

七月在线 python 基础第四节课作业

Author: 粽子

2016/11/11

第四次作业主要就是一道题，就是根据网上下载的数据集，分成训练集和测试集，然后用训练集训练分类模型，用测试集来检测模型的好坏。（以上这些老师都给出代码了）

然后我们的作业就是用计算模型的准确率。

（实现这个其实没什么难的，我觉得主要是检测对现有代码的理解）

前面 64 行都是老师给出的代码，我的实现从第 65 行开始，一共就 20 行。

主要就是做了两件事情：

- 1、把 X_test 这个列表中的所有元素都扔进模型里训练，得到的结果存在一个 list 里面

predict_result

- 2、根据 predict_result 和 y_test 这两个 list 中相同元素的个数来计算准确率。

最后算出的准确率见最后的图： 0.3722222222223

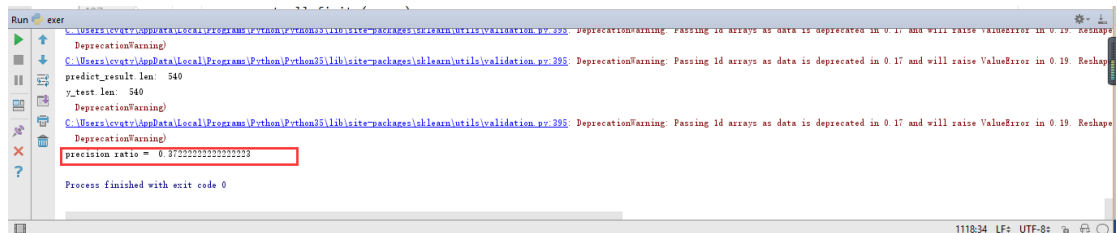
```
1 #Lesson4 Homework
2 from sklearn import svm, datasets
3
4 class Dataset:
5     # 我们创建一个 dataset 的类，这个类会帮我们下载相关的数据集，
6     # 并给我们分类好 x,y
7     def __init__(self, name):
8         # 告诉类，我们需要哪一个数据集
9         # 我们有两个选择，一个是 'iris' 一个是 'digits'
10        self.name = name
11
12    def download_data(self):
13        # 从 sklearn 的自带集中下载我们指定的数据集
14        if self.name == 'iris':
15            # 这里是 sklearn 自带的数据集下载方法，更多信息可以参照官网
16            self.downloaded_data = datasets.load_iris()
17        elif self.name == 'digits':
18            self.downloaded_data = datasets.load_digits()
19        else:
20            # 如果不是我们预想的两种数据集，则报错
21            print('Dataset Error: No named datasets')
22
23    def generate_xy(self):
24        # 通过这个过程来把我们的数据集分为原始数据以及他们的 Label
25        # 我们先把数据下载下来
26        self.download_data()
27        x = self.downloaded_data.data
28        y = self.downloaded_data.target
```

```

29     print('\nOriginal data looks like this: \n', x)
30     print('\nLabels looks like this: \n', y)
31     print('\nx: ', len(x), ' ', len(x[0]))
32     print('\ny: ', len(y))
33     return x, y
34
35     def get_train_test_set(self, ratio):
36         # 这里, 我们把所有的数据分成训练集和测试集
37         # 一个参数要求我们告知, 我们以多少的比例来分割训练和测试集
38         # 首先, 我们把XY 给generate 出来:
39         x, y = self.generate_xy()
40
41         # 有个比例, 我们首先得知道 一共有多少的数据
42         n_samples = len(x)
43         # 于是我们知道, 有多少应该是训练集, 多少应该是测试集
44         n_train = n_samples * ratio
45         # 好了, 接下来我们分割数据
46         X_train = x[:n_train]
47         y_train = y[:n_train]
48         X_test = x[n_train:]
49         y_test = y[n_train:]
50         # 好, 我们得到了所有想要的玩意儿
51         return X_train, y_train, X_test, y_test
52         # ===== 我们的 dataset 类创造完毕=====
53
54     # 比如, 我们使用 digits 数据集
55     data = Dataset('digits')
56     # 接着, 我们可以用0.7 的分割率把xy 给分割出来
57     X_train, y_train, X_test, y_test = data.get_train_test_set(0.7)
58     clf = svm.SVC()
59     print(clf.fit(X_train, y_train))
60     test_point = X_test[12]
61     y_true = y_test[12]
62     print(clf.predict(test_point))
63     print(y_true)
64
65     #根据训练出来的模型, 把X_test 中的每一个点都放在预测器中做预测, 然后放在一个list 里
66     #面(假设我的list 叫做predict_result)
67     predict_result = []
68     for element in X_test:
69         predict_result.append(clf.predict(element))
70     print('predict_result.len: ', len(predict_result))
71     print('y_test.len: ', len(y_test))
72

```

```
73 #计算准确率:
74 if len(predict_result) != len(y_test):
75     print('something wrong with the calculation, predict_result.len: ',
76         len(predict_result),
77         'y_test.len:%d', len(y_test))
78 else:
79     cnt = 0;
80     for i in range(0, len(predict_result), 1):
81         if predict_result[i] == y_test[i]:
82             cnt += 1
83
84 precision_ratio = cnt / len(predict_result)
85 print('precision ratio = ', precision_ratio)
```



```
Run: exec
C:\Users\crazy\AppData\Local\Programs\Python\Python35\lib\site-packages\sklearn\utils\validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise ValueError in 0.19. Reshape
predict_result len: 540
y_test len: 540
C:\Users\crazy\AppData\Local\Programs\Python\Python35\lib\site-packages\sklearn\utils\validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise ValueError in 0.19. Reshape
precision ratio = 0.3722222222222222
Process finished with exit code 0
1118:34 LF+ UTF-8: 11
```