Data Science and Visualization

Assignment 1

Aim : Pre-processing techniques

Title : To apply pre-processing techniques on the raw dataset.

Problem Statement : Access an open source dataset “Titanic”. Apply pre-processing techniques on the raw dataset.

Theory :

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Dataset Description :

DATASET specs :

NAME: titanic3

TYPE: Census

SIZE: 1309 Passengers, 14 Variables

DESCRIPTIVE ABSTRACT:

The titanic3 data frame describes the survival status of individual passengers on the Titanic. The titanic3 data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

VARIABLE DESCRIPTIONS :

Pclass : Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

survival : Survival (0 = No; 1 = Yes)

name : Name

sex : Sex

age : Age

sibsp : Number of Siblings/Spouses Aboard

parch : Number of Parents/Children Aboard

ticket : Ticket Number

fare : Passenger Fare (British pound)

cabin : Cabin

embarked : Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

boat : Lifeboat

body : Body Identification Number

home.dest : Home/Destination

Steps Involved in Data Preprocessing :

1. **Data Cleaning :** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
2. **Data Reduction :** Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.
3. **Data Transformation :** This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:
    - I. **Normalization :** It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0).
    - II. **Attribute Selection :** In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
    - III. **Discretization :** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
    - IV. **Concept Hierarchy Generation :** Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Example :

Loading dataset in pandas

```
df = pd.read_csv('train.csv')
```

Dropping Columns which are not useful

```
cols = ['Name', 'Ticket', 'Cabin']
df = df.drop(cols, axis=1)
```

Dropping rows having missing values

```
df = df.dropna()
```

Creating Dummy Variables

```
dummies = []
cols = ['Pclass', 'Sex', 'Embarked']
for col in cols:
dummies.append(pd.get_dummies(df[col]))
```

```
titanic_dummies = pd.concat(dummies, axis=1)
```

And finally we concatenate to the original dataframe column wise

```
df = pd.concat((df,titanic_dummies), axis=1)
```

Conclusion : Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

References :

[1] https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[2] https://medium.datadriveninvestor.com/implementation-of-data-preprocessing-on-titanic-dataset-6c553bef0bc6

Data Science and Visualization

Assignment 2

Title : Predict Probability

Aim : To predict the probability of a survival of a person**.**

Problem Statement : Build training and testing dataset of assignment 1 to predict the probability of a survival of a person based on gender, age and passenger-class.

Dataset Description :

DATASET SPECS :

NAME: titanic3

TYPE: Census

SIZE: 1309 Passengers, 14 Variables

DESCRIPTIVE ABSTRACT:

The titanic3 data frame describes the survival status of individual passengers on the Titanic. The titanic3 data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

VARIABLE DESCRIPTIONS :

Pclass : Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

survival : Survival (0 = No; 1 = Yes)

name : Name

sex : Sex

age : Age

sibsp : Number of Siblings/Spouses Aboard

parch : Number of Parents/Children Aboard

ticket : Ticket Number

fare : Passenger Fare (British pound)

cabin : Cabin

embarked : Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

boat : Lifeboat

body : Body Identification Number

home.dest : Home/Destination

Theory :

Dataset Splitting :

The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results. By default, the Test set is split into 30 % of actual data and the training set is split into 70% of the actual data.

The **scikit-learn library** provides us with the model_selection module in which we have the splitter function train_test_split().

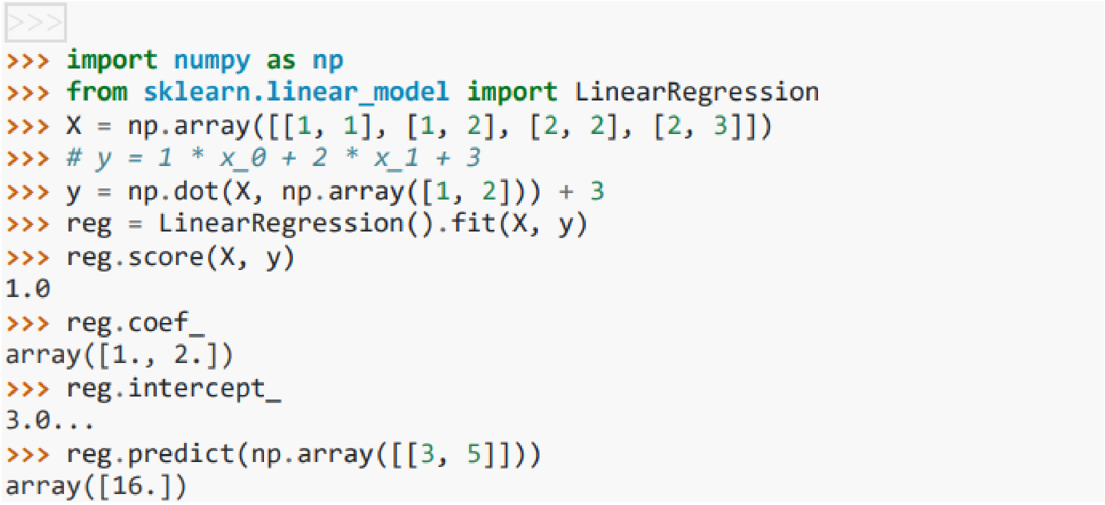
Syntax :

```
train_test_split(*arrays, test_size=None, train_size=None,
random_state=None, shuffle=True, stratify=None)
```

Regression :

Regression analysis is a statistical methodology that allows us to determine the strength and relationship of two variables. Regression is not limited to two variables, we could have 2 or more variables showing a relationship. The results from the regression help in predicting an unknown value depending on the relationship with the predicting variables. For example, someone's height and weight usually have a relationship. Generally, taller people tend to weigh more. We could use regression analysis to help predict the weight of an individual, given their height.

Linear Regression in sci-kit learn :



```
>>> import numpy as np
>>> from sklearn.linear_model import LinearRegression
>>> X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
>>> # y = 1 * x_0 + 2 * x_1 + 3
>>> y = np.dot(X, np.array([1, 2])) + 3
>>> reg = LinearRegression().fit(X, y)
>>> reg.score(X, y)
1.0
>>> reg.coef_
array([1., 2.])
>>> reg.intercept_
3.0...
>>> reg.predict(np.array([[3, 5]]))
array([16.])
```

Implementation :

Split the data into the target and feature variables.

```
X = titanic_data.drop(columns =['PassengerId','Name','Ticket','Survived'],axis=1)

Y = titanic_data['Survived']
```

Here, X is the feature variable, containing all the features like Pclass, Age, Sex, Embarked, etc. excluding the Survived column.

Y, on the other hand, is the target variable, as that is the result that we want to determine,i.e, whether a person is alive.

Now, we will be splitting the data into four variables, namely, X_train, Y_train, X_test, Y_test.

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, random_state=2)
```

Here 0.2 means that the data will be segregated in the X_train and X_test variables in a 80:20 ratio.

Create a model named model

```
model = LogisticRegression()

model.fit(X_train, Y_train)
```

Check for a random Person using random data

```
input_data = (3,0,35,0,0,8.05,0)

input_data_as_numpy_array = np.asarray(input_data)

input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
```

Predict using our model:

```
prediction = model.predict(input_data_reshaped)

print(prediction)
```

Conclusion : Linear regression model have been studied and implemented to predict probability of the survival of the person.

References :

[1] https://www.analyticsvidhya.com/blog/2021/07/titanic-survival-prediction-using-machine-learning/

[2] http://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pdf

Data Science and Visualization

Assignment 3

Title : Study of Abalone dataset.

Aim : To predict the age of abalone from physical measurements using linear regression.

Problem Statement : Load the data from data file and split it into training and test datasets. Summarize the properties in the training dataset. The number of rings is the value to predict: either as a continuous value or as a classification problem. Predict the age of abalone from physical measurements using linear regression or predict ring class as classification problem.

Dataset Description :

Dataset Information :

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

Attribute Information :

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Name / Data Type / Measurement Unit / Description

Sex / nominal / -- / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer / -- / +1.5 gives the age in years

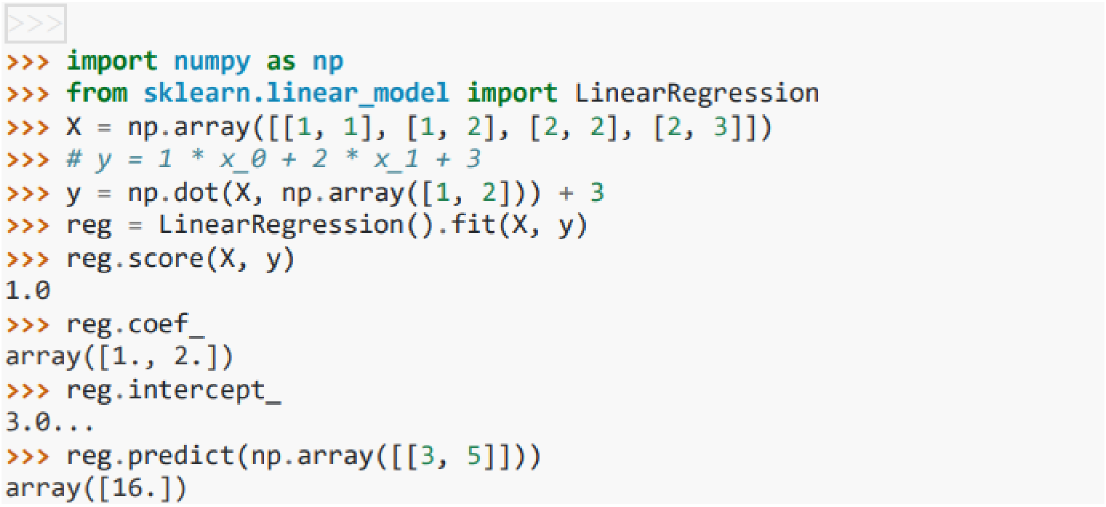
Theory :

Abalone is a shellfish considered a delicacy in many parts of the world. An excellent source of iron and pantothenic acid, and a nutritious food resource and farming in Australia, America and East Asia. 100 grams of abalone yields more than 20% recommended daily intake of these nutrients. The economic value of abalone is positively correlated with its age. Therefore, to detect the age of abalone accurately is important for both farmers and customers to determine its price. However, the current technology to decide the age is quite costly and inefficient. Farmers usually cut the shells and count the rings through microscopes to estimate the abalones age. Telling the age of abalone is therefore difficult mainly because their size depends not only on their age, but on the availability of food as well. Moreover, abalone sometimes form the so-called 'stunted' populations which have their growth characteristics very different from other abalone populations This complex method increases the cost and limits its popularity. Our goal in this report is to find out the best indicators to forecast the rings, then the age of abalones.

Regression :

Regression analysis is a statistical methodology that allows us to determine the strength and relationship of two variables. Regression is not limited to two variables, we could have 2 or more variables showing a relationship. The results from the regression help in predicting an unknown value depending on the relationship with the predicting variables. For example, someone's height and weight usually have a relationship. Generally, taller people tend to weigh more. We could use regression analysis to help predict the weight of an individual, given their height.

Example:



```
>>>
>>> import numpy as np
>>> from sklearn.linear_model import LinearRegression
>>> X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
>>> # y = 1 * x_0 + 2 * x_1 + 3
>>> y = np.dot(X, np.array([1, 2])) + 3
>>> reg = LinearRegression().fit(X, y)
>>> reg.score(X, y)
1.0
>>> reg.coef_
array([1., 2.])
>>> reg.intercept_
3.0...
>>> reg.predict(np.array([[3, 5]]))
array([16.])
```

Counters :

Counter is a container included in the collections module. It is a subclass of dict. Therefore it is an unordered collection where elements and their respective count are stored as a dictionary.

Syntax :

```
class collections.Counter([iterable-or-mapping])
```

Example:

```
# A Python program to show different ways to create
# Counter
from collections import Counter

# With sequence of items
print(Counter(['B','B','A','B','C','A','B','B','A','C']))

# with dictionary
print(Counter({'A':3, 'B':5, 'C':2}))

# with keyword arguments
print(Counter(A=3, B=5, C=2))
```

Implementation :

Importing dataset

```
import pandas as pd
col = ['sex','length','diameter','height','weight','sweight',
      'vweight','shweight','rings']
df = pd.read_csv('http://archive.ics.uci.edu/ml/machine-learning-
databases/abalone/abalone.data', names=col)
```

To count the number of rings make a new dataframe y

```
X = df.drop('rings', axis=1) #input
y = df['rings']    #output
```

Import Collection to count the number of rings and count rings using the counter function

```
from collections import Counter
Counter(y)
```

Split the data set into training and testing

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, random_state = 0, test_size = 0.20)
```

Now train the linear regression model

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression()
reg.fit(X_train, y_train)
y_pred = reg.predict(X_test)
```

Now measure the mean absolute error and r2 score to test the model’s accuracy.

```
from sklearn.metrics import mean_absolute_error
mean_absolute_error(y_test, y_pred)
```

```
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

Conclusion : Linear regression model have been studied and implemented to predict the age of abalone from physical measurements using abalone dataset.

References:

[1] https://www.geeksforgeeks.org/counters-in-python-set-1/

[2] https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset