**qq群标签分类器**

**一、数据**

1.train\_data.csv：训练数据4000

2.test\_data.csv：测试数据20462

**二、构建分类器的流程**

1.数据预处理

（1）词向量：基于“百度百科”中文词向量来构造。共300维

（2）词性标注：利用正态分布随机初始化。共50维

（3）连接词向量和词性标注：共350维

（4）在训练集中，利用5-cross\_validation划分训练集和验证集

|  |  |
| --- | --- |
| **Dataset** | **Number** |
| train dataset | 3201 |
| validation dataset | 799 |

2.分类器：lightgbm

|  |  |  |
| --- | --- | --- |
| **Hyper-Parameters** | **Description** | **Value** |
| max\_depth | 树的深度 | 10 |
| learning\_rate | 学习率 | 0.05 |
| n\_estimators | 评估器的数量 | 600 |
| objective | 目标函数 | binary |
| reg\_lambda | L2正则 | 0.5 |

**三、分类器效果（在验证集中评估）**

1.针对label 1(该标签需要去除的)的预测：

|  |  |  |  |
| --- | --- | --- | --- |
| **5 fold cross validation** | **Precision** | **Recall** | **F1-score** |
| 0 | 0.791667 | 0.500000 | 0.612903 |
| 1 | **0.875000** | 0.429825 | 0.576471 |
| 2 | 0.798611 | 0.504386 | 0.618280 |
| 3 | 0.823129 | 0.530702 | 0.645333 |
| 4 | 0.828947 | **0.555066** | **0.664908** |
| Mean | 0.823471 | 0.503996 | 0.623579 |

**四、预测**

1.根据构建的分类器，预测test\_data中的label不为1的数据。共预测**16896**条数据。

2.把预测值从大到小进行排列。

3.得到数据test\_pre\_label0.csv

|  |  |  |
| --- | --- | --- |
| **数量** | **描述** | **错误率** |
| 100个 | 100个中，有5个不应该删除 | 5% |