LOGISTIC REGRESSION

田思成 工程科学学院 U201714461

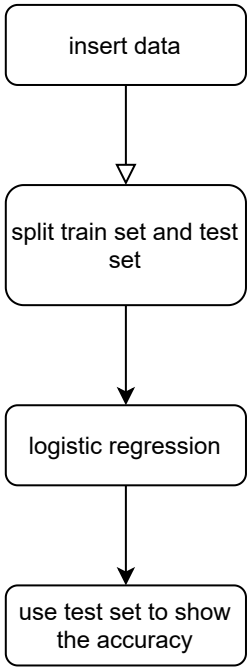
实验目的

1. 了解逻辑回归的概念

试验内容

使用逻辑回归的办法对乳腺癌数据进行处理，并且把数据分为训练集和测试集以便于测试。最后验证logistic regression的准确度。

流程如下:



代码如下:

```
#another way to insert data of breast cancer
from sklearn.datasets import load_breast_cancer
cancer = load_breast_cancer()
#show the data's format, so that we can deal with data
#print(cancer.data.shape)
#print(cancer.target.shape)
#show details of data
#print(cancer.DESCR)
#malignant = 0, benign = 1 if you wish, use print(cancer.target)
#split the train and test dataset
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(cancer.data,
cancer.target, test_size=0.25, random_state=42)
# train_data：所要划分的样本特征集
# train_target：所要划分的样本结果 此处为Benign or malignant
# test_size：样本占比，如果是整数的话就是样本的数量
# random_state：是随机数的种子。随机抽取，random_state 保证每次数据可以重复。
```

```
# 随机数种子：其实就是该组随机数的编号，在需要重复试验的时候，保证得到一组一样的随机数。比如你每次都填1，其他参数一样的情况下你得到的随机数组是一样的。但填0或不填，每次都会不一样。
from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
pred = log_reg.predict(X_test)
acc_score = log_reg.score(X_test, y_test)
print(acc_score)
list(cancer.target_names)
import pandas as pd
d = {'predictions': pred, 'real values': y_test}
data = pd.DataFrame(data=d)
print(data)
data.predictions == data['real values']
wrong_predictions= []
for i in range(0,143):
    if data.predictions[i] != data['real values'][i]:
        wrong_predictions.append(data.predictions[i])
        print("wrongly diagnosed patient number:", i, 'as', wrong_predictions[-1])
    i=i+1

```

代码中注释已经写得较为清楚，便不对思路进行赘述。

实验结果

```
0.965034965034965
     predictions  real values
0              1            1
1              0            0
2              0            0
3              1            1
4              1            1
..           ...          ...
138            1            1
139            0            0
140            1            1
141            0            0
142            1            1

[143 rows x 2 columns]
wrongly diagnosed patient number: 20 as 1
wrongly diagnosed patient number: 58 as 1
wrongly diagnosed patient number: 77 as 1
wrongly diagnosed patient number: 112 as 0
wrongly diagnosed patient number: 120 as 0
PS C:\Users\Administrator\Desktop\大三下\机器学习\每周试验\2nd\Breast-Cancer-predictions-master>
```

可以注意到，在验证集占比为25%的情况下，准确度为96.5%左右。其中错误的个体序号分别为：20,58,77,112,120

试验总结和反思

本次试验主要采取了panda, sklearn这两个模块进行测试，实际上有投机取巧的意思，因为内置了logistic regression的函数，所以甚至不用大费周章的计算数据。但是我认为造轮子要比借轮子好，之后有时间会做一个自己提取数据进行计算的版本。

但总的来说还是比较好的完成了任务，可以较为完善的体现logistic regression的含义。