Classification and Regression

The first assignment

Introduction and Important Date

Marks: 5

Due on: 23:59:59, 2019/04/03

Purpose: Use **Linear models/SVMs/Neural networks** to deal with **a specific data set**.

Notes: The key skill to train in this experiment is pre-processing the data, not to implement different models. Therefore, **the students can use models implemented by frameworks.**

Intention The data in real world is in different format or even **wrong** format. Which means in most of the time, data processing is a very important and time consuming work. In contract, with the help of many machine learning frameworks, people can utilize various models by several lines of code. Therefore, as a data scientist, sometimes processing data cost much more effort than implementing a machine learning model. In the serial assignments, we want the students to learn some basic ideas of data processing. In this assignment, we want the students to train these abilities:

1. How to read data in a unusual format specific;
2. How to deal with attributes with different ranges of values.
3. How to quick implement some common machine learning models by utilizing the documents of sklearn in internet.

Details

Data: check data folder.

Tasks:

1. Use "**Reason for absence**" as the label, adopt a machine learning model to predict the label for a given instance.
2. Use "**Absenteeism time in hours**" as the target, adopt a machine learning model to predict the target of a given instance.

The assignment contains 3 steps: analysis the data, pre-process the data and classification/regression.

Analysis: Read the dataset description to figure out what the meaning of the attributions.

Pre-process:

1) Read the csv format data; **Note that the data cannot be read by existing APIs (like pandas.read_csv)**, you need to found out why and read it (Hint: pay attention to the type of separator of the CSV format).

2) some values of attributes are very large (>100) and some are small (about 0 or 1), use normalization to process the data attributes;

3) Should some attributes be transformed to other styles? For example, how about using one-hot coding for the date related attributes?

Classification and regression: Which model performs the best in each task? Try it.

Report **micro-average F1 scores** (https://blog.csdn.net/sinat_28576553/article/details/80258619?utm_source=blogxgwz8) for the classification task and report **mean squared error** (MSE) for the regression task.

Note: We provide the **train.csv** for training your model and the **valid.csv** for validating your model. HOWEVER, we keep a **test.csv** for testing your code, and the test.csv will be published after the submission is finished.

Requirement

The code should accept the path of the data file as input, take python as the example (Similar for other languages):

```
$> python exp.py ./train.csv ./valid.csv
```

The **first** file is the train dataset the **second** is the validation data set or the test data set.

And output the micro-average F1 scores and the MSE of regression like this:

```
Micro-average F1 of classification:
57%
Mean squared error of regression:
1.2
```

The provided package should contain:

- one simple report
- All the source code
- The additional components for running the code

The **simple report** should contain at least 3 section:

1. how you pre-process the data
2. how you design the machine learning models (For example, how many neurons and layers in a neural network, shown these parameters in tables)
3. the results.

You can write the report follow the style of 《软件学报 2016 年排版样例.doc》, and feel free to use English or Chinese.

The **provided code** should be self-contained. **Specifically**:

1. If the code use all the packages contained in Anaconda 3, no additional packages are needed;
2. If the code use other python packages that can be download by pip, provide a requirement.txt, check the blog (https://www.cnblogs.com/zhaoyingjie/p/6645811.html) for an example;
3. If the code based on other machine learning libraries (which also provide python install packages) like pytorch/mxnet/tensorflow/theano et al., please provide the specific .whl packages of the used framework and list the environment (Windows/linux, the version of the system/python) in the report;
4. Otherwise, provide the detailed steps of installing the running environment of the code in the report.

For simplicity, we suggest to use **sklean** directly.

Remarks

1. **python exp.py ./train.csv ./valid.csv** runs successfully and outputs the micro-average F1 scores and the MSE, 3 marks.
2. **python exp.py ./train.csv ./test.csv (The test.csv will be released after submission)** runs successfully and outputs the micro-average F1 scores and the MSE, 1 marks.
3. If the micro-average F1 score or MSE is in the **top 90% of all the submissions,** 1 marks.
4. If the work is not submitted in time, it will not be accepted anymore and **-1 marks.**

To-do List

- The code for classification and regression.
- The report to explain your work.

How to Submit

Send an email with a zip file, the report and the code should be in the zip package.

Contact: 雷云文

Email address: leiyw@sustc.edu.cn