数学学院案例分析报告

案例名称：鸢尾花分类

报告负责人：魏丹怡（202032164）

完成时间：2021.03.22

目　录

一、 数据来源和基本情况

数据来源：

https://raw.githubusercontent.com/ageron/handson-ml/master/datasets/housing/housing.tgz

原始数据的基本情况：

数据集是关于不同种类鸢尾花的数据，包含 3 个类别共 150 个样本，每类各 50 个样本。每个样本有 4 项特征：花萼长度（sepal length）、花萼宽度（sepal width）、花瓣长度（petal length）及花瓣宽度（petal width）；还有一个分类变量：类别（Label）。

鸢尾花的类别为：山鸢尾（iris setosa）、变色鸢尾（iris versicolour）及维吉尼亚鸢尾（iris virginica），分别由数字 0、1、2 表示，其中 0 表示山鸢尾，1 表示变色鸢尾，2 表示维吉尼亚鸢尾。

二、 特征分析和可视化

1、数据的统计特征分析

对于鸢尾花的四个特征进行描述性统计，如下图所示：

表 1 鸢尾花特征描述性统计

	sepallength	sepalwidth	petallength	petalwidth
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

从表中可以看出花萼长度、花萼宽度、花瓣长度和花瓣宽度的均值各不相同，且对于均方误差来说，其中花瓣宽度方差最大为 1.76，花萼宽度是方差最小的为 1.43。也就是说这几个特征中花瓣宽度的差异最大，在之后选择分析特征时可以给予着重考虑。

2、特征可视化

将数据描述表示从表格转化为图像形式如下小提琴图所示。小提琴图结合了箱形图和密度图的特征，主要用来显示数据的分布形状。中间白点为中位数，中间的黑色粗条表示四分位数范围。上下贯穿小提琴图的黑线代表最小非异常值 min 到最大非异常值 max 的区间，线上下端分别代表上限和下限，超出此范围为异常数据，其中颜色的部分为密度图宽。

其中纵坐标代表鸢尾花特征，横坐标代表鸢尾花的三种不同种类，其中四幅图分别代表四个不同的鸢尾花特征。左上角的图为花萼长度，右上角为花萼宽度，左下角为花瓣长度，右下角为花瓣宽度。可以看到三个类别的花萼长度和花萼宽度近似相等，但对于第三种花的花萼长度和第一种花的花萼宽度自身有着较大的差异。对于花瓣长度和花瓣宽度而言三种不同种类的鸢尾花有着相对较大的差距，其中第一种花即山鸢尾（iris setosa）有着最短的花瓣长度和宽度，而第三种花维吉尼亚鸢尾（iris virginica）有着最长的花瓣长度和宽度。由此来看第三第四个特征更适合用来进行不同鸢尾花种类的区分。

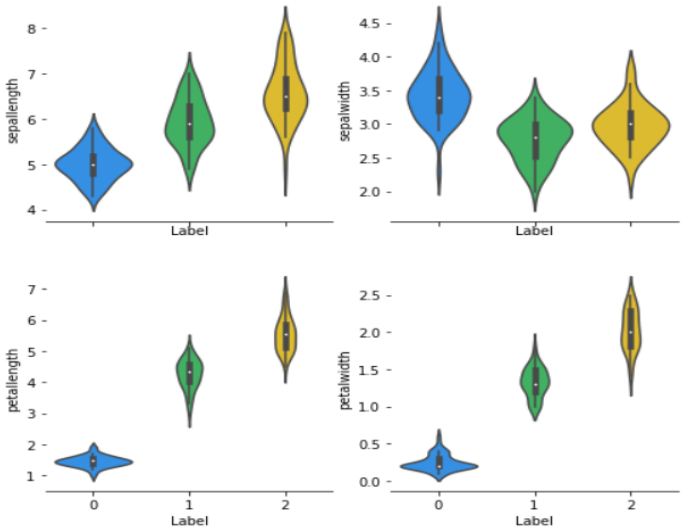

图 1 鸢尾花特征小提琴图

3、特征和特征之间的关系

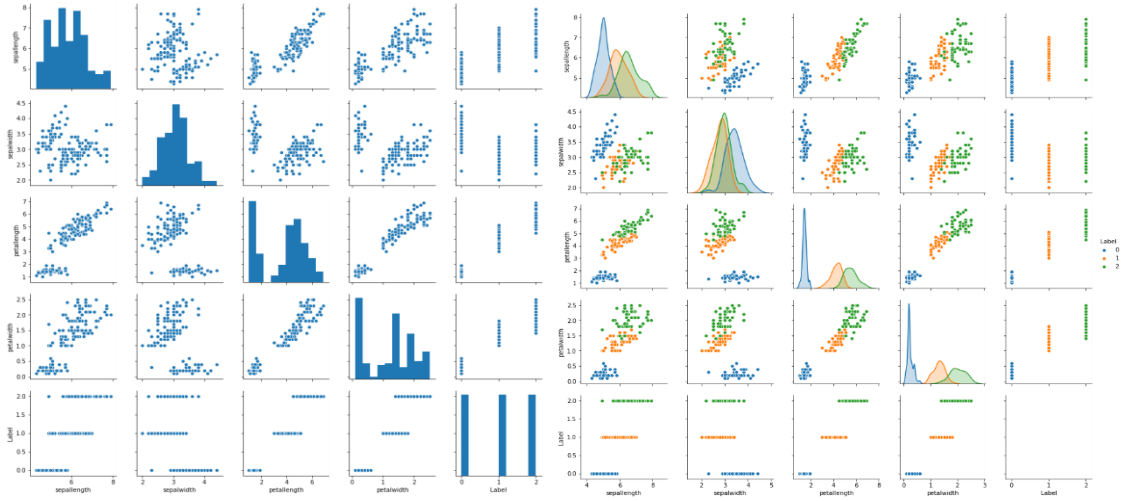

图 2 鸢尾花特征和特征之间的关系

我采取绘制散点图的方法来表示特征和特征之间的关系，将一个特征作为 x 轴，另一个特征作为 y 轴，将每一个数据点绘制为图上的一个点。由于每副散点图只能展式 2 个（或 3 个）特征之间的关系，难以对多于 3 个特征的数据集作图，为了解决这个问题，绘制了散点图矩阵，从而可以两两查看所有的特征。由上图可以看到对角线上是各个属性的直方图（分布图），而非对角线上是两个不同属性之间的相关图

如上图所示，左侧的图为未进行鸢尾花种类区分的散点图。从图中发现，花瓣的长度和宽度之间以及萼片的长短和花瓣的长、宽之间具有比较明显的相关关系，但花萼宽度和长度之间的相关性非常低。右图为经过 hue 分类后的 pairplot 中发现，不论是从对角线上的分布图还是从分类后的散点图，都可以看出对于不同种类的花，其萼片长、花瓣长、花瓣宽的分布差异较大。换句话说，这些属性是可以帮助我去识别不同种类的花的。比如，对于萼片、花瓣长度较短，花瓣宽度较窄的花，那么它大概率是山鸢尾。此外，山鸢尾（第 0 类）总是明显区分于其他两类鸢尾花。

4、结论

通过对于鸢尾花各个特征的描述性统计、绘制小提琴图以及对于鸢尾花各个特征之间的散点图。可以看出鸢尾花的种类是可区分的，且其瓣长与花瓣宽以及萼片长宽与花瓣长宽之间对于三种鸢尾花有着很好的区分作用，可以在后续特征选择和提取以及鸢尾花数据分类中起到提示与佐证作用。

三、特征工程（特征选择和提取）

1、卡方检验

表 2　卡方检验

得分	10.8178	3.7107	116.3126	67.0484
P 值	4.47651499e-03	1.56395980e-01	5.53397228e-26	2.75824965e-15

根据特征的得分可以看出，后两个特征是得分最高，因此选择后两个特征；根据 P 值可以得到一致的结果，后两个特征的 P 值最小，更能拒绝原假设，置信度更高。

2、交叉验证

从 4 个特征中随机选取 2 个特征进行组合，共 6 种组合方式，使用逻辑回归模型，进行十折交叉验证得到最大分类准确率并绘制散点图如下，其中横坐标代表选取的特征，纵坐标代表在选取该组特征时可达到的最高准确率：

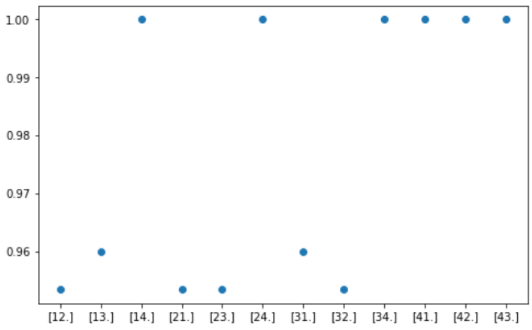

图 3 鸢尾花特征选取图

从上图可以看出，在选取第三、四个特征（花瓣长度和宽度）时可达到的准确率最高，与卡方检验得到的结果一致，故选取后两个特征（花瓣长度和宽度）进行接下来的分析。

四、 鸢尾花数据的分类

1、形式化和分类器数学推导

我们拟用逻辑回归及 SVM（支持向量机）这两种机器学习算法对于鸢尾花数据进行分类。

（1）Logistic Regression

逻辑回归是监督学习中的一种，它根据大量带有分类标签的特征变量来训练优化模型，在根据模型来预测只有特征变量的分类标签。在鸢尾花案例中，我们通过许多带有分类标签（鸢尾花的三种类别）的特征变量数据来训练预测鸢尾花类别的模型，这些特征变量有：花瓣的长度、花瓣的宽度、花萼的长度、花萼的宽度。为实现预测分类问题的目的我们利用了 Logistic 函数(或者成为 Sigmoid 函数)：

$$log_i(Z) = \frac{1}{1+e^{-z}}$$

整个函数的图象如下：

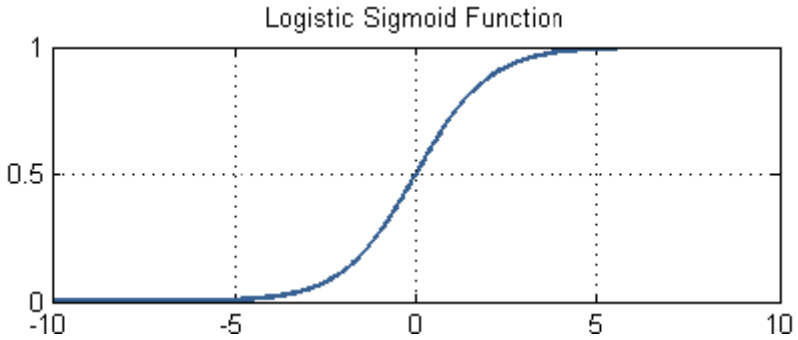

图 4 Sigmoid 函数图像

Sigmoid 函数有一些特点，比如当$z = 0$是$log_i(z) = 0.5$当$z < 0$时，$0 < log_i(z) < 0.5$；当$z > 0$时，$0.5 < log_i(z) < 1$。所以$log_i(z)$函数的取值范围为$(0,1)$。

其中回归的基本方程为:

$$z = w_0 + \sum_i^N w_i x_i$$

可以把$log_i(z)$的函数值看成类别为 1 的概率预测值，当$log_i(z) < 0.5$时，预测的分类为 0；当$log_i(z) >= 0.5$时，预测的分类为 1。对于鸢尾花数据这种三分类问题，只需将两个二分类的逻辑回归组合即可实现。即通过一个逻辑回归先区分是否为山鸢尾，再通过一个逻辑回归区分是否是变色鸢尾或是维吉尼亚鸢尾。

(2) SVM

运用支持向量机实现一个多分类问题，我使用一对多的方法，即训练时依次把某个类别的样本归为一类,其他剩余的样本归为另一类，这样 k 个类别的样本就构造出了 k 个 SVM。在此案例中，要将鸢尾花分为三类（记为 0、1、2），于是抽取训练集的时候，分别抽取：

（a）0 所对应的向量作为正集，1、2 所对应的向量作为负集；

（b）1 所对应的向量作为正集，0，3 所对应的向量作为负集；

（c）3 所对应的向量作为正集，1，2 所对应的向量作为负集；

使用这三个训练集分别进行训练，得到三个训练结果。在测试的时候，把对应的测试向量分别利用这三个训练结果文件进行测试。最后每个测试都有一个结果，最终的结果便是这三个值中最大的一个作为分类结果。

将鸢尾花的属性以坐标形式表示，建立以下支持向量机模型：

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$subject\ to\ y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0. i = 1, \cdots, n$$

其中参数 C 代表在线性不可分的情况下，对分类错误的惩罚程度。ζ_i是对于第 i 样本点的分类损失，如果分类正确则是 0，如果分类有所偏差则对应一个线性的值，ζ_i的求和是总误差，优化的目标当这个值越小越好，越小代表对训练集的分类越精准。目标函数中另一项的最小化的优化方向则是使间隔大小最大。

2、计算机仿真过程

(1) Logistic Regression

在 jupyter notebook 中调用 sklearn.linear_model 包构建基于逻辑回归的鸢尾花分类模型。

(2) SVM

在 jupyter notebook 中调用 svc 构建基于支持向量机的鸢尾花分类模型。

3、评估方法和数据划分

此案例为监督学习，可采用混淆矩阵进行评估，通过特征选择，分类依据包含 2 个特征变量，通过这些特征将鸢尾花归类。首先，将数据导入，并观察其各类的分布情况：

```
2    50
1    50
0    50
Name: Label, dtype: int64
```

可见，150 个样本中，三个种类的鸢尾花比例为 1：1：1，均为 50 个样本。

下面需要将数据分为训练集和测试集并且使得训练集和测试集中三种花的比例尽量满足 1：1：1，其中 70%的数据划分为训练集，30%的数据划分为测试集：

```
2    37
1    35
0    33
Name: Label, dtype: int64
0    17
1    15
2    13
Name: Label, dtype: int64
```

4、分类结果和比较统计

（1）Logistic Regression

通过 Sklearn 的 LogisticRegression 模型，取 C=1000，solver='lbfgs'，对鸢尾花数据集进行分类，结果如下：

	precision	recall	f1-score	support
0	1.00	1.00	1.00	14
1	0.92	1.00	0.96	12
2	1.00	0.95	0.97	19
accuracy			0.98	45
macro avg	0.97	0.98	0.98	45
weighted avg	0.98	0.98	0.98	45

可以看出，模型分类的精度达到了 98%。其次，通过混淆矩阵来观察预测分类和实际分类情况，绘制此混淆矩阵的热点图如下：

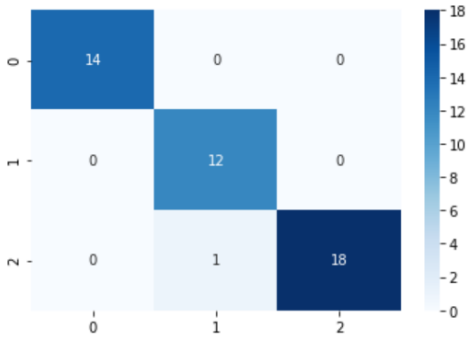

图 5　混淆矩阵热力图（a）

由上图可知，山鸢尾（第 0 类）和变色鸢尾（第 1 类）的分类完全正确，而本应该归类为维吉尼亚鸢尾（第 2 类）的 1 个样本被误分为变色鸢尾（第 1 类）。

(2) SVM

通过 Sklearn 的 SVC 模型对鸢尾花数据集进行分类，分类结果如下：

```
混淆矩阵列 [[16  0  0]
 [ 0 17  1]
 [ 0  0 11]]
准确率 0.9777777777777777
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        16
           1       1.00      0.94      0.97        18
           2       0.92      1.00      0.96        11

    accuracy                           0.98        45
   macro avg       0.97      0.98      0.98        45
weighted avg       0.98      0.98      0.98        45
```

可以看出，模型分类的精度同样达到了 98%，通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

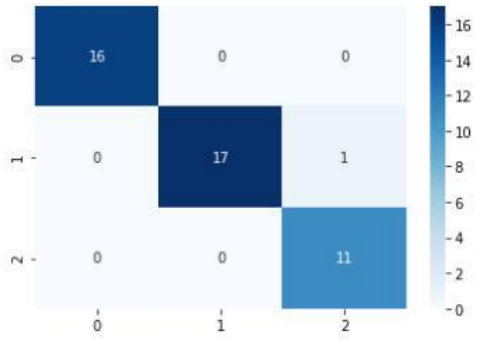

图 6　混淆矩阵热力图（b）

由上图可知，山鸢尾（第 0 类）和维吉尼亚鸢尾（第 2 类）的分类完全正确，而本应该归类为变色鸢尾（第 1 类）的 1 个样本被误分为维吉尼亚鸢尾（第 2 类）。

(3) 比较统计

基于 10 折交叉验证法，绘制两种方法的学习曲线，对比鸢尾花数据集分类准确率，结果如下：

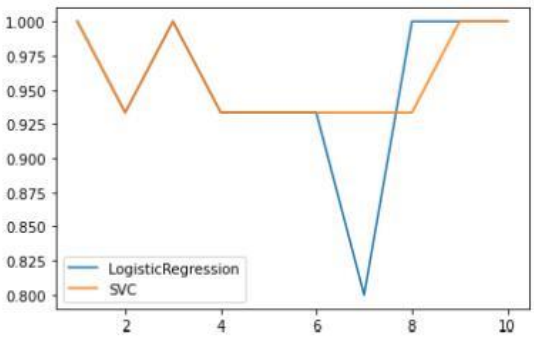

图 7 学习曲线

由上图可以看出，两种分类模型的准确率总体差距不大，其分类准确率均较高，但相较于 SVM 模型，逻辑回归模型的分类准确率偶尔会出现较低的情况。

5、结果可视化

（1）Logistic Regression

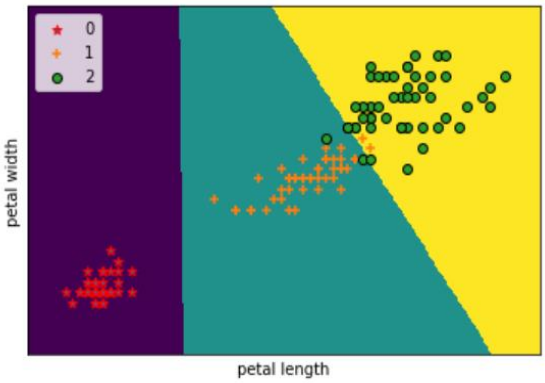

图 8　逻辑回归分类结果

（2）SVM

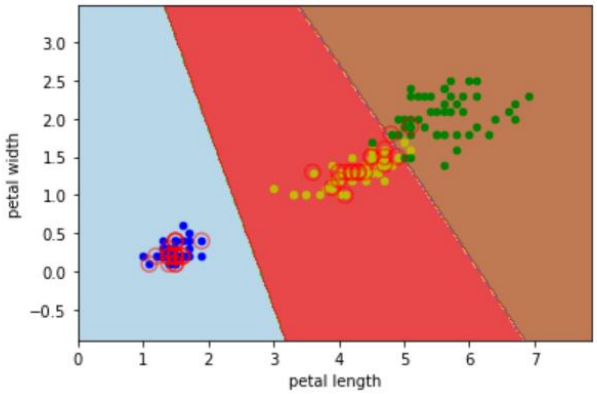

图 9　SVM 分类结果

五、结果分析和得出结论

本文以鸢尾花数据集为实验数据，采用逻辑回归和 SVM 模型对数据集进行分类，由于数据集特征较少，且无缺失值等，较好分类，两种分类模型的准确率总体差距不大，分类结果平均准确率达到 98%。

此外，由结果可视化得知，山鸢尾（第 0 类）总是明显区分于其他两类鸢尾花，在分类结果中的准确率每次均可达到 100%，而其余两类鸢尾花的特征区分不够明显，常出现误分的状况，因此，在之后的研究中，若想提高分类准确率，可以着重考虑如何更好地区分变色鸢尾（第 1 类）与维吉尼亚鸢尾（第 2 类）。

六、实践总结

通过对于鸢尾花数据分类问题的研究，学会了利用绘图和描述性统计预判问题的可行性以及对于数据建立初步的认识，其次利用卡方检验与交叉验证的特征选择和提取方法，剔除对于分类影响不大的特征。进而利用逻辑回归和 SVM 机器学习算法对鸢尾花进行分类。在分类的过程中，我加深了对逻辑回归和 SVM 理论知识的理解，同时在不断的尝试与纠错中，增加了许多编程经验。

此外，不足在于划分数据集时，直接调用 sklearn 中的模型选择划分数据，没有使用交叉验证划分数据，选择了比较简单的划分数据方法。在之后的学习中，可以更加注重对于细节的处理。