Machine Learning
Laboratory 01

Submission Due: 7-Aug-2019

Submission Procedure:

1. Compress all files into a zip file named “Lab1_XXX.zip”
2. Email “Lab_XXX.zip” to xavier.xhq@qq.com with title “Machine Learning: Lab1_XXX”. Write down your name and student ID in the email.

XXX is your first name + last name in small letter without spacing.
E.g. Chen Tao : Lab1_chentao.zip

Exercise 1

A dataset D can be found in the file "lab1_Ex1_data.csv". You are required to finish exercise 1 using NumPy and Pandas.

All the answers (i.e. values or datasets) should be stored in one dictionary variable named Lab1_Ex1. Save the dictionary to a "pkl" file named as "Lab1_Ex1.pkl".

1. Find out the maximum of each numerical feature. (Store the answer to key "Q1")
2. Find out the mean of each numerical feature. (Store the answer to key "Q2")
3. How many different feature values of "Sex". (Store the answer to key "Q3")
4. How many missing values (i.e. nan) in the dataset? (Store the answer to key "Q4")
5. Replace the missing values of "Age" by the median of "Age". (Store the new dataset to key "Q5")
6. Create a new dataset which containing the records of D containing "male" in "Sex" feature. The samples should be sorted according to "PassengerId" in an ascending order. (Store the new dataset to key "Q6")
7. Create a new dataset which containing the records of D containing "male" in "Sex" feature and Age > 20. The samples should be sorted according to "PassengerId" in an ascending order. (Store the new dataset to key "Q7")
8. Create a new feature to D, named "newSex" to store the following data. The samples should be sorted according to "PassengerId" in an ascending order. (Store the new dataset to key "Q8")
   a) If "Sex" = "male", the value is 1
   b) If "Sex" = "female", the value is 2
9. Create three new features to D, named "nE1", "nE1", and "nE3", to store the following data. The samples should be sorted according to "PassengerId" in an ascending order. (Store the new dataset to key "Q9") (* this question is independent to Q6)
   a) If "Embarked" = "Q", the values are 1, 0, and 0 for "nE1", "nE1", and "nE3"
   b) If "Embarked" = "S", the values are 0, 1, and 0 for "nE1", "nE1", and "nE3"
   c) If "Embarked" = "C", the values are 0, 0, and 1 for "nE1", "nE1", and "nE3"

The structure of dictionary should be as the same as follows. Be noted that it is just an example on structure and the values are not the real answers.

```
{
    "Q1": {
        "Pclass": 5,
        "Age": 99,
        "SibSp": 5,
        "Parch": 5,
        "Fare": 999.9
    },
```

```
    "Q2": {
        "Pclass": 3.33,
        "Age": 44.44,
        "SibSp": 3.33,
        "Parch": 3.33,
        "Fare": 444.44
    },
    "Q3": 5,
    "Q4": 1000,
    "Q5": DataFrameObject,
    "Q6": DataFrameObject,
    "Q7": DataFrameObject,
    "Q8": DataFrameObject,
    "Q9": DataFrameObject
}
```

Exercise 2

A 10-class image dataset can be found in the file “lab1_Ex2_data.zip”. 1001 images can be found after unzip. ID_X_YYY.png represents a training sample, while test.png is a test sample, where ID is the sample ID from 0001 to 1000, X is the class ID from 0, 1, ..., 9, YYY is the sample number from 001, 002, ..., 100. You are required to finish exercise 2 using NumPy and Matplotlib.

All the answers (i.e. values or IDs) should be stored in one dictionary variable named Lab1_Ex2. Save the dictionary to a “pkl” file named as “Lab1_Ex2.pkl”, and save the image of Q4 as “png” file named as “Lab1_Ex2.png”.

1. Calculate the Euclidean distances between the test sample and all training samples. The answer should be stored in an array. The i^{th} value represents the distance to the i^{th} training sample (sorted by ID in an ascending order) (Store the answer to key “Q1”)
2. Find out the most similar image from the training set to the test sample. What is the class ID of the test sample if 1-nn is used as a classifier? (Store the sample ID of the most similar image to key “Q2_1”, and store the class ID of the test sample to key “Q2_2”)
3. Find out the most three similar images from the training set to the test sample. What is the class ID of the test sample if 3-nn is used as a classifier? (Store the sample IDs of the three most similar images as a set to key “Q3_1”, and the class ID of the test sample to key “Q3_2”)
4. Plot the most three similar images and the test sample on one figure. Set each image title as its file name. (Save the figure as a file)

The structure of dictionary should be as the same as follows. Be noted that it is just an example on structure and the values are not the real answers.

```
{
    "Q1": array([3.456, 4.5674, ..., 2.234]),
    "Q2_1": '0001',
    "Q2_2": 0,
    "Q3_1": {'0001', '0002', '0003'},
    "Q3_2": 0
}
```

Exercise 3

"lab1_Ex3_data.zip" contains a training set (lab1_Ex3_train.csv) and a test set (lab1_Ex3_test.csv) of a 2-class problem. We assume prior probabilities of both classes are 0.5. The samples in each class follow the Gaussian distribution. A Bayes Classifier using Maximum Likelihood Parameter should be implemented.

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

You are required to finish this exercise by using NumPy, Pandas and Matplotlib.

All the answers (i.e. values) should be stored in one dictionary variable named Lab1_Ex3. Save the dictionary to a "pkl" file named as "Lab1_Ex3.pkl", and save the image of Q4 as "png" file named as "Lab1_Ex3.png"

1. Write a function for a Bayes Classifier by using Maximum Likelihood Parameter Estimation (you should assume the features are independent). Here is the function header. (Store the answer to key "Q1")

   ```
   def bayes (train_data, train_label, test_data)
       return test_label
   ```

   (Parameters)

   `train_data`: an array of size by n times 2, contains data of n training samples

   `train_label`: an array of size by n times 1, contains labels of n training samples

   `test_data`: an array of size by m times 2, contains data of m testing samples

   (Return value)

   `test_label`: an array of size by m times 1, contains labels of m testing samples
2. What is the training error? (Store the answer to key "Q2")
3. Determine the class IDs of the test samples in an NumPy array (sorted according to "test id" in an ascending order) (Store the answer to key "Q3")
4. Plot a figure containing (Save the figure as a file)
   a) the decision boundary of the Bayes Classifier (solid line, in black)
   b) Training samples of class 1 (blue circle)
   c) Training samples of class 2 (green square)
   d) Test samples (red triangle)

The structure of dictionary should be as the same as follows. Be noted that it is just an example on structure and the values are not the real answers.

```
{
    "Q1":    <function    __main__.bayes(train_data,    train_label,
test_data)>,
    "Q2": 0.123,
    "Q3": array([1, 0, ..., 1])
}
```