数据科学基础第二次作业

王晨曦

2018 年 4 月 20 日

1 问题重述

探索 iris（鸢尾花）数据集（图1）。

setosa　　versicolor　　virginica

图 1: iris 数据集类别

- 探索 iris 的以下属性:
  - DESCR：数据集总体描述
  - data：所有观测实例的特征数据（numpy.ndarray 类型）
  - data.shape：维信息
  - feature_names：特征名（sepal 萼片,petal 花瓣）
  - target：所有实例的类别值（numpy.ndarray 类型），0/1/2
  - target.shape：维信息
  - target_names：类别名称（setosa/versicolor/virginica）
- 探索各特征的最小值，最大值，均值，中位数，标准差。
- 探索各特征之间，以及特征与目标之间的相关性（相关系数）。

2 问题解答

解答分为三部分:数据集合属性、特征统计参数计算和相关系数计算。完整代码用 jupyter notebook 格式的文件附在文件夹中，文件名为 hw2.ipynb。

2.1 探索 iris 数据集的属性

对 iris 数据集的属性进行分析，可以得到关于数据和类别的基本情况：

- iris 是关于鸢尾花的植物参数和对应类别的数据集，共有 150 个样本由 R. A. Fisher 发布于 1988 年 7 月，是模式识别领域的著名数据集。
- iris 数据集包含了萼片长宽、花瓣长宽四种属性，单位均为厘米，数据类型为 numpy.ndarray。
- iris 数据集包含了 setosa、versicolor 和 virginica 三种类别，每种类别分别有 50 个实例，标签分别用 0、1 和 2 表示，标签数据类型为 numpy.ndarray。

详细代码如下，为方便起见，代码中省去了导入库和数据集的相关操作。

```
In [1]:  # 数据集总体描述
         iris.DESCR.split('\n')
Out[1]: ['Iris Plants Database',
         '====================',
         '',
         'Notes',
         '_____',
         'Data Set Characteristics:',
         '    :Number of Instances: 150 (50 in each of three classes)',
         '    :Number of Attributes: 4 numeric, predictive attributes and the
                                              class',
         '    :Attribute Information:',
         '        - sepal length in cm',
         '        - sepal width in cm',
         '        - petal length in cm',
         '        - petal width in cm',
         '        - class:',
         '                - Iris-Setosa',
         '                - Iris-Versicolour',
         '                - Iris-Virginica',
         '    :Summary Statistics:',
         '',
         '    ============== ==== ==== ======= ===== ====================',
         '                    Min  Max   Mean    SD   Class Correlation',
         '    ============== ==== ==== ======= ===== ====================',
         '    sepal length:   4.3  7.9   5.84   0.83    0.7826',
         '    sepal width:    2.0  4.4   3.05   0.43   -0.4194',
         '    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)',
         '    petal width:    0.1  2.5   1.20  0.76     0.9565  (high!)',
         '    ============== ==== ==== ======= ===== ====================',
         '',
         '    :Missing Attribute Values: None',
         '    :Class Distribution: 33.3% for each of 3 classes.',
         '    :Creator: R.A. Fisher',
         '    :Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)',
         '    :Date: July, 1988',
         '',
         'This is a copy of UCI ML iris datasets.',
         'http://archive.ics.uci.edu/ml/datasets/Iris',
         '',
```

```
        'The famous Iris database, first used by Sir R.A Fisher',
        '',
        'This is perhaps the best known database to be found in the',
        "pattern recognition literature.  Fisher's paper is a classic in the
                                                field and",
        'is referenced frequently to this day.  (See Duda & Hart, for example.)
                                                 The',
        'data set contains 3 classes of 50 instances each, where each class
                                                refers to a',
        'type of iris plant.  One class is linearly separable from the other 2;
                                                 the',
        'latter are NOT linearly separable from each other.',
        '',
        'References',
        '----------',
        '   - Fisher,R.A. "The use of multiple measurements in taxonomic
                                                problems"',
        '     Annual Eugenics, 7, Part II, 179-188 (1936); also in "
                                                Contributions to',
        '     Mathematical Statistics" (John Wiley, NY, 1950).',
        '   - Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene
                                                Analysis.',
        '     (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
                                                ',
        '   - Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New
                                                System',
        '     Structure and Classification Rule for Recognition in Partially
                                                Exposed',
        '     Environments".  IEEE Transactions on Pattern Analysis and Machine
                                                ',
        '     Intelligence, Vol. PAMI-2, No. 1, 67-71.',
        '   - Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule".  IEEE
                                                Transactions',
        '     on Information Theory, May 1972, 431-433.',
        '   - See also: 1988 MLC Proceedings, 54-64.  Cheeseman et al"s
                                                AUTOCLASS II',
        '     conceptual clustering system finds 3 classes in the data.',
        '   - Many, many more ...',
        '']


In [2]: # 数据类型
        type(iris.data)
Out[2]: numpy.ndarray


In [3]: # 数据维度信息
        iris.data.shape
Out[3]: (150L, 4L)


In [4]: # 特征名称
        iris.feature_names
```

```
Out[4]: ['sepal length (cm)',
         'sepal width (cm)',
         'petal length (cm)',
         'petal width (cm)']


In [5]: # 所有实例的类别值
        iris.target
Out[5]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
               0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
               0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
               1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
               1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2])


In [6]: # 类别维度信息
        iris.target.shape
Out[6]: (150L,)


In [7]: # 类别名称
        iris.target_names
Out[7]: array(['setosa', 'versicolor', 'virginica'],
              dtype='|S10')
```

2.2 探索各特征的统计参数

对四种特征分别计算最值、均值、中位数、标准差和四分位数（距），结果见表1。为了更好地观察特征值的总体情况，可以用箱线图来刻画特征的分布状况，结果见图2，对于每个特征，自下而上的五条横线分别表示特征值的最小值（不含异常值）、下四分位数、中位数（红线）、上四分位数和最大值（不含异常值），圆圈代表数据中的异常值。

表 1: iris 数据集各特征的统计参数

特征名称	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
最小值	4.30	2.00	1.00	1.00
最大值	7.90	4.40	6.90	2.50
均值	5.84	3.05	3.76	1.20
中位数	5.80	3.00	4.35	1.30
标准差	0.82	0.43	1.76	0.76
下四分位数	5.10	2.80	1.60	0.30
上四分位数	6.40	3.30	5.10	1.80
四分位距	1.30	0.50	3.50	1.50

从图2可以看出，除了萼片宽度（sepal width）以外，其余三个属性均无异常值出现；只有萼片长度（sepal length）呈对称分布，其余三个属性为偏斜分布。

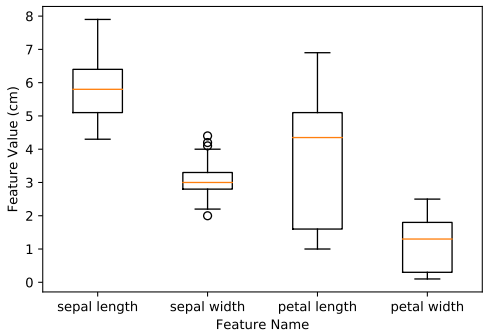

图 2: 特征值分布箱线图

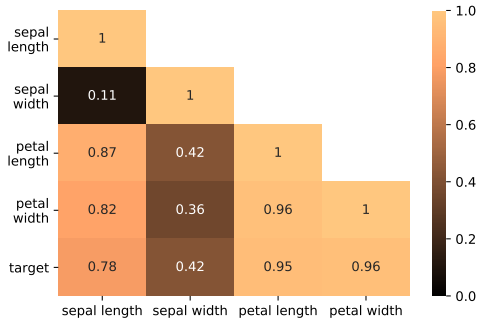

图 3: 相关系数热度图

2.3 探索特征与特征、特征与目标间的相关性

如表2所示，前四行是四个特征之间的相关系数（重复值已省略），第五行是四个特征与类别（数值化为0,1,2）之间的相关系数。可以看出，萼片长度与花瓣长宽、花瓣宽度与花瓣长度存在显著正相关性；萼片长度、花瓣长宽与类别存在显著正相关性，萼片宽度与类别存在负相关，但不明显。

表 2: iris 数据集特征与特征、特征与目标间的相关系数

特征名称	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
sepal length (cm)	1.000	-0.109	0.872	0.818
sepal width (cm)	–	1.000	-0.421	-0.357
petal length (cm)	–	–	1.000	0.963
petal width (cm)	–	–	–	1.000
target	0.783	-0.419	0.949	0.956

图3展示了相关系数绝对值的可视化结果，图中颜色越浅代表相关性越强，颜色越深代表相关性越弱。可以发现，萼片宽度与其他特征以及目标值的相关性都非常低，结合2.2中萼片宽度有不少异常值出现，可以推断鸢尾花的萼片宽度对花的性状影响很小。