链滴

城市计算 AI 挑战赛 - 算法选择入门

作者：cttmayi

原文链接：https://ld246.com/article/1573145457499

来源网站：链滴

算法选择

参考scikit-learn算法选择图进行算法选择

1. 数据量大于50(>50 samples)

Yes

2. 预测类别(predicting a **category**)

No

3. 预测数量(predicting a **quantity**)

Yes

4. 数据量小于100K(<100K samples)

这个问题比较难回答, 整体数据有大于100K, 数据简化整理以后, 发现数据远远小于100K. 暂定回答为**Y**
s

5. 有一些特征很重要(few features should be important)

问题太抽象,不懂. 先回答**No**吧

6. 线性支持向量机(SVR(kernel="linear"))

就先选定它吧, 参考文档

特征选择

第一个版本, 先选择简单点. IN,OUT和星期几, 和当前的时间, 站点ID有关

特征:

-- 特征输入

星期几, 时间(小时, 分钟), 站点ID

-- 结果输出

IN, OUT

数据准备

```
train_result = train.groupby(['stationID','wkday','days','hours','minutes'])['status'].sum().to_fra
e('inNums').reset_index()
train_result['outNums'] = train.groupby(['stationID','wkday','days','hours','minutes'])['status'].c
unt().values
train_result['outNums'] = train_result['outNums'] - train_result['inNums']

import numpy as np
test_data = train_result.loc[:,['stationID', 'wkday', 'hours', 'minutes']]
test_data = test_data.values

test_target = train_result.loc[:,['inNums']]
test_target = test_target.values
test_target = test_target.reshape(test_target.size)

from sklearn.model_selection import train_test_split
```

```
test_data_A, test_data_B, test_target_A, test_target_B = train_test_split(test_data, test_target, t
st_size=0.1, random_state=0)
```

SVR

```
from sklearn import svm

clf = svm.SVR(kernel='linear')
clf.fit(test_data_A, test_target_A)

pred = clf.predict(test_data_B)
pred = pred.astype(np.int64)
```

评估答案

平均绝对误差: 84.7, 成绩一般, 符合预期吧, 再想办法进一步改进.

```
# 用mean_absolute_error, 评估答案
from sklearn import metrics
import matplotlib.pyplot as plt
plt.figure(figsize=(24, 13))

plt.plot(range(0,len(pred)), pred - test_target_B)

metrics.mean_absolute_error(test_target_B, pred)
```