

第一次作业

- 秦绪博
- 2017000621

数据集和模型概述

使用`sklearn`自带的数据包提供的手写体数字识别数据集`load_digits`。输入`digits.data.shape`可以看到数据规模为(1797,64)。即一共有1797组数据，每幅图片由 $8 \times 8 = 64$ 的像素矩阵表示。

使用`sklearn`自带模块进行交叉验证数据切分，按照训练集：测试集=3:1的规模进行分割。

本次实验使用基于线性假设的支持向量机分类器`LinearSVC`，随机森林分类器`RandomForestClassifier`和`XGBOOST`分类器`XGBClassifier`。

代码

```
from sklearn.datasets import load_digits
# 从通过数据加载器获得手写体数字的数码图像数据并储存在digits变量中。
digits = load_digits()

from sklearn.model_selection import train_test_split

# 随机选取75%的数据作为训练样本；其余25%的数据作为测试样本。
X_train, X_test, y_train, y_test = train_test_split(digits.data,
                                                    digits.target, test_size=0.25, random_state=33)

from sklearn.preprocessing import StandardScaler

# 从sklearn.svm里导入基于线性假设的支持向量机分类器LinearSVC。
from sklearn.svm import LinearSVC

# 从仍然需要对训练和测试的特征数据进行标准化。
ss = StandardScaler()
X_train = ss.fit_transform(X_train)
X_test = ss.transform(X_test)

# 初始化线性假设的支持向量机分类器LinearSVC。
lsvc = LinearSVC()
#进行模型训练
lsvc.fit(X_train, y_train)
# 利用训练好的模型对测试样本的数字类别进行预测，预测结果储存在变量y_predict中。
lsvc_y_predict = lsvc.predict(X_test)
print('The Accuracy of Linear SVC is', lsvc.score(X_test, y_test))
from sklearn.metrics import classification_report
print(classification_report(y_test, lsvc_y_predict))

from sklearn.ensemble import RandomForestClassifier
```

```

rfc=RandomForestClassifier()
rfc.fit(X_train,y_train)
rfc_y_predict=rfc.predict(X_test)

print ('The Accuracy of Random Forest is', rfc.score(X_test, y_test))
from sklearn.metrics import classification_report
print (classification_report(y_test, rfc_y_predict))

from xgboost import XGBClassifier
xgbc=XGBClassifier()
xgbc.fit(X_train,y_train)
xgb_y_pred = xgbc.predict(X_test)
xgb_predictions = [round(value) for value in xgb_y_pred]
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, xgb_predictions)
print("The Accuracy of XGB00ST is: %.2f%%" % (accuracy * 100.0))
print (classification_report(y_test,xgb_predictions))

```

输出结果

使用sklearn自带的classification_report模块输出报告：

```

The Accuracy of Linear SVC is 0.95333333333333
      precision    recall  f1-score   support

0         0.92        1.00        0.96         35
1         0.96        0.98        0.97         54
2         0.98        1.00        0.99         44
3         0.93        0.93        0.93         46
4         0.97        1.00        0.99         35
5         0.94        0.94        0.94         48
6         0.96        0.98        0.97         51
7         0.92        1.00        0.96         35
8         0.98        0.84        0.91         58
9         0.95        0.91        0.93         44

avg / total         0.95        0.95        0.95        450

The Accuracy of Random Forest is 0.9355555555556
      precision    recall  f1-score   support

0         0.92        1.00        0.96         35
1         0.93        0.96        0.95         54
2         0.95        0.95        0.95         44
3         0.95        0.85        0.90         46
4         0.94        0.91        0.93         35
5         0.87        0.94        0.90         48
6         1.00        0.98        0.99         51
7         1.00        1.00        1.00         35
8         0.90        0.93        0.92         58

```

9	0.93	0.84	0.88	44
avg / total	0.94	0.94	0.94	450
The Accuracy of XGB00ST is: 94.89%				
	precision	recall	f1-score	support
0	0.92	1.00	0.96	35
1	0.96	0.98	0.97	54
2	0.98	0.93	0.95	44
3	0.98	0.91	0.94	46
4	0.97	0.94	0.96	35
5	0.96	0.90	0.92	48
6	0.96	0.96	0.96	51
7	0.92	0.97	0.94	35
8	0.93	0.95	0.94	58
9	0.91	0.95	0.93	44
avg / total	0.95	0.95	0.95	450