Lab Examination on Machine Learning Model Implementation using Python (VAC-MLP 791)

Q1. Data Frame

1. Read the iris dataset at https://github.com/neurospin/pystatsml/tree/master/datasets/iris.csv (https://github.com/neurospin/pystatsml/tree/master/datasets/)

2. Print column names
3. Get numerical columns
4. For each species compute the mean of numerical columns and store it in a stats table with the following columns:

species sepal_length sepal_width petal_length petal_width

Q2. Write a Python program to give the weights and bias for a McCulloch-Pitts (M-P) neuron with inputs $\boldsymbol{x, y, z} \in \{-1, 1\}$ and whose output is $\boldsymbol{z}$ if $\boldsymbol{x} = -1$ and $\boldsymbol{y} = 1$ and is -1 otherwise

Q3. Data preparation:

Download heart dataset from following link.
https://www.kaggle.com/zhaoyingzhu/heartcsv (https://www.kaggle.com/zhaoyingzhu/heartcsv)

(https://www.kaggle.com/zhaoyingzhu/heartcsv)
Perform following operation on given dataset:

a) Find Shape of Data
b) Find Missing Values
c) Find data type of each column
d) Finding out Zero's
e) Find Mean age of patients
f) Now extract only Age, Sex, ChestPain, RestBP, Chol. Randomly divide the dataset in training (75%) and testing (25%).

Through the diagnosis test 100 reports were predicted as COVID positive, but only 45 of those were actually positive. Total 50 people in the sample were actually COVID positive. There were a total of

500 samples. Create confusion matrix based on above data and find

I. Accuracy
II. Precision
III. Recall
IV. F-1 score

Q4. Consider the following matrix, $\boldsymbol{M}$ and vector, $\boldsymbol{v}$:

$$\boldsymbol{M} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 0 \\ 1 & 3 & 3 \end{bmatrix} \qquad \boldsymbol{v} = \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$$

Compute the following matrix-vector and vector-vector products explaining how you arrived at each answer:

a) $\boldsymbol{M} \cdot \boldsymbol{v} =$

b) $\boldsymbol{v}^{\boldsymbol{T}} \cdot \boldsymbol{M} =$

c) $\boldsymbol{v}^{\boldsymbol{T}} \cdot \boldsymbol{v} =$

Install Python (https://www.python.org/ (https://www.python.org/)), Numpy (http://www.numpy. org/ (http://www.numpy.%20org/)) and Scipy (https://www.scipy.org/ (https://www.scipy.org/)) on your computer. Write a short program that defines $\boldsymbol{M}$ and $\boldsymbol{v}$ and computes the answers to the problem above using Numpy.

Q5. Regression Technique

Download temperature data from below link. https://www.kaggle.com/venky73/temperaturesof-india?select=temperatures.csv (https://www.kaggle.com/venky73/temperaturesof-india?select=temperatures.csv)

This data consists of temperatures of INDIA averaging the temperatures of all places month wise. Temperatures values are recorded in CELSIUS

a.Apply Linear Regression using suitable library function and predict the Month-wise temperature.
b. Assess the performance of regression models using MSE, MAE and R-Square metrics
c. Visualize simple regression model.

Q6. Classification Technique

Every year many students give the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is the target variable.

Data Set Available on Kaggle (The last column of the dataset needs to be changed to 0 or 1)

Data Set : **https://www.kaggle.com/mohansacharya/graduate-admissions (https://www.kaggle.com/mohansacharya/graduate-admissions)**

(https://www.kaggle.com/mohansacharya/graduate-admissions)
The counselor of the firm is supposed check whether the student will get an admission or not based on his/her GRE score and Academic Score. So to help the counselor to take appropriate decisions build a machine learning model classifier using Decision tree to predict whether a student will get admission or not.

A. Apply Data pre-processing (Label Encoding, Data Transformation….) techniques if necessary.

B. Perform data-preparation (Train-Test Split)
C. Apply Machine Learning Algorithm
D. Evaluate Model.

Q7. Clustering Techniques

Download the following customer dataset from below link:

Data Set: **https://www.kaggle.com/shwetabh123/mall-customers (https://www.kaggle.com/shwetabh123/mall-customers)**

(https://www.kaggle.com/shwetabh123/mall-customers)
This dataset gives the data of Income and money spent by the customers visiting a Shopping Mall. The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, as a

mallbowner youneed to find the group of people who are the profitable customers for the mall owner. Apply at least two clustering algorithms (based on Spending Score) to find the group of customers.

a. Apply Data pre-processing (Label Encoding , Data Transformation….) techniques if necessary.
b. Perform data-preparation(Train-Test Split)
c. Apply Machine Learning Algorithm
d. Evaluate Model.
e. Apply Cross-Validation and Evaluate Model

Q8. Association Rule Learning

Download Market Basket Optimization dataset from below link.
Data Set: **https://www.kaggle.com/hemanthkumar05/market-basket-optimization (https://www.kaggle.com/hemanthkumar05/market-basket-optimization)**

(https://www.kaggle.com/hemanthkumar05/market-basket-optimization)
This dataset comprises the list of transactions of a retail company over the period of one week. It contains a total of 7501 transaction records where each record consists of the list of items sold in one transaction. Using this record of transactions and items in each transaction, find the association rules between items.

There is no header in the dataset and the first row contains the first transaction, so mentioned header = None here while loading dataset.

Follow the steps given below:
a. Perform Data Preprocessing
b. Generate the list of transactions from the dataset
c. Train Apriori algorithm on the dataset
d. Visualize the list of rules

Q9. Clustering Techniques

Download the following customer dataset from below link:

Data Set: **https://www.kaggle.com/shwetabh123/mall-customers (https://www.kaggle.com/shwetabh123/mall-customers)**

(https://www.kaggle.com/shwetabh123/mall-customers)

This dataset gives the data of Income and money spent by the customers visiting a Shopping Mall. The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, as a mall owner you need to find the group of people who are the profitable customers for the mall owner. Apply at least two clustering algorithms (based on Spending Score) to find the group of customers.

a. Apply Data pre-processing (Label Encoding , Data Transformation….) techniques if necessary.
b. Perform data-preparation(Train-Test Split)
c. Apply Machine Learning Algorithm
d. Evaluate Model.
e. Apply Cross-Validation and Evaluate Model

Q10. Apply PCA on the iris dataset

The data set is available at:
https://github.com/duchesnay/pystatsml/raw/master/datasets/iris.csv (https://github.com/duchesnay/pystatsml/raw/master/datasets/iris.csv)

a. Describe the data set. Should the dataset been standardized ?
b. Describe the structure of correlations among variables.
c. Compute a PCA with the maximum number of components K by your computed values.

d. Print the K principal components directions and correlations of the K principal components with the original variables. Interpret the contribution of the original variables into the PC.
e. Plot the samples projected into the K first PCs.
f. Color samples by their species.