Machine Learning大作业报告

2018013398 软件82 段凡

环境依赖

- scikit-learn==0.24.0
- pandas==1.1.5
- numpy==1.19.4
- matplotlib==3.3.3
- seaborn==0.11.1
- sklearn==0.0.0

代码环境

代码在src/LogisticRegression文件夹下，在这里可以运行代码。

- 安装依赖:

```
pip install -r requirements.txt
```

- 运行特征选择（热力图）:

```
python data_process.py
```

- 运行grid_search:

```
python test.py
```

算法选择

该相亲问题是一个二分类问题，这里选用相对基础简单的逻辑回归模型进行问题的处理。

特征分析与选择

- 数据分析

  首先根据训练数据，注意到其中大部分列的特征数据均有数据缺失，因此我们首先通过筛选，将缺失率较高(>0.7)的特征筛选剔除,其缺失情况如下图：

```
-----------------------------------
          column_name  percent_missing
shar2_3       shar2_3         0.759816
intel5_3     intel5_3         0.759816
attr5_3       attr5_3         0.759816
amb7_3         amb7_3         0.759816
fun7_3         fun7_3         0.759816
intel7_3     intel7_3         0.759816
sinc7_3       sinc7_3         0.759816
attr7_3       attr7_3         0.759816
sinc5_3       sinc5_3         0.759816
fun5_3         fun5_3         0.759816
shar7_3       shar7_3         0.759816
amb5_3         amb5_3         0.759816
attr7_2       attr7_2         0.763078
fun7_2         fun7_2         0.763078
intel7_2     intel7_2         0.763078
shar7_2       shar7_2         0.764287
amb7_2         amb7_2         0.766582
sinc7_2       sinc7_2         0.766582
expnum         expnum         0.784825
numdat_3     numdat_3         0.821675
num_in_3     num_in_3         0.920140
```

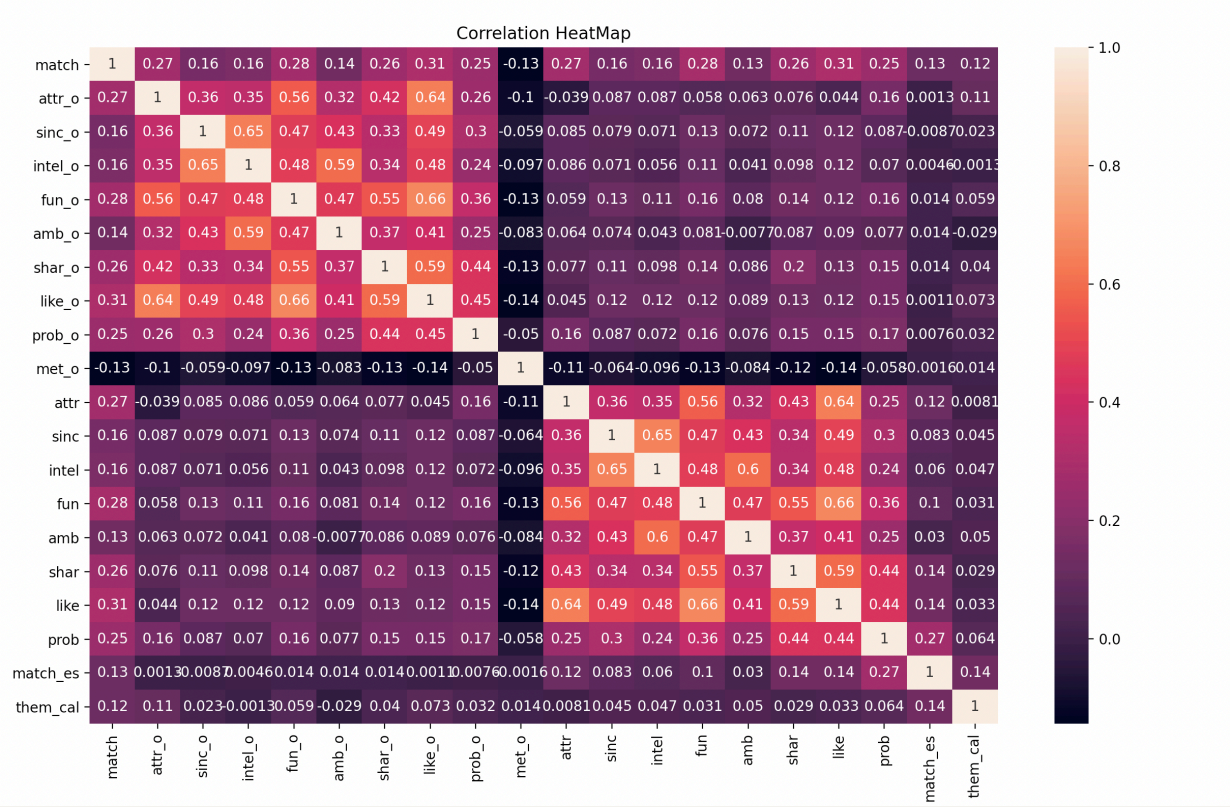
- 特征筛选

根据语义相关性，将语义无关列以及可以直接与运算得到结果的'dec'与'dec_o'特征剔除。

```
['dec_o', 'dec', 'iid', 'id', 'gender', 'idg',
            'condtn', 'wave', 'round', 'position',
'positin1',
            'order', 'partner', 'pid', 'field', 'tuition',
'career']
```

- 协相关系数计算

对精简过后的特征我们计算所有特征与结果的相关性，并对其中缺省值用众数进行填充，同时我们将相关性绝对值较大(>0.1)的结果保留在热力图中，其可视化图像如下所示：



- 语义以及相关性辅助分组

根据语义我们将特征进行分组，其中attr/attr_o/sinc/sinc_o等特征是第一次约会后双方之间的评价，语义上与结果相关度较高，因此我们将其作为基础组1，同时我们将约会后一天、约会后3-4周填写问卷的相关数据分组为组2、组3，将基础'match'函数与这些组进行组合来选择效果更优的不同分组，参数未进行调整时的运行结果如下：

	1	2	3	1+2	1+3	2+3	1+2+3
train_accuracy	85.6591	83.6293	83.6293	86.1061	86.1665	83.6293	86.2027
test_accuracy	85.1485	75.2475	75.2475	87.1287	88.1188	75.2475	88.1188

超参数调优

通过`grid_search`网格调优，最终我们得到的最佳逻辑回归模型参数如下：

```
{'model': LogisticRegression(C=0.8, max_iter=500, penalty='l1',
solver='liblinear',tol=0.01),
 'train_accuracy': 0.8615440376948169, 'test_accuracy':
0.8910891089108911,'feature_id': 5}
```

最终我们的最优测试准确率达到89.11%。

总结

逻辑回归在这个二分类问题上表现较为一般，这也与我们选择使用线性模型来进行数据拟合有一定的相关性，尽管最终正确率并不够高，排除排行榜上通过dec&dec_o直接获取结果的成绩中这个正确率只能算一般，这也与模型的限制相关。在这次解决实际问题的实验中，我对实际问题解决也有了更多的经验和认知，算法本身并不是解决这类分类的实际问题最重要的，特征选择、数据分析以及模型选择才是我们最应该关注的问题。