while reading external data

4. Preprocessing

Preprocessing is a crucial step in data analysis and machine learning. It involves cleaning, transforming, and preparing raw data into a format suitable for further analysis and modeling. Preprocessing helps in improving the quality of the data, removing noise, dealing with missing values, and reducing the impact of outliers.

a) Remove Outliers

Removing outliers is an essential preprocessing step to handle data points that deviate significantly from the majority of the data. Outliers can adversely affect the performance of machine learning models, as they can introduce noise and bias the results. There are different approaches to remove outliers, and the choice depends on the nature of the data and the problem you are trying to solve.

b) Normalize Datasets, Data encoding

Normalization is the process of scaling numerical features in the dataset to a common range. This step is important because many machine learning algorithms work better when all features are on a similar scale. Two common methods for normalization are Min-Max Scaling and Z-score normalization (Standardization).

Data encoding is the process of converting categorical variables into numerical representations.

c) Handling Missing Data

Handling missing data is a crucial step in data preprocessing since real-world datasets often contain missing values due to various reasons. It's essential to deal with missing data appropriately to ensure the accuracy and reliability of your analysis or machine learning models.

5. Machine Models

a) Types of machine learning models – Supervised learning, Unsupervised learning, reinforcement learning.

Supervised learning is a type of machine learning where the algorithm learns from a labeled dataset. The goal of supervised learning is to learn a mapping function that can predict the correct output label for new, unseen input data.

Unsupervised learning is a type of machine learning where the algorithm learns from an unlabeled dataset. The goal of unsupervised learning is to find patterns or structures within the data without any explicit guidance.

Reinforcement learning is a type of machine learning where an agent learns to interact with an environment to achieve a specific goal. The agent receives feedback in the form of rewards or penalties based on its actions. The goal of reinforcement learning is to learn a strategy or policy that maximizes the cumulative reward over time.

b) Parameters of machine learning model (Learning rate, regularization, etc.)

The model has parameters that are learned during the training process. These parameters define the relationships between the input features and the output labels or predictions.

The learning rate is a hyperparameter that controls the step size at which a machine learning algorithm updates the model's parameters during the optimization process.

Regularization is a technique used to prevent overfitting in machine learning models. It adds a penalty term to the loss function, discouraging the model from becoming overly complex and fitting the noise in the training data.

6. Test-train data split: using constant ration, k-fold cross validation

Test-Train Data Split and K-Fold Cross-Validation are two common techniques used to evaluate the performance of machine learning models and assess their generalization ability.

The test-train data split using constant ration is a simple and widely used method to evaluate machine learning models. It involves dividing the available dataset into two separate sets: the training set and the test set. The test-train split is typically done using a constant ratio, such as 70%-30%, 80%-20%, or 90%-10%, depending on the size of the dataset.

K-Fold Cross-Validation is a resampling technique that helps provide a more robust estimate of a model's performance by repeatedly dividing the dataset into K subsets or "folds." The model is trained and evaluated K times, each time using a different fold as the test set and the remaining folds as the training set.

7. Output Inference

Output inference in the context of machine learning refers to the process of interpreting and analyzing the predictions made by a trained machine learning model. Once a model is trained on the training data and evaluated on the test data, it is deployed to make predictions on new, unseen data. The output of the model is typically a predicted label, value, or class probability.

8. Validation: different metrics – Confusion Matrix, Precision, Recall, F1score

Validation is an essential part of assessing the performance of a machine learning model. Different metrics are used to evaluate how well the model is making predictions and to gain insights into its strengths and weaknesses.

A confusion matrix is a table that summarizes the performance of a classification model. It shows the counts of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions.

Precision:

Precision is the proportion of correctly predicted positive instances (TP) out of all instances predicted as positive (TP + FP). It measures the model's accuracy for positive predictions.

Recall (Sensitivity or True Positive Rate):

Recall is the proportion of correctly predicted positive instances (TP) out of all actual positive instances (TP + FN). It measures the model's ability to correctly identify positive cases.

F1-Score:

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that takes both precision and recall into account. F1-score is useful when the data is imbalanced, i.e., one class dominates the other.