

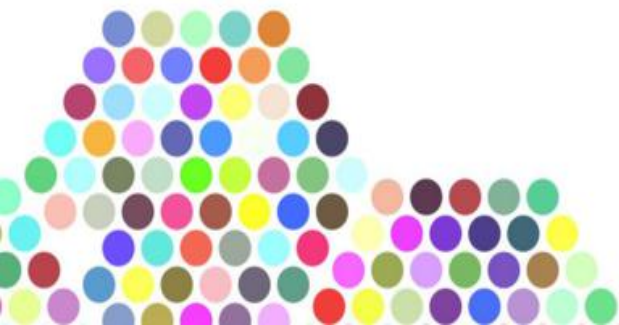
What's cooking Homework

The Home of Data Science

COMPETITIONS • CUSTOMER SOLUTIONS • JOBS BOARD

Get started »

DM14301033



1:preparation:

- environment: Linux Ubuntu 15.04
- tool: python2.7

• 流程:



2: begin

```
[ { "id": 10259, "cuisine": "greek", "ingredients": [
  "olives", "grape tomatoes", "garlic", "pepper", "id",
  "garbanzo beans", "feta cheese crumbles" ] }, { "id": 20130, "cuisine": "filipino", "ingredients": [
  "plain flour", "ground pepper", "black pepper", "thyme", "eggs", "green tomatoes",
  "vegetable oil" ] }, { "id": 22213, "cuisine": "indian", "ingredients": [
  "water", "vegetable oil", "wheat", "salt" ] }, { "id": 6602, "cuisine": "jamaican", "ingredients": [
  "black pepper", "shallots", "pepper", "onions", "garlic paste", "milk", "butter",
  "juice", "water", "chili powder", "passata", "boneless chicken skinless thigh", "garam masala", "dough",
  "bay leaf" ] }, { "id": 6602, "cuisine": "jamaican", "ingredients": [
  "sugar", "butter", "eggs", "fresh ginger root", "..." ] }
```

- 我们拿到了两个数据集， **test.json** 以及 **train.json**
- 目的： **upup.csv**

3.deal with the json

- we use "json.load()" to read the json into the python,and get the datas:
- 字符串分离: "hot milk","salt"->"hot milk salt" or "hot_milk salt"
- then we use vectorize(掉包) 来进行向量化:
- 其他处理方法: 控制train的输入:

4.method -----begin with logistic

- 对于方法，我一开始掉包使用logistic_regression.
- 结果：77.3;
- 后来自己实现了logistic_regression:
- 结果：73;

5.others method

- 我遇到了起初的瓶颈，与上层建筑差太多。

- 于是，我多样性尝试各种方法

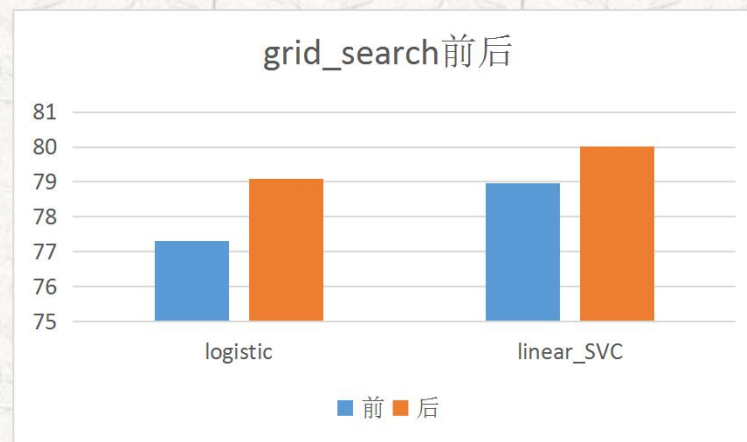
- | | |
|----|----------------------------------------|
| 正常 | <code>grid_search()</code> 【一种自动调参数机制】 |
|----|----------------------------------------|

- random_forest: 69.3

- logistic 77.3 78.8

- linear_SVC: 78.9 79.1

- bys 74.7



6,then

- 后来，我又尝试了一开始的xgboost: 79.0
- 再进行了excel下的vote:
- xgboost+linearSVC+randomforest+bys(myself) 79.99 better

8,next

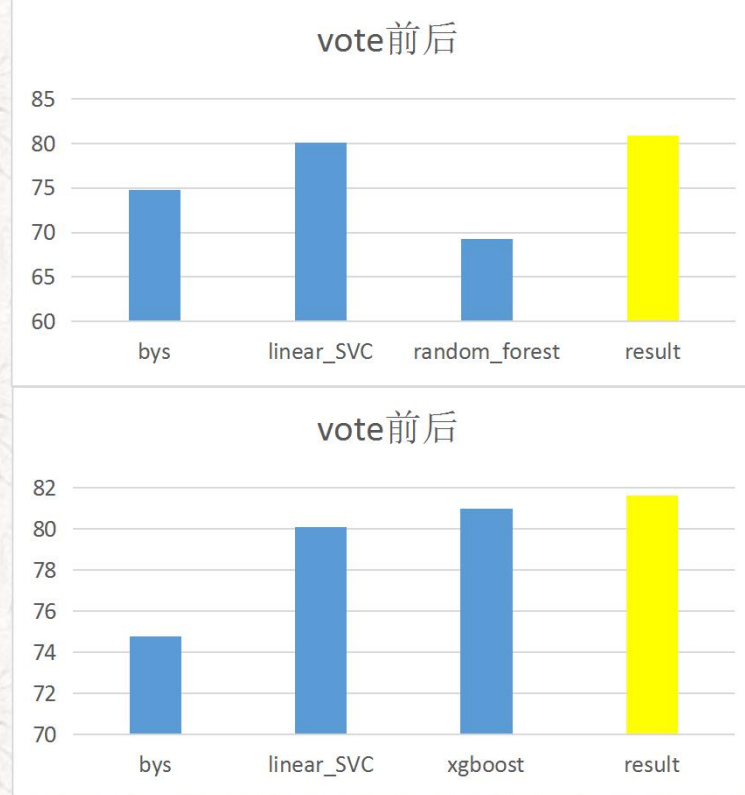
- 剔除必要的干扰项
- 进行再优化，而且利用不同方式进行训练，取训练集和测试集中交集作为菜式，进行学习和分类。
- **xgboost**改良：结合**grid_search**,先提取**ingredient**特征，再训练特征模型，之后再将特征模型整合进菜系模型

9, better vote

- 有了更好的子项，我再进行vote。

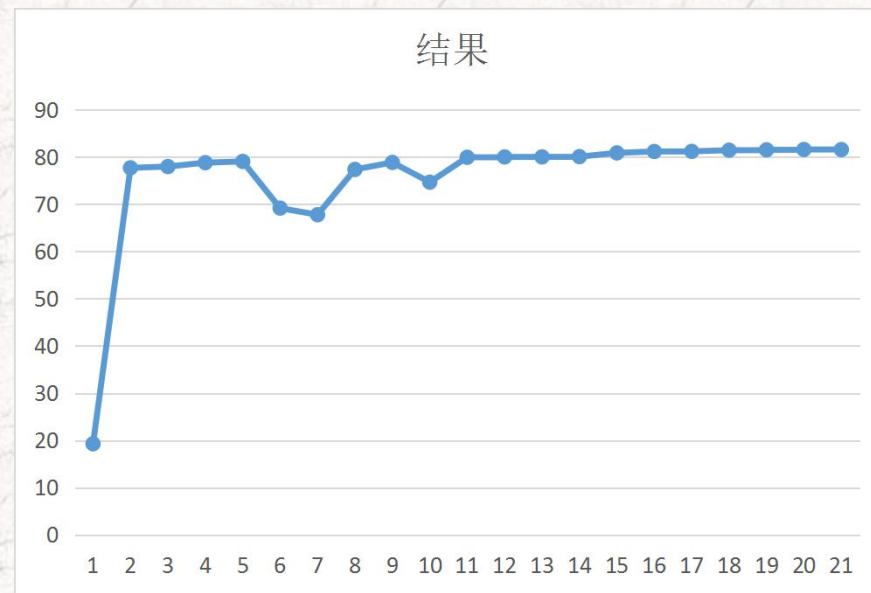
- 简单的excel函数：
`=IF(B2=C2,B2,A2)`

- xgboost+linearSVC+randomForest+bys（myself）： 81.547;



10, final

- 再发现每个国家中有些菜本来就含有国家名。
- 改善程序，有出现国家名直接判定成功。



- 结论：我们要持之以恒，不断提升自我。