数学学院案例分析报告

案例名称：乳腺癌数据集分类

报告负责人：魏丹怡（202032164）

完成时间：2021.03.30

目　录

一、 数据来源和基本情况

1、数据来源

我使用的数据来自美国威斯康星州的乳腺癌诊断数据集，由威廉·H·沃尔伯格博士创建。此数据集包含 569 例细胞活检案例，医疗人员采集了患者乳腺肿块经过细针穿刺后的数字化图像，并且对这些数字图像进行了特征提取，这些特征可以描述图像中的细胞核呈现，同时肿瘤可以分为良性、恶性两类，所有的数据都是匿名的。

2、基本情况

此数据集共有 569 个样本，每个样本有 30 个乳房肿块活检图像显示的细胞核的特征，肿瘤的分类包括良性、恶行两类，我们分别用“0”和“1”来表示良性和恶性。其中良性肿瘤有 212 例，恶性肿瘤有 357 例。

30 个特征包括细胞核的半径(Radius)、质地(Texture)、周长(Perimeter)、面积(Area)和光滑度(Smoothness)等的均值、标准差和最大值，具体展式如下：

表 1 案例特征

特征	解释
mean radius	半径（点中心到边缘的距离）平均值
mean texture	纹理（灰度值的标准差）平均值
mean perimeter	周长 平均值
mean area	面积 平均值
mean smoothness	平滑程度（半径内的局部变化）平均值
mean compactness	紧密度 平均值
mean concavity	凹度（轮廓凹部的严重程度）平均值
mean concave points	凹缝（轮廓的凹部分）平均值
mean symmetry	对称性 平均值
mean fractal dimension	分形维数 平均值
radius error	半径（点中心到边缘的距离）标准差
texture error	文理（灰度值的标准差）标准差
perimeter error	周长 标准差
area error	面积 标准差
smoothness error	平滑程度（半径内的局部变化）标准差
compactness error	紧密度 标准差
concavity error	凹度（轮廓凹部的严重程度）标准差
concave points error	凹缝（轮廓的凹部分）标准差
symmetry error	对称性标准差
fractal dimension error	分形维数 标准差
worst radius	半径（点中心到边缘的距离）最大值
worst texture	文理（灰度值的标准差）最大值
worst perimeter	周长 最大值
worst area	面积 最大值
worst smoothness	平滑程度（半径内的局部变化）最大值
worst compactness	紧密度 最大值

worst concavity	凹度（轮廓凹部的严重程度）最大值
worst concave points	凹缝（轮廓的凹部分）最大值
worst symmetry	对称性 最大值
worst fractal dimension	分形维数 最大值

二、 特征分析

1、数据的统计特征分析

对样本的 30 个特征进行描述性统计，如下表所示：

表 1 特征描述性统计

特征	mean	std	min	0.25	0.50	0.75	max
mean radius	14.13	3.52	6.98	11.70	13.37	15.78	28.11
mean texture	19.29	4.30	9.71	16.17	18.84	21.80	39.28
mean perimeter	91.97	24.30	43.79	75.17	86.24	104.10	188.50
mean area	654.89	351.91	143.50	420.30	551.10	782.70	2501.0
mean smoothness	0.10	0.01	0.05	0.09	0.10	0.11	0.16
mean compactness	0.10	0.05	0.02	0.06	0.09	0.13	0.35
mean concavity	0.09	0.08	0.00	0.03	0.06	0.13	0.43
mean concave points	0.05	0.04	0.00	0.02	0.03	0.07	0.20
mean symmetry	0.18	0.03	0.11	0.16	0.18	0.20	0.30
mean fractal dimension	0.06	0.01	0.05	0.06	0.06	0.07	0.10
radius error	0.41	0.28	0.11	0.23	0.32	0.48	2.87
texture error	1.22	0.55	0.36	0.83	1.11	1.47	4.89
perimeter error	2.87	2.02	0.76	1.61	2.29	3.36	21.98
area error	40.34	45.49	6.80	17.85	24.53	45.19	542.20
smoothness error	0.01	0.00	0.00	0.01	0.01	0.01	0.03
compactness error	0.03	0.02	0.00	0.01	0.02	0.03	0.14
concavity error	0.03	0.03	0.00	0.02	0.03	0.04	0.40
concave points error	0.01	0.01	0.00	0.01	0.01	0.01	0.05
symmetry error	0.02	0.01	0.01	0.02	0.02	0.02	0.08
fractal dimension error	0.00	0.00	0.00	0.00	0.00	0.00	0.03
worst radius	16.27	4.83	7.93	13.01	14.97	18.79	36.04
worst texture	25.68	6.15	12.02	21.08	25.41	29.72	49.54
worst perimeter	107.26	33.60	50.41	84.11	97.66	125.40	251.20
worst area	880.58	569.36	185.20	515.30	686.50	1084.0	4254.0
worst smoothness	0.13	0.02	0.07	0.12	0.13	0.15	0.22
worst compactness	0.25	0.16	0.03	0.15	0.21	0.34	1.06
worst concavity	0.27	0.21	0.00	0.11	0.23	0.38	1.25
worst concave points	0.11	0.07	0.00	0.06	0.10	0.16	0.29
worst symmetry	0.29	0.06	0.16	0.25	0.28	0.32	0.66
worst fractal dimension	0.08	0.02	0.06	0.07	0.08	0.09	0.21

上表展式了 30 个特征的平均值，标准差，最大最小值以及四分位数。对于标准差来说，mean area（面积平均值）、worst perimeter（周长最大值）和 worst area（面积最大值）的值较大，分别为 654.89、107.26 和 880.58，也就是说不同样本之间这几个特征的差异最大，在之后选择分析特征时可以给予着重考虑。

2、特征相关性

在本案例中，每个样本有 30 个特征，若将所有特征纳入考虑，无疑大大增加了计算复杂度，因此，首先对特征相关性进行分析。绘制特征相关性热力图如下：

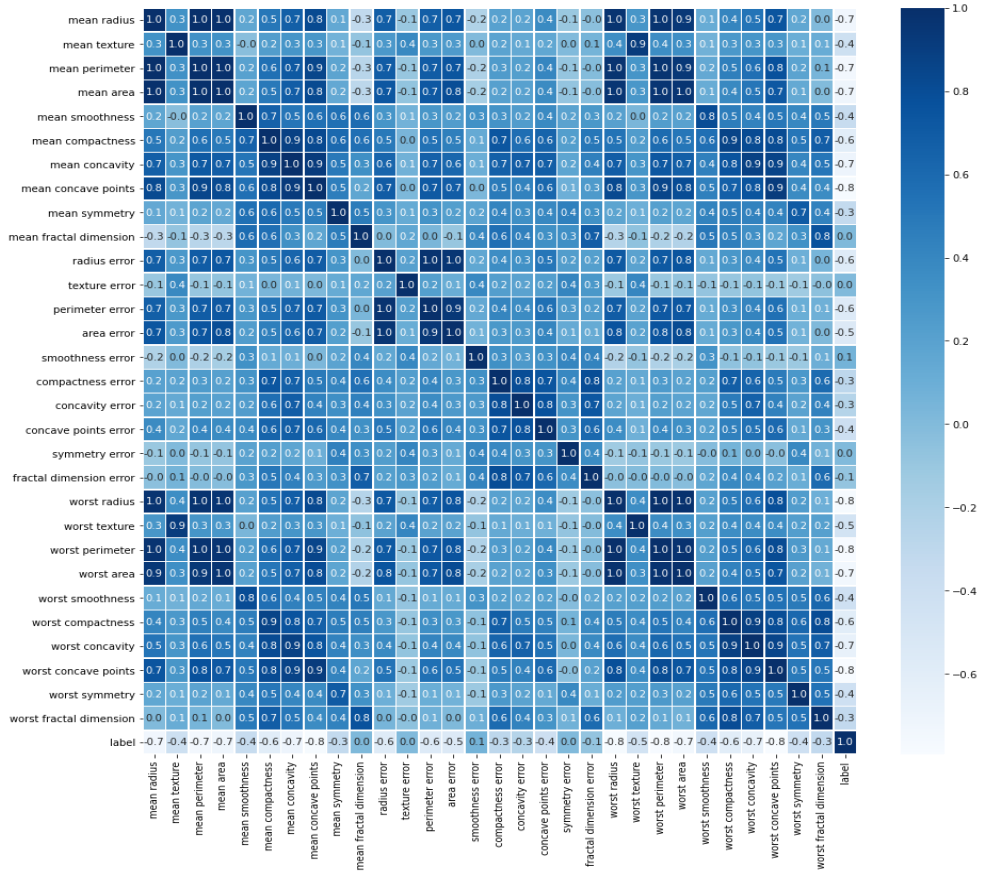

图 1 特征相关性

上图表现出了样本 30 个特征之间的相关性。首先，每个特征与自身的完全相关性是毋庸置疑的；其次，各个特征之间也有一定的相关性，如 mean radius（半径平均值）与（mean perimeter（周长平均值）、mean area（面积平均值）、worst radius（半径最大值）、worst perimeter （周长最大值）之间的相关系数为 1，与 worst area（面积最大值）之间的相关系数为 0.9……

不难看出，大多数特征之间都存在较高的相关性，因此，在后续实验中，无需将所有特征都纳入考虑，而是从中选取代表性较高的特征即可。

三、 特征工程

1、卡方检验

采取卡方检验的方法选取特征，选择得分高且 P 值低的特征，首先尝试选取 6 个特征，选择结果返回值如下：

```
array([[ 122.8  , 1001.   ,  153.4  ,   25.38 ,  184.6  , 2019.   ],
       [ 132.9  , 1326.   ,   74.08 ,   24.99 ,  158.8  , 1956.   ],
       [ 130.   , 1203.   ,   94.03 ,   23.57 ,  152.5  , 1709.   ],
       ...,
       [ 108.3  ,  858.1  ,   48.55 ,   18.98 ,  126.7  , 1124.   ],
       [ 140.1  , 1265.   ,   86.22 ,   25.74 ,  184.6  , 1821.   ],
       [  47.92 ,  181.   ,   19.15 ,    9.456,   59.16 ,  268.6  ]])
```

即选择 mean perimeter（周长平均值）、mean area（面积平均值）、area error（面积标准差）、worst radius（半径最大值）、worst perimeter （最大周长）、worst area（面积最大值）这 6 个特征。

通过特征分析，我们已知 mean perimeter（周长平均值）、mean area（面积平均值）、worst radius（半径最大值）、worst perimeter （最大周长）、worst area（面积最大值）之间的相关系数为 1，即完全相关，因此上述特征无需全部纳入考虑，尝试选取 3 个特征，选择结果返回值如下：

```
array([[1001.  ,  153.4 , 2019.  ],
       [1326.  ,   74.08, 1956.  ],
       [1203.  ,   94.03, 1709.  ],
       ...,
       [ 858.1 ,   48.55, 1124.  ],
       [1265.  ,   86.22, 1821.  ],
       [ 181.  ,   19.15,  268.6 ]])
```

即选择 mean area（面积平均值）、area error（面积标准差）、worst area（面积最大值）这三个特征。

2、方差选择法

采取方差选择法选取特征，选择方差大于 1200 的特征，选择结果返回值如下：

```
array([[1001.  ,  153.4 , 2019.  ],
       [1326.  ,   74.08, 1956.  ],
       [1203.  ,   94.03, 1709.  ],
       ...,
       [ 858.1 ,   48.55, 1124.  ],
       [1265.  ,   86.22, 1821.  ],
       [ 181.  ,   19.15,  268.6 ]])
```

即选择 mean area（面积平均值）、area error（面积标准差）、worst area（面积最大值）这 3 个特征，与卡方检验的选择结果一致。

3、递归特征消除法

采取递归特征消除法选取特征，选择结果返回值如下：

```
array([[1.095 ,  0.7119,  0.2654],
       [0.5435,  0.2416,  0.186 ],
       [0.7456,  0.4504,  0.243 ],
       ...,
       [0.4564,  0.3403,  0.1418],
       [0.726 ,  0.9387,  0.265 ],
       [0.3857,  0.    ,  0.    ]])
```

即选择 radius error（半径标准差）、worst concavity（凹度最大值）、worst concave points（凹缝最大值）这 4 个特征，与前两种方法选择结果不一致。

4、结论

通过上述 3 种特征选择方法，得到了两种不同的结论，为了保证后续实验顺利进行，对选取特征之间的关系进行分析。

（1）绘制散点图矩阵及三维图分析 mean area（面积平均值）、area error（面积标准差）、worst area（面积最大值）这 3 个特征之间的关系：

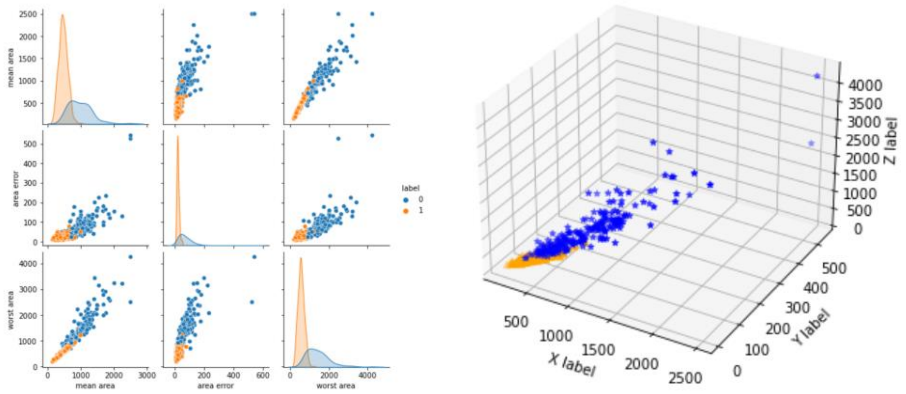

图 2 特征之间的关系（a）

由上面两幅图可以看出，在选择这 3 个特征时，样本的分类边界十分清晰，几乎没有交织在一起的两类样本点，因此在后续分类中，采取这 3 个特征时的结果精确度应该较高。

（2）绘制散点图矩阵分析 radius error（半径标准差）、worst concavity（凹度最大值）、worst concave points（凹缝最大值）这 3 个特征之间的关系：

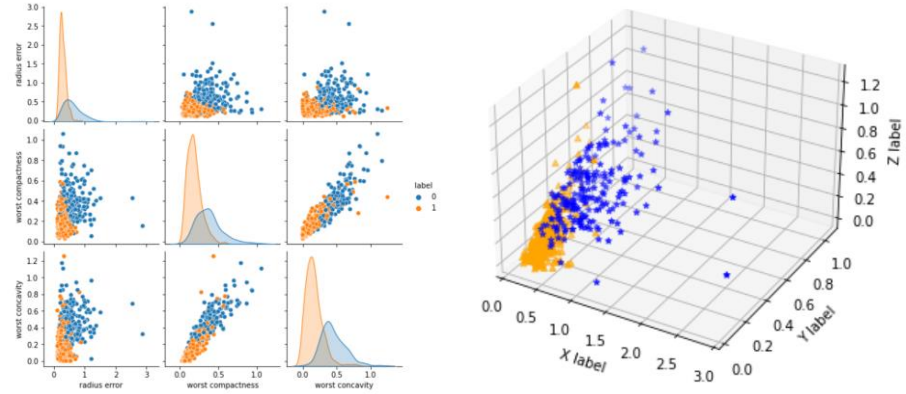

图 3 特征之间的关系（b）

由上图可以看出，在选择这 3 个特征时，有大量不同类别的样本点相互交织，不易分类，

因此在后续分类中，不应采取这一组特征。

综上所述，选择的特征为 mean area（面积平均值）、area error（面积标准差）和 worst area（面积最大值）。

四、乳腺癌数据的分类

1、形式化和分类器数学推导

（1）KNN 算法

KNN（K-NearestNeighbor）分类算法是数据挖掘分类技术中最简单的方法之一。所谓 K 最近邻，就是 K 个最近的邻居的意思，说的是每个样本都可以用它最接近的 K 个邻近值来代表。近邻算法就是将数据集合中每一个记录进行分类的方法。

在本案例中，KNN 算法步骤如下：

第一步:计算需要预测的没有标签的乳腺癌测试集数据点与所有给定乳腺癌训练集数据点的欧式距离；

第二步：对第一步中所求的距离进行排序，并选择前面 K 个距离对应的点。这里对于 K 的取值选取是一个重点问题，我们通过交叉验证的方法绘制交叉验证曲线，并选取在训练集上最稳定最优的 K 值；

第三步：统计这 k 个点中三种乳腺癌种类所对应的的频数，并将出现最大频数的乳腺癌类别作为该没有标签的乳腺癌测试集数据点的标签。

（2）SVM 算法

SVM 的核心在于找到一个超平面将两类样本准确的分开，同时保证间隔尽可能的大，这样会有更好的泛化能力。乳腺癌数据是一个线性可分数据集，于是我们无需用到 SVM 中的核函数。

设超平面方程为:$w^T x + b = 0$,我们需要做的找到这样一个超平面划分两类乳腺癌并使得乳腺癌训练集数据上的点到这个超平面的间隔距离尽可能的远。接着将原问题通过拉格朗日对偶算法转换为其对偶问题，其主要原因是自然引入核函数从而降低求解复杂度，虽然乳腺癌数据并不利用核函数，但这种方法已经约定俗成。其对偶优化问题如下，我们解决该优化问题即可。

$$max\ L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s.t.$$

$$\begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i > 0, i = 1,2 \ldots n \end{cases}$$

（3）Logistic Regression 算法

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。在乳腺癌案例中，我们通过许多带有分类标签（乳腺癌有两种类别）的特征变量数据来训练预测乳腺癌类别的模型。为实现预测分类问题，我们使用了 Logistic 模型及 Sigmoid 函数:

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

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:

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

我们可以把$log_i(z)$的函数值看成类别为 1 的概率预测值，当$log_i(z)<0.5$时，我们预测的分类为 0；当$log_i(z)>=0.5$时，我们预测的分类为 1。这样我们就可以很容易的对乳腺癌数据集分类，即良性或者恶性。

2、计算机仿真过程

（1）KNN

在 jupyter notebook 中调用 sklearn.linear_model 包构建基于 KNN 的乳腺癌分类模型。

（2）SVM

在 jupyter notebook 中调用 svc 构建基于支持向量机的乳腺癌分类模型。

（3）Logistic Regression

在 jupyter notebook 中调用 sklearn.linear_model 包构建基于逻辑回归的乳腺癌分类模型。

3、评估方法和数据划分

（1）混淆矩阵

混淆矩阵向我们展式了查准率(准确率)与查全率(召回率)：

查准率(P) = $\frac{TP}{TP+FP}$ ，即在被判别为正类别的样本中，确实为正类别的比例是多少；

查全率(R) = $\frac{TP}{TP+FN}$ ，即在所有正类别样本中，被正确判别为正类别的比例是多少。

（2）ROC 曲线

模型训练完成之后，每个样本都会获得对应的两个概率值，一个是样本为正样本的概率，一个是样本为负样本的概率。把每个样本为正样本的概率取出来，进行排序，然后选定一个阈值，将大于这个阈值的样本判定为正样本，小于阈值的样本判定为负样本，可以得到两个值，一个是真正率，一个是假正率：

真正率（TPR）= $\frac{TP}{TP+FN}$ ，即模型判定为正样本且实际为正样本的样本数与所有的正样本数之;

假正率（FPR）= $\frac{FP}{TN+FP}$ ，即模型判定为正样本实际为负样本的样本数与所有的负样本数之比。

每选定一个阈值，就能得到一对真正率和假正率，由于判定为正样本的概率值区间为

[0,1]，那么阈值必然在这个区间内选择，因此在此区间内不停地选择不同的阈值，重复这个过程，就能得到一系列的真正率和假正率，以这两个序列作为横纵坐标，即可得到 ROC 曲线了。而 ROC 曲线下方的面积，即为 AUC 值。

(3)数据划分

在本案例中，共有 569 个样本，其中正例 212 个，负例 357 个。将样本的 70%划分为训练集，30%划分为测试集，并训练集、测试集中正、负例的比例与原数据集尽量一致。划分结果如下：

```
1    245
0    153
Name: label, dtype: int64
1    112
0     59
Name: label, dtype: int64
```

4、分类结果和比较统计

（1）KNN 算法

通过 Sklearn 的 KNeighborsClassifier 模型，取 K=14，对乳腺癌数据集进行分类，结果如下：

```
准确率 0.9707602339181286
             precision    recall  f1-score   support

          0       0.98      0.94      0.96        63
          1       0.96      0.99      0.98       108

avg / total       0.97      0.97      0.97       171
```

可以看出，模型分类的精度达到了 97%。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：

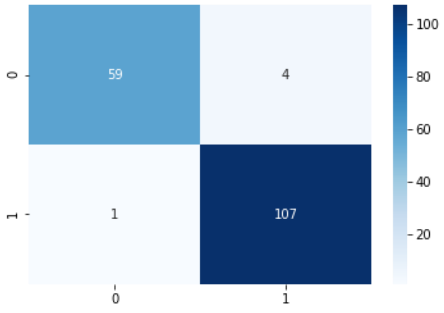

图 4 KNN 算法混淆矩阵

由上图可知，此模型将本应分类为良性肿瘤（标签为 0）的 4 个样本误分为恶性肿瘤（标签为 1），同时将本应分类为恶性肿瘤（标签为 1）的 1 个样本误分为良性肿瘤（标签为 0）。也就是说，此模型在良性肿瘤的分类中，查准率(P)达到了 98%，查全率（R）达到了 94%；在恶性肿瘤的分类中，查准率(P)达到了 96%，查全率（R）达到了 99%。

对比来看，对于良性肿瘤（标签为 0）的检测结果没有恶性肿瘤（标签为 1）的检测结果好，原因可能是因为 K 阶邻近法中如果两个类别的个数相差大的话。在取 K 个距离时大概率会取到类别样本数大的那个类别，从而影响判断。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

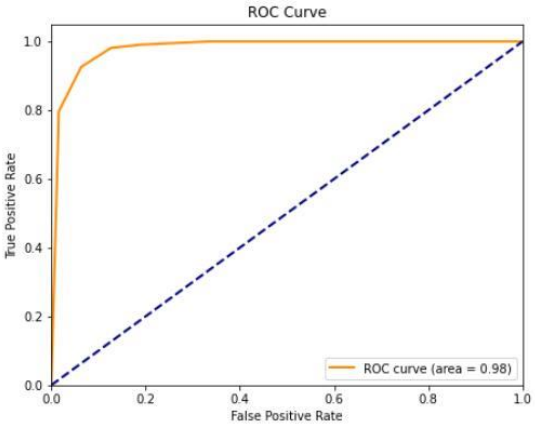

图 5 KNN 算法 ROC 曲线

由上图可知，曲线十分接近于左上角，曲线下方的面积（AUC）为 0.98，说明此模型分类准确率较高。

（2）SVM

通过 Sklearn 的 LogisticRegression 模型，取 C=1000，solver='lbfgs'，对乳腺癌数据集进行分类，结果如下：

准确率 0.9415204678362573

	precision	recall	f1-score	support
0	0.96	0.87	0.92	63
1	0.93	0.98	0.95	108
accuracy			0.94	171
macro avg	0.95	0.93	0.94	171
weighted avg	0.94	0.94	0.94	171

可以看出，模型分类的精度达到了 94%。其次，我们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：

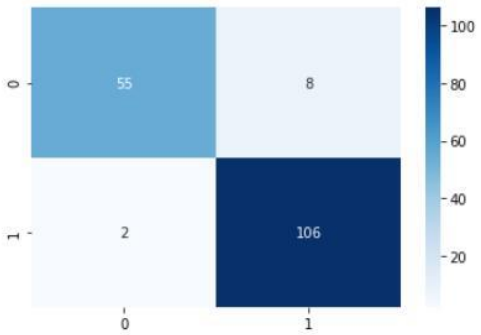

图 6 SVM 算法混淆矩阵

由上图可知，此模型将本应分类为良性肿瘤（标签为 0）的 8 个样本误分为恶性肿瘤（标签为 1），同时将本应分类为恶性肿瘤（标签为 1）的 2 个样本误分为良性肿瘤（标签为 0）。也就是说，此模型在良性肿瘤的分类中，查准率(P)达到了 96%，查全率（R）达到了 87%；在恶性肿瘤的分类中，查准率(P)达到了 93%，查全率（R）达到了 98%。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

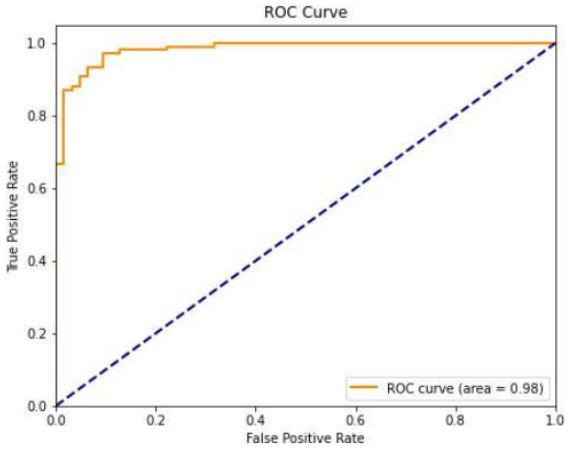

图 7 SVM 算法 ROC 曲线

由上图可知，曲线十分接近于左上角，曲线下方的面积（AUC）为 0.98，说明此模型分类准确率较高。

（3）Logistic Regression

通过 Sklearn 的 SVC 模型对乳腺癌数据集进行分类，分类结果如下：

准确率: 0.9473684210526315

	precision	recall	f1-score	support
0	0.94	0.92	0.93	63
1	0.95	0.96	0.96	108
accuracy			0.95	171
macro avg	0.94	0.94	0.94	171
weighted avg	0.95	0.95	0.95	171

可以看出，模型分类的精度达到了 93%，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：

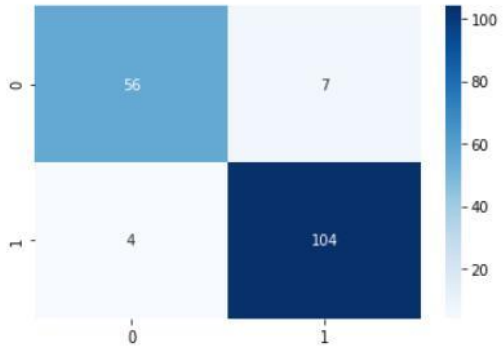

图 8 LR 算法混淆矩阵

由上图可知，此模型将本应分类为良性肿瘤（标签为 0）的 7 个样本误分为恶性肿瘤（标签为 1），同时将本应分类为恶性肿瘤（标签为 1）的 4 个样本误分为良性肿瘤（标签为 0）。也就是说，此模型在良性肿瘤的分类中，查准率(P)达到了 94%，查全率（R）达到了 92%；在恶性肿瘤的分类中，查准率(P)达到了 95%，查全率（R）达到了 96%。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

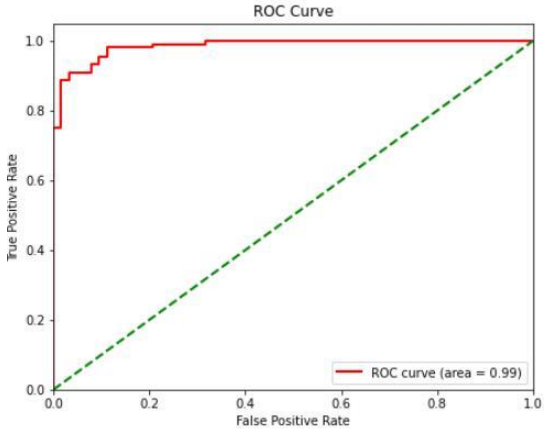

图 9 LR 算法 ROC 曲线

由上图可知，曲线十分接近于左上角，曲线下方的面积（AUC）为 0.99。

5、结果可视化

（1）KNN

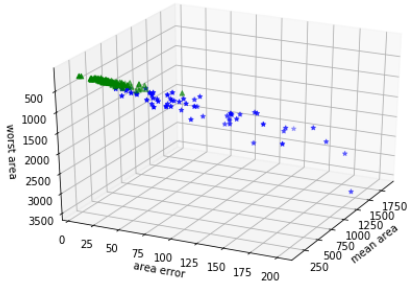

图 10 KNN 分类结果

（2）SVM

SVM 模型可以在两类数据之间找到一个超平面作为决策边界，将两类数据分隔开。如下图，中间蓝色的平面将两类数据分隔。

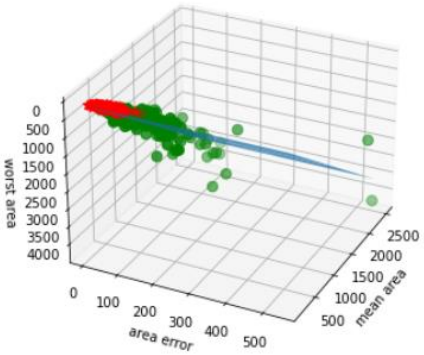

图 11 SVM 分类结果

（3）Logistic Regression

逻辑回归模型在两类数据之间找到一个超平面作为决策边界，将两类数据分隔开。如下图，中间蓝色的平面将两类数据分隔。

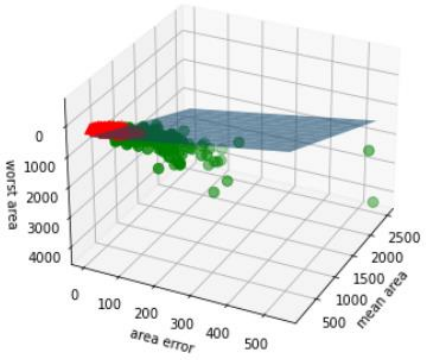

图 12 LR 分类结果

五、结果分析和得出结论

本文以乳腺癌数据集为实验数据，采用 KNN、SVM 和逻辑回归模型对数据集进行分类，该数据集无缺失值，较好分类。KNN 模型的分类准确率为 97%，SVM 模型的分类准确率为 94%，逻辑回归模型的分类准确率为 93%，比较来看，KNN 模型的分类准确率较优于其他两种模型的分类准确率。

基于 10 折交叉验证法，绘制三种方法的学习曲线，对比乳腺癌数据集分类准确率，结果如下：

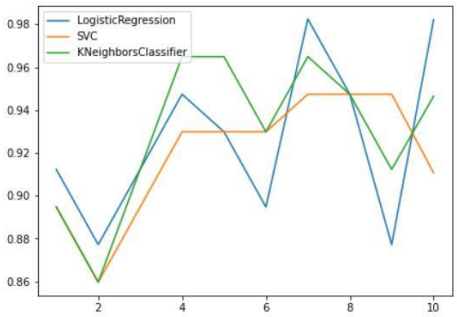

图 13 分类结果对比

由上图可以看出，三种方法总体分类准确率差距不大，但是 KNN 模型分类准确率多次高于 SVM 和逻辑回归模型的分类准确率，与直接训练得到的分类准确率结果一致。

六、实践总结

本次实践通过逻辑回归、SVM 以及 KNN 算法实现了对于乳腺癌数据的良性恶性肿瘤分类。相比较上一次实践中的鸢尾花数据，乳腺癌数据集有更多的特征，需要在特征选取上下功夫，我应用了卡方检验、方差选择法以及递归特征消除法进行特征选择。在分类的过程中，我加深了对逻辑回归、SVM 以及 KNN 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，针对于 KNN 算法在划分数据集中未注意训练集中各个分类下数据的数量对于结果的影响。在之后的修改补充中应考虑不同算法对于训练集各个分类下数据个数有没有要求，从而使得训练得到的分类器具有更好的分类效果。