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.

Experiment:

Problem Statement:

In a financially volatile market, as the stock market, it is important to have a very precise prediction of a future trend. Because of the financial crisis and scoring profits, it is mandatory to have a secure prediction of the values of the stocks. Predicting a non-linear signal requires advanced algorithms of machine learning. The literature contains studies with different machine learning algorithms such as ANN (artificial neural networks) with different feature selection. The results of this study will show that the algorithm of classification SVM (Support Vector Machines) with the help of feature selection PCA (Principal component analysis) will have the success of making a profit.

ML Tasks:

Classification: Find whether at that day, company experienced profit or loss.

Regression:

Predict the Closing price for next few days.

Experiment No. 01

Clustering:

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

Data Preprocessing

```
import quandl
df = quandl.get('WIKI/AMZN')
```

```
df.head()
```

Date	Open	High	Low	Close	Volume	Ex-Dividend	Split Ratio	Adj. Open	Adj. High	Adj. Low	Adj. Close	Adj. Volume
1997-05-16	22.38	23.75	20.50	20.75	1225000.0	0.0	1.0	1.865000	1.979167	1.708333	1.729167	14700000.0
1997-05-19	20.50	21.25	19.50	20.50	508900.0	0.0	1.0	1.708333	1.770833	1.625000	1.708333	6106800.0
1997-05-20	20.75	21.00	19.63	19.63	455600.0	0.0	1.0	1.729167	1.750000	1.635833	1.635833	5467200.0
1997-05-21	19.25	19.75	16.50	17.13	1571100.0	0.0	1.0	1.604167	1.645833	1.375000	1.427500	18853200.0
1997-05-22	17.25	17.38	15.75	16.75	981400.0	0.0	1.0	1.437500	1.448333	1.312500	1.395833	11776800.0

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 5248 entries, 1997-05-16 to 2018-03-27
Data columns (total 12 columns):
Open           5248 non-null float64
High           5248 non-null float64
Low            5248 non-null float64
Close          5248 non-null float64
Volume         5248 non-null float64
Ex-Dividend    5248 non-null float64
Split Ratio    5248 non-null float64
Adj. Open      5248 non-null float64
Adj. High      5248 non-null float64
Adj. Low       5248 non-null float64
Adj. Close     5248 non-null float64
Adj. Volume    5248 non-null float64
dtypes: float64(12)
memory usage: 533.0 KB
```

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.