综合实训 机器学习之特征提炼 任务二

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

一、任务要求

- 任务二：实现一个机器学习算法（我负责做分类算法），选择合适的特征，训练模型，记录模型预测准确率。
- 截止时间：11.27
- 提交内容：
  - 一份可读的代码，
  - 对于每一个数据集，输出模型预测误差
- 工作量：代码量比较大，需要了解机器学习算法并完成实现

二、KNN算法

由于所给数据的分类为多类别，所以我最终选择了KNN算法来进行实验。

1.概述

KNN（K- Nearest Neighbor）法即K最邻近法，最初由 Cover和Hart于1968年提出，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。

KNN是通过测量不同特征值之间的距离进行分类。它的思路是：如果一个样本在特征空间中的k个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中K通常是不大于20的整数。

KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

2.算法流程

总体来说，KNN分类算法包括以下4个步骤：

- 准备数据，对数据进行预处理
- 计算测试样本点（也就是待分类点）到其他每个样本点的距离 。
- 对每个距离进行排序，然后选择出距离最小的K个点 。
- 对K个点所属的类别进行比较，根据少数服从多数的原则，将测试样本点归入在K个点中占比最高的那一类

3. 优缺点

- 优点
  - 思路简单，易于理解，易于实现，无需估计参数
- 缺点
  - 不均衡性：当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。
  - 计算量较大：对每一个待分类的样本点都要计算它到全体已知样本的距离，才能求得它的K个最近邻点
  - 输出可解释性不强

三、实验过程和结果

1. 实验步骤

- 特征选择，分别选择相关程度最高的16、12、8、4个特征来进行后续训练和测试
- 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

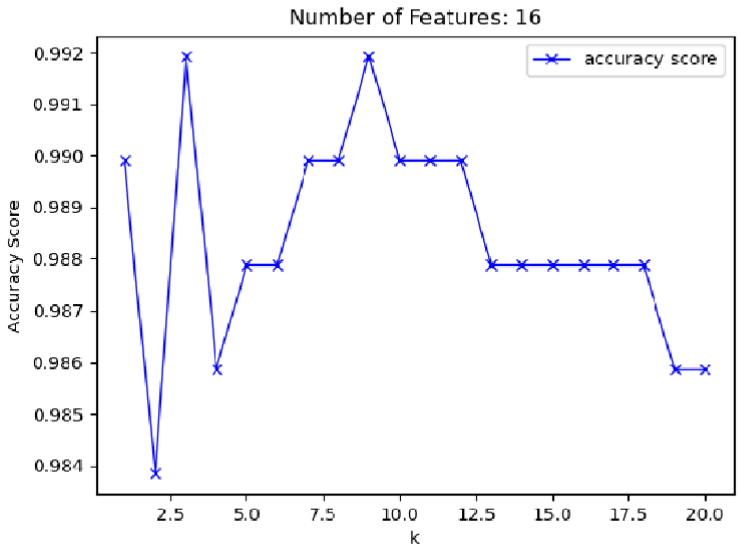
2. 实验结果

（**1**）各表具体结果（以**2-AS_cla.txt**为例）

①16个特征值

- 选择了相关系数最大的16个特征，即全部特征
- 选择特征： max_degree，fail_node_degree，fail_neber_degree，fail_degree_sum，max_load，big_load_num，fail_load_sum，fail_num，first_round_fail，neber_fail_num，fail_round，subgraph_num，fail_node_load，load_change，degree_change，fail_neber_load
- 所得的结果如下

k	Accuracy Score
1	0.98989898989899
2	0.9838383838383838
3	0.9919191919191919
4	0.9858585858585859
5	0.9878787878787879
6	0.9878787878787879
7	0.98989898989899
8	0.98989898989899
9	0.9919191919191919
10	0.98989898989899
11	0.98989898989899
12	0.98989898989899
13	0.9878787878787879
14	0.9878787878787879
15	0.9878787878787879
16	0.9878787878787879
17	0.9878787878787879
18	0.9878787878787879
19	0.9858585858585859
20	0.9858585858585859



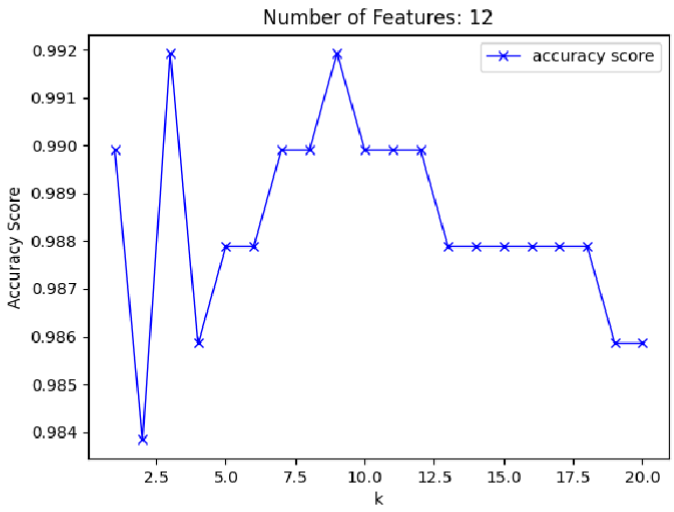
- 由图可知，当选择全部特征对模型进行训练时，准确率均达到98%以上，当k=3或9时，准确率最高。

②12个特征值

- 选择了相关系数最大的12个特征
- 选择特征：fail_neber_degree，fail_degree_sum，max_load，fail_load_sum，fail_num，first_round_fail，neber_fail_num，subgraph_num，fail_node_load，load_change，degree_change，fail_neber_load

- 所得结果如下：

k	Accuracy Score
1	0.98989898989899
2	0.9838383838383838
3	0.9919191919191919
4	0.9858585858585859
5	0.9878787878787879
6	0.9878787878787879
7	0.98989898989899
8	0.98989898989899
9	0.9919191919191919
10	0.98989898989899
11	0.98989898989899
12	0.98989898989899
13	0.9878787878787879
14	0.9878787878787879
15	0.9878787878787879
16	0.9878787878787879
17	0.9878787878787879
18	0.9878787878787879
19	0.9858585858585859
20	0.9858585858585859



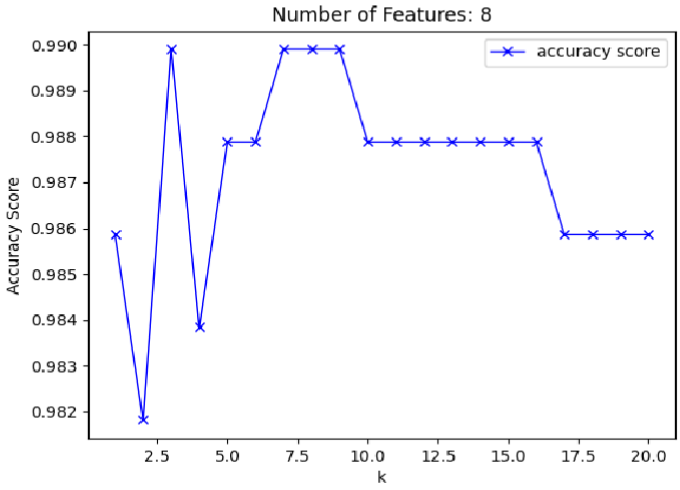
- 由图可知，当选择全部特征对模型进行训练时，准确率均达到98%以上，当k=3或9时，准确率最高。

③**8**个特征值

- 选择了相关系数最大的8个特征
- 选择特征：fail_degree_sum，fail_load_sum，fail_num，first_round_fail，

  fail_node_load，load_change，degree_change，fail_neber_load
- 所得结果如下：

k	Accuracy Score
1	0.9858585858585859
2	0.9818181818181818
3	0.98989898989899
4	0.9838383838383838
5	0.9878787878787879
6	0.9878787878787879
7	0.98989898989899
8	0.98989898989899
9	0.98989898989899
10	0.9878787878787879
11	0.9878787878787879
12	0.9878787878787879
13	0.9878787878787879
14	0.9878787878787879
15	0.9878787878787879
16	0.9878787878787879
17	0.9858585858585859
18	0.9858585858585859
19	0.9858585858585859
20	0.9858585858585859

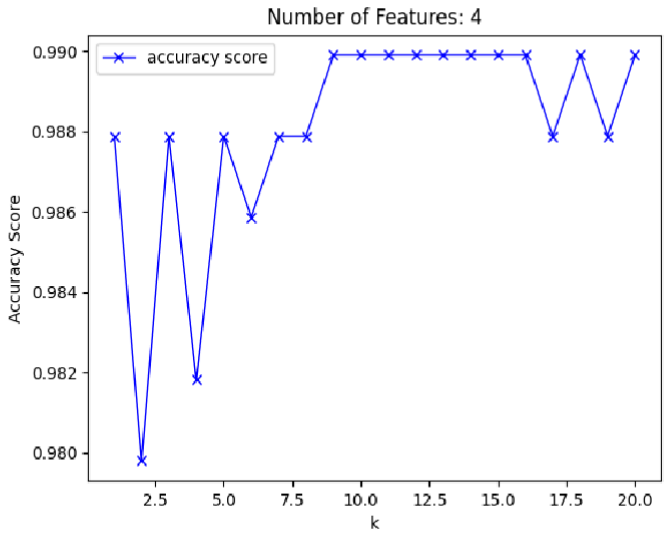


- 由图可知，当选择全部特征对模型进行训练时，准确率均达到98%以上，当k=3、7、8或9时，准确率最高，但是最高准确率反而没有选16或12个特征值时高了

④4个特征值

- 选择了相关系数最大的4个特征
- 选择特征：fail_load_sum，first_round_fail，load_change，fail_neber_load
- 所得结果如下：

k	Accuracy Score
1	0.9878787878787879
2	0.9797979797979798
3	0.9878787878787879
4	0.9818181818181818
5	0.9878787878787879
6	0.9858585858585859
7	0.9878787878787879
8	0.9878787878787879
9	0.98989898989899
10	0.98989898989899
11	0.98989898989899
12	0.98989898989899
13	0.98989898989899
14	0.98989898989899
15	0.98989898989899
16	0.98989898989899
17	0.9878787878787879
18	0.98989898989899
19	0.9878787878787879
20	0.98989898989899

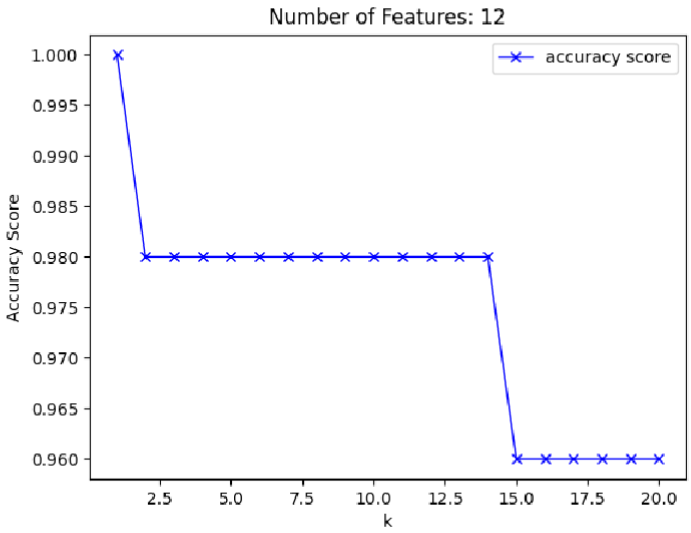


- 由图可知，当选择全部特征对模型进行训练时，准确率均达到97%以上，当k=9-16时，准确率最高，但是最高准确率同意没有选16或12个特征值时高了

（**2**）各表结果汇总（以选择相关系数最高的**12**个特征值为例）

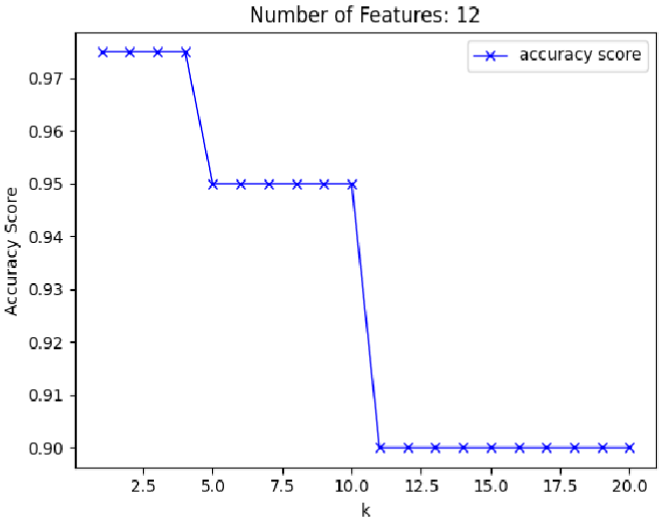
- 1-AS_cla

k	Accuracy Score
1	1.0
2	0.98
3	0.98
4	0.98
5	0.98
6	0.98
7	0.98
8	0.98
9	0.98
10	0.98
11	0.98
12	0.98
13	0.98
14	0.98
15	0.96
16	0.96
17	0.96
18	0.96
19	0.96
20	0.96



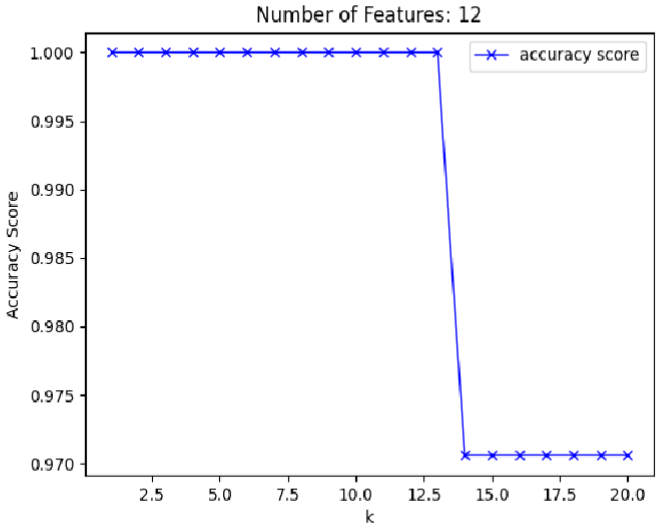
- 1-BA_cla

k	Accuracy Score
1	0.975
2	0.975
3	0.975
4	0.975
5	0.95
6	0.95
7	0.95
8	0.95
9	0.95
10	0.95
11	0.9
12	0.9
13	0.9
14	0.9
15	0.9
16	0.9
17	0.9
18	0.9
19	0.9
20	0.9



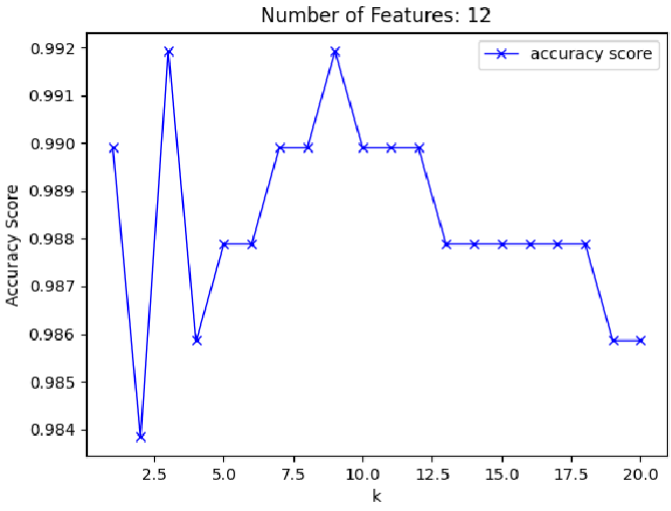
- 1-USAirlines_cla

k	Accuracy Score
1	1.0
2	1.0
3	1.0
4	1.0
5	1.0
6	1.0
7	1.0
8	1.0
9	1.0
10	1.0
11	1.0
12	1.0
13	1.0
14	0.9705882352941176
15	0.9705882352941176
16	0.9705882352941176
17	0.9705882352941176
18	0.9705882352941176
19	0.9705882352941176
20	0.9705882352941176



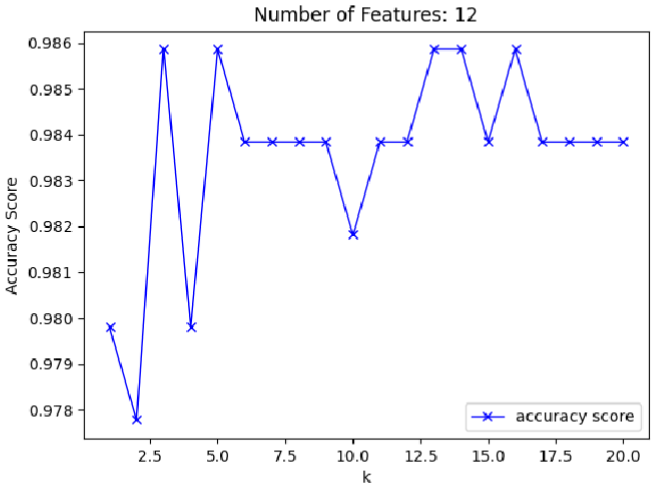
- 2-AS_cla

k	Accuracy Score
1	0.98989898989899
2	0.9838383838383838
3	0.9919191919191919
4	0.9858585858585859
5	0.9878787878787879
6	0.9878787878787879
7	0.98989898989899
8	0.98989898989899
9	0.9919191919191919
10	0.98989898989899
11	0.98989898989899
12	0.98989898989899
13	0.9878787878787879
14	0.9878787878787879
15	0.9878787878787879
16	0.9878787878787879
17	0.9878787878787879
18	0.9878787878787879
19	0.9858585858585859
20	0.9858585858585859



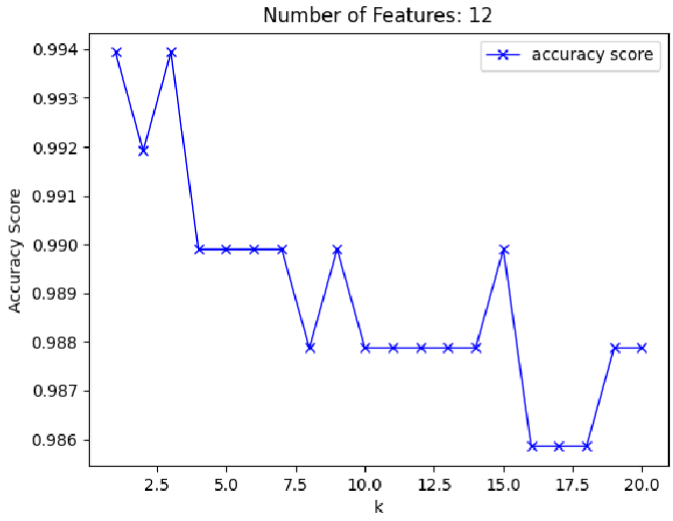
- 2-BA_cla

k	Accuracy Score
1	0.9797979797979798
2	0.9777777777777777
3	0.9858585858585859
4	0.9797979797979798
5	0.9858585858585859
6	0.9838383838383838
7	0.9838383838383838
8	0.9838383838383838
9	0.9838383838383838
10	0.9818181818181818
11	0.9838383838383838
12	0.9838383838383838
13	0.9858585858585859
14	0.9858585858585859
15	0.9838383838383838
16	0.9858585858585859
17	0.9838383838383838
18	0.9838383838383838
19	0.9838383838383838
20	0.9838383838383838



- 2-USAirlines_cla

k	Accuracy Score
1	0.9939393939393939
2	0.9919191919191919
3	0.9939393939393939
4	0.98989898989899
5	0.98989898989899
6	0.98989898989899
7	0.98989898989899
8	0.9878787878787879
9	0.98989898989899
10	0.9878787878787879
11	0.9878787878787879
12	0.9878787878787879
13	0.9878787878787879
14	0.9878787878787879
15	0.98989898989899
16	0.9858585858585859
17	0.9858585858585859
18	0.9858585858585859
19	0.9878787878787879
20	0.9878787878787879



四、分析与总结

从实验结果（1）可以看出，选择的特征数量不同，模型的准确率也会不同。在2-AS_cla.txt表的例子中，显然当特征数量为12或16时，模型的总体准确率更高。

从实验结果（2）可以看出，对于KNN模型，选择的k不同，模型的准确率也会不同，对于大部分表来说，k=3时，模型的准确率是最高的。

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