ST.FRANCIS INSTITUTE OF TECHNOLOGY

Mount Poinsur, S.V.P. Road, Borivli (West), Mumbai - 400103

Computer Engineering Department (2018-19)
Experiment No. 01

Aim:

To define problem statement of the mini project

To collect dataset for the problem statement

To do data preprocessing on collected dataset

Theory:

What is Machine Learning?

Machine learning is turning data into information. Machine learning lies at the intersection of computer science, engineering, and statistics and often appears in other disciplines. It can be applied to many fields from politics to geosciences. Any field that needs to interpret and act on data can benefit from machine learning techniques.

Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed. Two of the most widely adopted machine learning methods is supervised learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. A common use of supervised learning is to use historical data to predict statistically likely future events.

Example: An algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

Unsupervised Learning

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable. The goal of unsupervised learning is discovering hidden patterns within a dataset.

Example : With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Experiment:

Problem Statement:

Iris data set consists of the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris Setosa, Iris Virginica and Iris Versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

ML Tasks:

Classification: Find which of the IRIS flower type given: Sepal Length, Sepal Width, Petal Length, Petal Width

Computer Engineering Department (2018-19)
Experiment No. 01

Regression:

Predict the Sepal Width, given: Sepal Length, Petal Length, Petal Width

Clustering:

Form Three Clusters using any unsupervised learning algorithm, given: Sepal Length, Sepal Width, Petal Length, Petal Width

Dataset:

https://archive.ics.uci.edu/ml/datasets/iris

Installations:

import pip as pip

!pip install numpy

!pip install scipy

!pip install pandas

!pip install matplotlib

!pip install -U scikit-learn

Data Preprocessing

Cleaning and Filling Missing Data: Pandas provide different methods for filling the missing values. The fillna function can "fill in" NaN values with non-null data in different methods.

Conclusion: The problem statement for mini-project is finalized and data preprocessing experiments were performed on the dataset.