综合实训 机器学习之特征提炼 任务三

年级	**2018**级	专业	软件工程
学号	18342115	姓名	杨玲

一、任务要求

- 任务三：探究特征个数与预测准确率的关系
- 截止时间：12.31
- 提交内容：
  - 一份可读的代码，
  - 对于每一个数据集，输出特征个数与模型预测之间的关系
- 工作量：代码量较小，但需要一定的时间跑代码。

二、实验过程和结果

1. 实验步骤

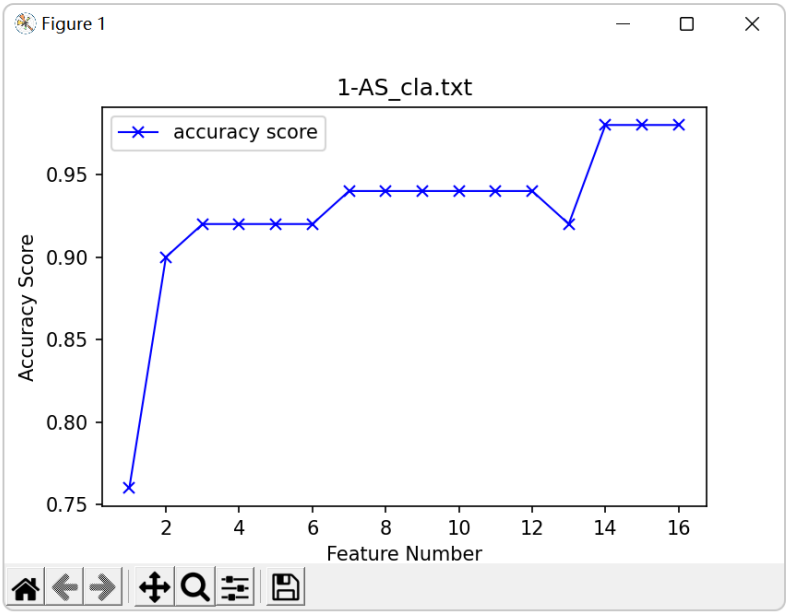
- 按照数据表中的前后顺序，分别选择1-16个特征数量来进行训练和测试
  - （按照相关度顺序的测试已在任务二中做过）
- 按照训练、测试比9:1的比例将数据集拆分为训练集和测试集，每10条数据中选取第二条数据作为测试集
- 调用sklearn库的KNeighborsClassifier来进行训练及测试
- 输出相关结果

2. 实验结果

- 1-AS_cla

最终结果：

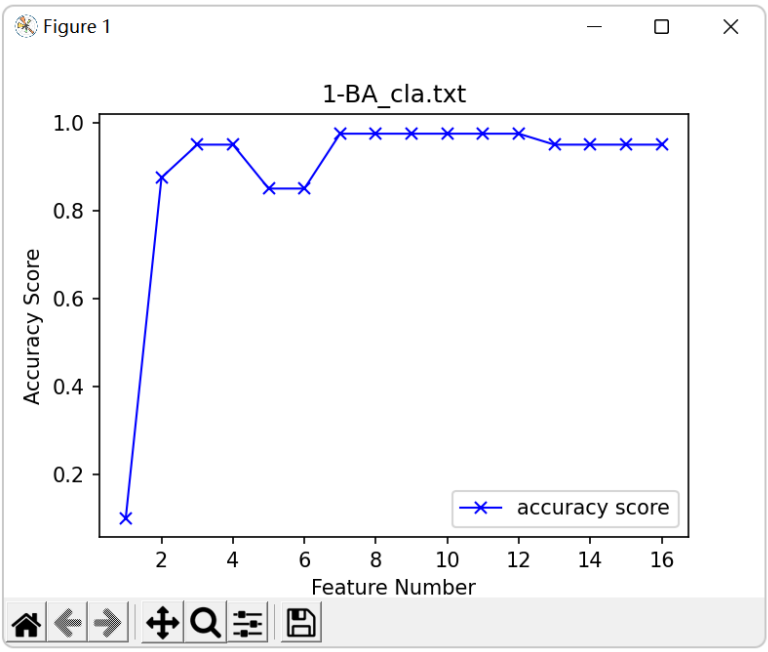
Feature Number	Accuracy Score
1	0.76
2	0.9
3	0.92
4	0.92
5	0.92
6	0.92
7	0.94
8	0.94
9	0.94
10	0.94
11	0.94
12	0.94
13	0.92
14	0.98
15	0.98
16	0.98



- 1-BA_cla

最终结果：

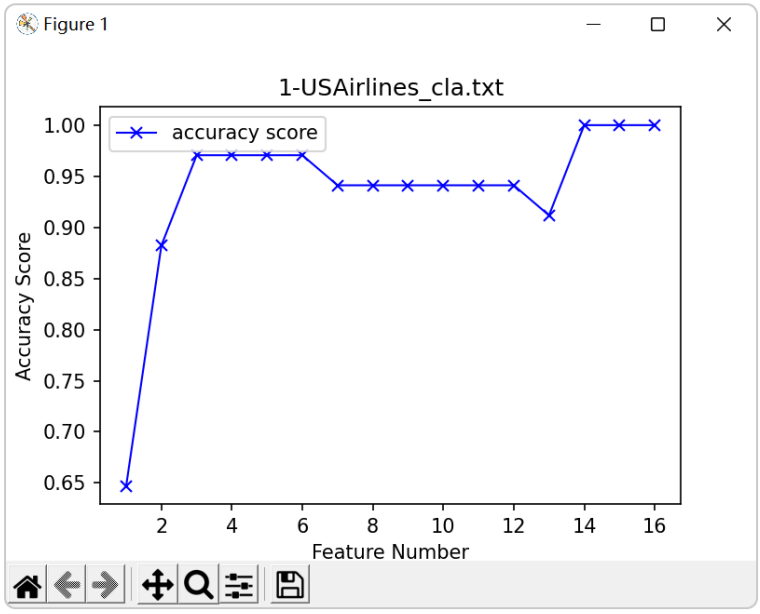
Feature Number	Accuracy Score
1	0.1
2	0.875
3	0.95
4	0.95
5	0.85
6	0.85
7	0.975
8	0.975
9	0.975
10	0.975
11	0.975
12	0.975
13	0.95
14	0.95
15	0.95
16	0.95



- 1-USAirlines_cla

最终结果：

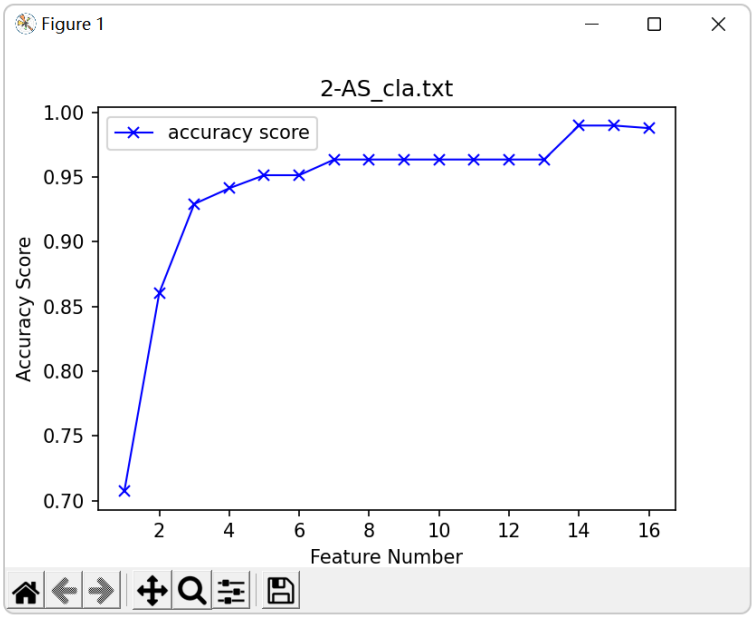
Feature Number	Accuracy Score
1	0.6470588235294118
2	0.8823529411764706
3	0.9705882352941176
4	0.9705882352941176
5	0.9705882352941176
6	0.9705882352941176
7	0.9411764705882353
8	0.9411764705882353
9	0.9411764705882353
10	0.9411764705882353
11	0.9411764705882353
12	0.9411764705882353
13	0.9117647058823529
14	1.0
15	1.0
16	1.0



- 2-AS_cla

最终结果：

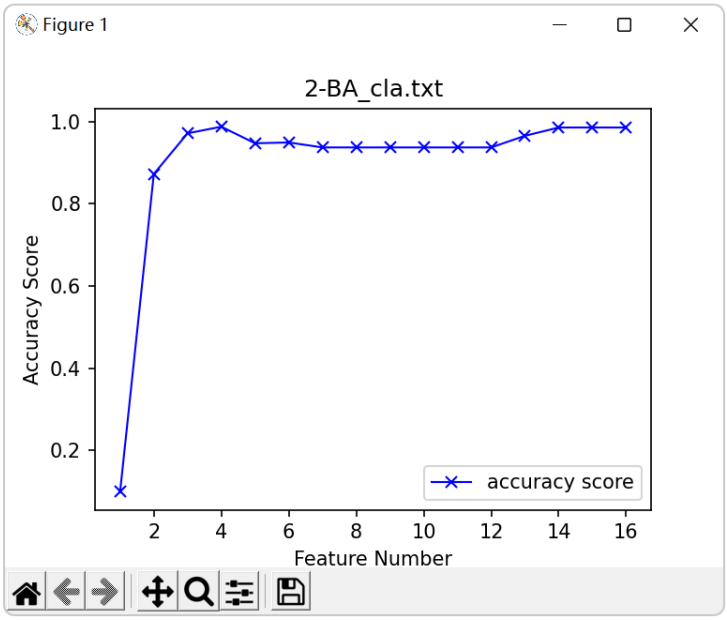
Feature Number	Accuracy Score
1	0.7070707070707071
2	0.8606060606060606
3	0.9292929292929293
4	0.9414141414141414
5	0.9515151515151515
6	0.9515151515151515
7	0.9636363636363636
8	0.9636363636363636
9	0.9636363636363636
10	0.9636363636363636
11	0.9636363636363636
12	0.9636363636363636
13	0.9636363636363636
14	0.98989898989899
15	0.98989898989899
16	0.9878787878787879



- 2-BA_cla

最终结果：

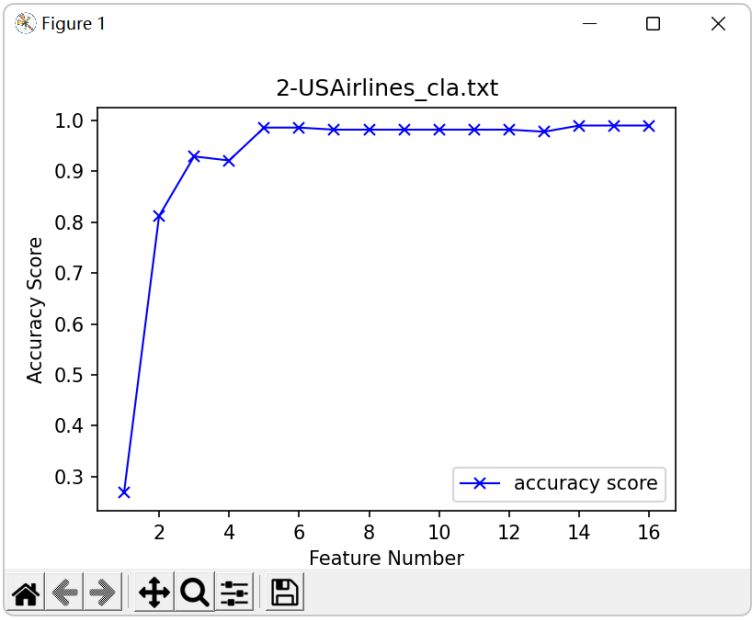
Feature Number	Accuracy Score
1	0.09898989898989899
2	0.8727272727272727
3	0.9717171717171718
4	0.9878787878787879
5	0.9474747474747475
6	0.9494949494949495
7	0.9373737373737374
8	0.9373737373737374
9	0.9373737373737374
10	0.9373737373737374
11	0.9373737373737374
12	0.9373737373737374
13	0.9656565656565657
14	0.9858585858585859
15	0.9858585858585859
16	0.9858585858585859



- 2-USAirlines_cla

最终结果：

Feature Number	Accuracy Score
1	0.2686868686868687
2	0.8121212121212121
3	0.9292929292929293
4	0.9212121212121213
5	0.9858585858585859
6	0.9858585858585859
7	0.9818181818181818
8	0.9818181818181818
9	0.9818181818181818
10	0.9818181818181818
11	0.9818181818181818
12	0.9818181818181818
13	0.9777777777777777
14	0.98989898989899
15	0.98989898989899
16	0.98989898989899



四、分析与总结

从实验结果可以看出，选择的特征数量不同，模型的准确率也会不同。

- 总体来说，随着选择的特征数量变多，模型的准确率也会上升。
- 一般来说，当选择的特征数量大于等于3时，模型的准确率都会高于90%（1-BA_cla除外）。
- 个别相关系数不高的特征向量可能反而会导致准确率下降。

五、参考文档

机器学习如何计算特征的重要性_简介机器学习中的特征工程

机器学习如何计算特征的重要性_机器学习之特征工程

sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.0.1 documentation