**Assignment2 报告**

姓名：夏一凡

班级：2019计算机科学与技术二班

摘要：下载可以用于分类的多个数据集，并学习多个分类模型的算法并自定义构造分类器，如：逻辑回归分类器、决策树分类器、朴素贝叶斯分类器、神经网络分类器等。对于比较复杂的模型SVM分类器，调用该模型即可。选择下载好的数据集对模型进行训练和测试，查看该模型的相关评价指标和最终的准确率得分来评估该模型的性能。

# 引言

本文中涉及到了五个分类器：逻辑回归分类器、决策树分类器、朴素贝叶斯分类器、神经网络分类器和SVM分类器。旨在学习每个分类器的算法，可以对简单的分类器实现算法自定义；学会用混淆矩阵对每个模型的结果可视化，可以对可视化结果有清晰的认识和分析；并用同一数据集对比某些模

型的accuracy,precision,recall,F1-measure等相关指标，分析模型的性能。

# 算法

逻辑回归算法[1]：

逻辑回归模型能够用于解决二分类问题。

从线性回归到逻辑回归：从线性回归模型出发，线性回归预测函数输出值y是连续变量，值域为R；逻辑回归的输出值y是离散型变量，值域为{0， 1}；逻辑回归模型可以被认为是一个被Sigmoid函数（logistic方程）所归一化后的线性回归。

逻辑回归原理：逻辑回归假设函数为：

hθ(X) = g(Xθ) = 1/(1+e-Xθ) 公式(2.1.1) 公式(2.1.1)中，X为样本输入，hθ(X)为模型输出，θ为要求解的模型参数。设0.5为临界值，当hθ(X)>0.5时，即Xθ>0时，y为1；当hθ(X)<0.5 时，即Xθ<0时，y为0.模型输出值hθ(X)在[0,1]区间内取值，因此可从概率角度进行解释：hθ(X)越接近于 0，则分类为 0 的概率高；hθ(X)越接近于 1，则分类为1的概率越高；hθ(X)越接近于临界值0.5，则无法判度，分类准确率会下降。

决策树算法（CART算法）[2]：

为了构建决策树，先定义一个节点类，类中包括以下几个变量：划分维度、划分阙值、是否是叶结点、左支、右支、分类。

基尼值的计算公式：

Gini(D) = 1-∑𝑑𝑣=1 𝑝(𝑣)^2 公式(2.2.1)

公式(2.2.1)中，基尼值Gini越小说明该数据集中不同类的数据越少， p(v)代表了v类数据在总类中的频率基尼系数的计算公式：

GiniIndex(D,a) = ∑𝑉𝑣=1 𝑁(𝑣) ∗ 𝐺𝑖𝑛𝑖(𝐷(𝑣)/𝑁

公式(2.2.2) 因为D1数据集和D2数据集的属性都是浮点数，为了二分化，需要一个阙值，这里用枚举左右的划分情况的方法。步骤如下：

(1). 排序给定维度下的属性；

(2). 选取相邻属性值的平均值作为候选阙值并去重；

(3). 遍历所有可能的阙值，选取基尼系数最小的划分阙值，返回基尼系数和划分阙值。

下一步是寻找划分维度：因为D1数据集和D2数据集中都有多个维度的数据，所以需要确定哪个维度的数据作为依据。依次计算各个维度下的最小基尼系数，并选取最小基尼系数最小的维度作为划分维度。

最后一步构建决策树：

用以上的算法，再加递归的方法就可以构建决策树。递归的终点有两种情况：dataset只有一个元素了，那就不用再分了；dataset里有很多元素，但都是同一类型的，体现在GiniIndex=0，说明已经纯洁，不用继续递归。

朴素贝叶斯算法[3]：

贝叶斯定理的公式：

P(A|B) = P(B|A) \* P(A) / P(B) 公式(2.3.1) 常用于解决分类问题。朴素贝叶斯的朴素就体现在它假设所有的属性（即特征）之间相互独立，这一假设可以表达为公式(2.3.2)：

P(X|Y=y) =  公式(2.3.2)

这样一来，前面的贝叶斯定理公式(2.3.1)就可以描述为公式(2.3.3)：

P(Y|X) = P(Y) \*  / P(X) 公式(2.3.3)

使P(Y|X)最大的类别Y就是样本X所述的类别。而对于每个样本来说，

P(X)是不随标签Yi改变的，所以，只需比较P(Y) \* 就可以了。

神经网络算法[4]：

神经网络分为三层：输入层、隐含层和输出层。激活函数默认为sigmoid 函数。

前向传播：

(1). 输入层到隐含层的前向传播：计算神经元的输入加权和，并计算该

神经元的输出（此处用到sigmoid函数）；同理计算每个隐含层神经元的输出。

(2). 隐含层到输出层：计算输出层的每个神经元。

反向传播：

(1). 计算总误差公式(2.4.1)：

Etotal = ∑(𝑡𝑎𝑟𝑔𝑒𝑡 − 𝑜𝑢𝑡𝑝𝑢𝑡)2/2 公式(2.4.1)

(2). 做隐含层到输出层的权值更新；

(3). 做隐含层到隐含层的权值更新。

这样误差反向传播就完成了。最后再把更新的权值重新计算，不停的迭代，总误差会下降。

SVM算法[5]：

SVM 学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。

SVM是一种监督学习算法，主要思想是建立一个最优决策超平面，使得

该平面两侧距平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力。以图2-5-1为例，黄色和蓝色是两种决策超平面，而黄色平面两侧距平面最近的两类样本之间的距离较大，所以可以称黄色是最优决策超平面。

而“支持向量”指训练集中的一些训练点，这些训练点最靠近决策面，是最难分类的数据点。比如图2-5-1中，画了虚线的四个点就是这种点。

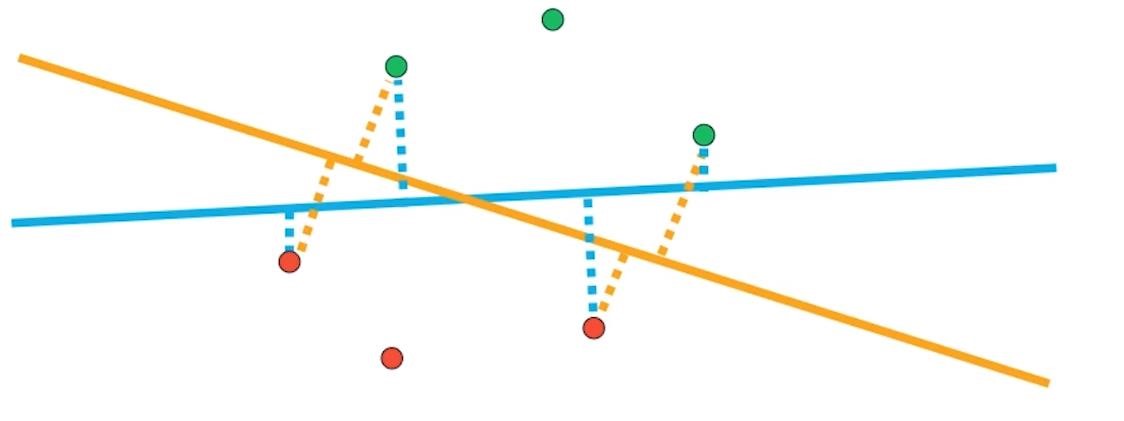


图 2-5-1

寻找到这类超平面后，我们假设超平面方程为公式(2.5.1)：

WT \* X+b=0 公式(2.5.1)

X为输入向量，W为权值向量，b为偏置，则可根据以下两个标准分为两类：WT \* X + b > 0和WT \* X + b < 0

# 实验及结果分析

逻辑回归：

自定义一个logisticRegression类，其中包括以下几个方法：sigmoid 函数的实现的方法，根据提供的训练数据对模型进行训练的方法，和预测方法等。

将D4乳腺癌数据集按照训练集：测试集=8：2的比例划分出训练集和测试集；

调用定义好的模型，输出根据预测属性得到的预测结果；

根 据 预 测 结 果 和 真 实 结 果 ， 绘 制 混 淆 矩 阵 ， 并 计 算 出 accuracy,precision,recall,F1\_measure这四个评价指标。

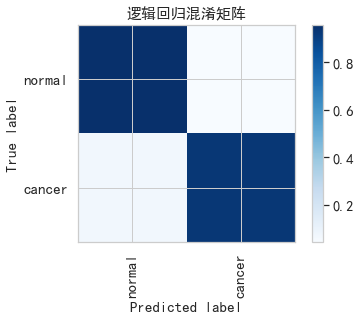


图 3-1-1

图3-1-1是最终得到的混淆矩阵，横轴是预测标签，纵轴是真正标签，该矩阵中的四个数字的意义：66代表TN(真实值为0，预测结果也为0),0代表FP(真实值为0，预测结果为1),22代表FN(真实值为1，预测结果为0),26 代表TP(真实值为1，预测结果也为1).四个数的总和代表总的测试类样本数，从左向右对角线的总和是预测结果正确的样本数。下面四个指标的计算需要用到这四个变量。四个指标的计算公式：

accuracy = (TP+TN)/(TP+TN+FP+FN) precision = TP/(TP+FP) recall = TP/(TP+FN)

F1\_measure = (2\*precision\*recall)/(precision+recall)

结果如图3-1-2：

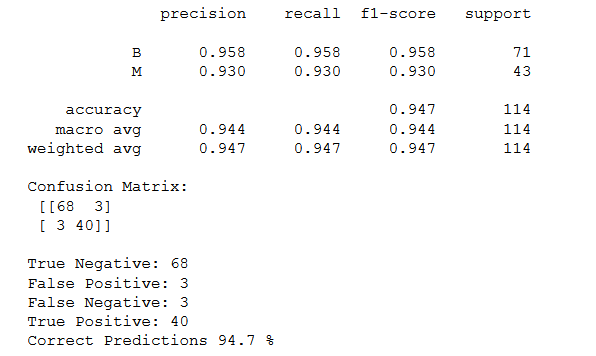


图 3-1-2

对应的柱状图（图3-1-3）：

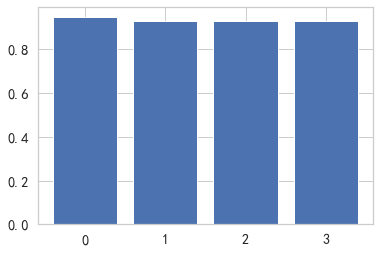


图 2-1-3

图3-1-2是各个指标的柱状图，但是由于各个指标之间的数值相差不大，

所以这里转成折线图（图3-1-3）再进行分析。

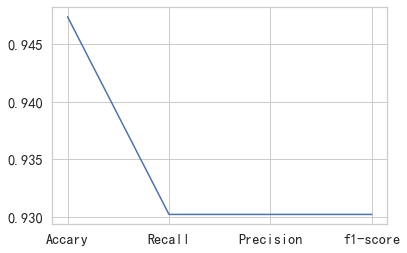


图 3-1-3

从折线图（图3-1-3）可以看出，accuracy在0.947左右，而

precision = recall = F1\_measure < accuracy ,由计算公式可知，造成三个指标相同的原因是混淆矩阵中 FP = FN。以上四个评价指标都至少大于

0.930，所以该模型适合该数据集。

决策树：

定义节点类并构造相关函数，函数和相关作用参见代码；读取数据集D1，做好预处理后划分训练集和测试集；

调用定义好的相关方法，构建对应的决策树，并用X\_test做测试，对比原y\_test数据；

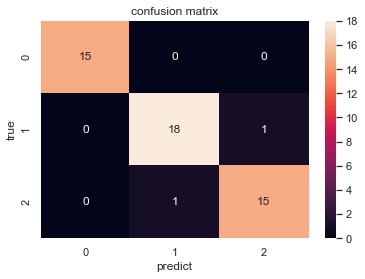
用预测结果和真实结果绘制混淆矩阵，并计算相关指标，绘制对应的折线图。

（D2数据集采用同样的步骤）得到如下结果：

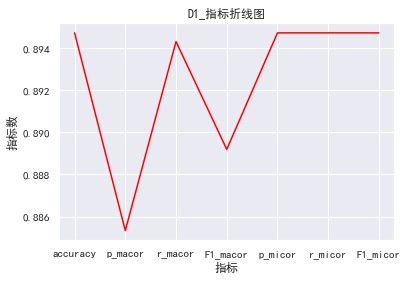
图 3-2-1

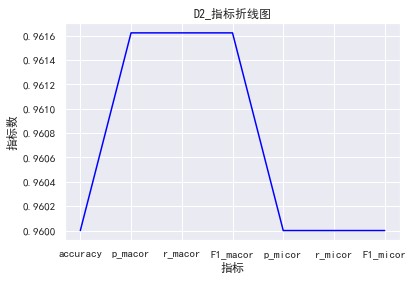
由预测结果的混淆矩阵（图3-2-1）可以看出，测试数据集中共有346组数据，其中分布在对角线上的数据占了绝大多数，即根据该模型得到的预测正确的样本占绝大多数，准确度为0.8930635838150289。

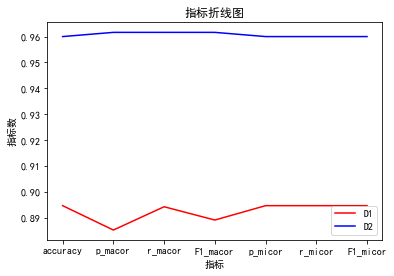
没有分布在对角线上的数据都在对角线附近且数量较少，可以得知该模型的准确率较高。

图 3-2-2

根据该混淆矩阵（图3-2-2）可以看出，共有50个预测样本，且预测结果中预测错误的样本只有两个，分别对应真实特征值为2时的一个样本，模型将其预测成了1，真实特征值为1的一个样本，模型将其预测成了2。且每一行的元素和都差不多相同，所以测试集中这三个类的样本数量差别不大。

图 3-2-3

图 3-2-4

图 3-2-5

从以上折线图（图3-2-5）中可以看出，蓝色折线明显在红色折线上方。蓝色折线对应的是D2数据集，红色折线对应的是D1数据集，所以该模型更适合D2数据集的训练和预测，但对于数据集D1来说，它的每个指标的值都在0.88到0.90之间，所以对于数据集D1，该模型也比较合适，只是更适用于数据集D2。且每条折线后三个指标部分都是平行于x轴的，所以可以得出结论：precision\_micor = recall\_micor = F1\_measure\_micor.还可以从每个数据集单个的折线图中更容易得到这三个指标还与accuracy相等。

朴素贝叶斯：

定义朴素贝叶斯类，包括朴素贝叶斯的相关算法步骤；读取数据集D1，做好预处理后划分训练集和测试集；

实例化朴素贝叶斯类，训练模型，并用 X\_test 做测试，对比原 y\_test 数据；

用预测结果和真实结果绘制混淆矩阵，并计算相关指标，绘制对应的折线图。

（D2数据集采用同样的步骤）

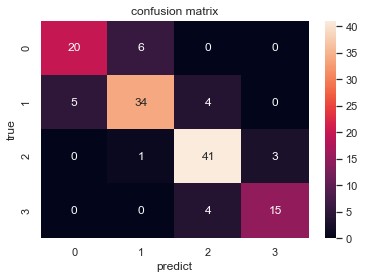
图 3-3-1

图 3-3-1 是使用朴素贝叶斯的方法，对数据集 D1 的预测结果的混淆矩阵。相较于用决策树模型得到的结果，该混淆矩阵中对角线上的元素和明显少于图中对角线元素之和，且每一类都有预测错误的样本，其中，真实特征值为1时，预测错误的样本数量最多。

图 3-3-2

图 3-3-2 是使用朴素贝叶斯的方法，对数据集 D2 的预测结果的混淆矩阵。总体来说对角线的元素之和远大于非对角线的元素之和，所以该模型是适合该数据集的。

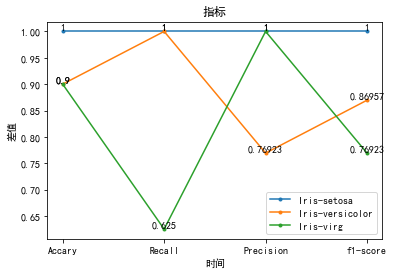
图 3-3-3

图 3-3-4

图 3-3-5

由图3-3-5可知，该模型用于数据集D1和数据集D2时，D2对应的折线图在D1之上，且每项指标都在0.92附近，所以该模型更适合数据集D2。而对于数据集D1，该模型对应的折线上的每个点都在0.82附近，比数据集D2 的每项指标都相差大概0.1，所以该模型更适用于D2。且根据单个的放大图

（图3-3-3和图3-3-4）可知，指标还是符合accuracy = precision\_micor = recall\_micor = F1\_measure\_micor.

神经网络：

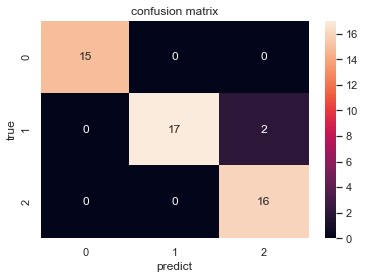
定义神经网络相关的方法：初始化参数、前向传播、计算代价函数、反向传播、更新参数、模型预测、建立神经网络等方法；

读取数据集D2，做好预处理后划分训练集和测试集；

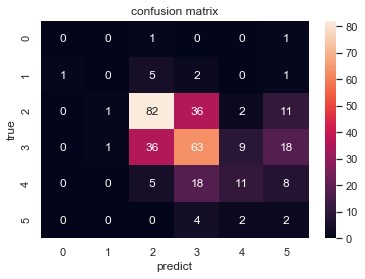
根据以上定义函数建立模型，训练模型，并用 X\_test 做测试，对比原 y\_test数据；

用预测结果和真实结果绘制混淆矩阵，并计算相关指标，绘制对应的折线图。

（D3数据集采用同样的步骤）

图 3-4-1

该混淆矩阵（图3-4-1）是在神经网络模型下，对数据集D2进行的预测结果与真实结果的比较。从图中可知，原样本类别是0和2时，预测结果都正确；原样本类别为 1 时，预测结果中有两个样本被预测成了类别 2；对角线元素之和为48，而非对角线元素之和只有2，准确率达到了0.96，该模型适合该数据集。

图 3-4-2

该混淆矩阵（图3-4-2）是神经网络模型对数据集D3做的样本分类。从每一行可以看出，类别为0，1，5的样本并不多，大多数样本都集中在2，3， 4 这三个类别，且真实类别为 2 的样本中，通过神经网络模型预测正确的结果最多，而样本0和样本1中，样本量很少，且都没有预测正确。对角线的元素之和与非对角线的元素之和相差不大，初步估计该模型的准确率在 0.5 附近，所以该模型对数据集D3不是很合适。

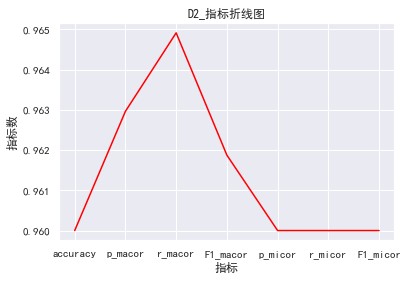


图 3-4-3

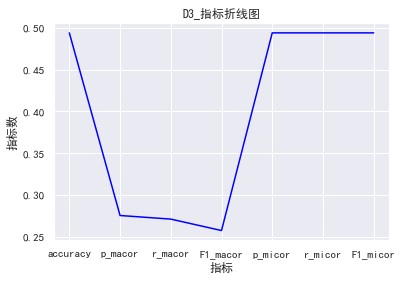
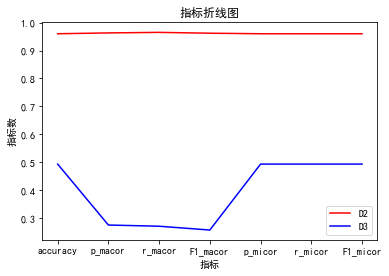


图 3-4-4

图 3-4-5

从图 3-4-5 中可以得出以下结论：该神经网络模型对于数据集 D2 更合适，因为数据集D2根据该模型预测得到的各个指标的数值都在0.95到1.0 之间；而对于数据集D3，通过该模型预测得到的各个指标却在02到0.5之间，所以并不适用于数据集D3。但对于每个一数据集，还是符合accuracy = precision\_micor = recall\_micor = F1\_measure\_micor.

SVM：

定义函数：建立模型的函数、训练模型的函数、预测函数、打印分数函数；

读取数据集D2，做好预处理后划分训练集和测试集；

根据以上定义函数建立模型，训练模型，并用 X\_test 做测试，对比原 y\_test数据；

用预测结果和真实结果绘制混淆矩阵，并计算相关指标，绘制对应的折线图。

（D3数据集采用同样的步骤）

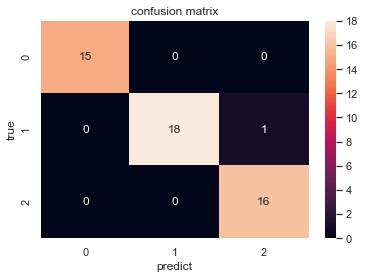
图 3-5-1

图 3-5-1 是通过 SVM 模型对数据集 D2训练预测得到的预测结果和真实样本属性的混淆矩阵。从该混淆矩阵可以看出，预测结果中只有一个是错误的，即把真实样本的1类别预测成了2类别。且从混淆矩阵中可以看出，在测试集中，1 类别的样本数量最多。该模型的对角线数据之和远大于非对角线数据之和，准确率高，适用于数据集D2.

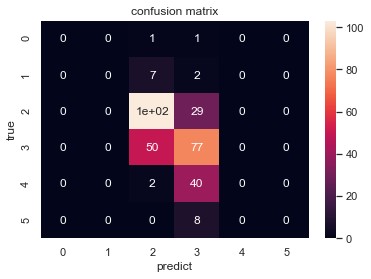
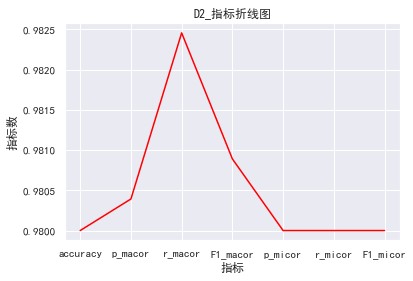
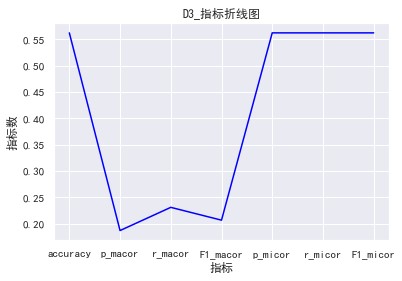
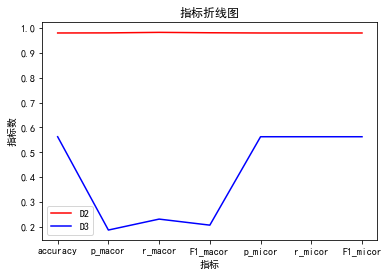
图 3-5-2

图 3-5-2 是通过 SVM 模型对数据集 D3训练预测得到的预测结果和真实样本属性的混淆矩阵。可以很明显的看出，测试样本中，大多数样本都是2， 3 类，这两个类被对于准确率的影响很大。而在这两个类别中，类别 2 的样本预测正确的结果更多。还可以看出类别0，1，4，5中虽然在测试数据集中有这4个样本，但在预测结果中这四个类的所有样本都预测错误。初步估计该SVM模型的对于数据集D3而言，准确率在0.6附近。

图 3-5-3

图 3-5-4

图 3-5-5

从这个折线图（图3-5-5）可以看出，数据集D2对应的各项指标都在0.97 到1.0之间，数据集D3对应的各指标都在0.2或0.6附近，所以数据集D2 对应的各指标的值都在数据集D3各项指标之上，且远大于D3的各项指标，所以SVM模型对数据集D2很合适，对数据集D3的效果并不好。最后各数据

集上的指标都符合 accuracy = precision\_micor = recall\_micor = F1\_measure\_micor.

# 结论

本文主要是对逻辑回归模型、决策树分类模型、朴素贝叶斯分类器模型、神经网络分类器模型和SVM分类器模型等五个分类模型进行学习。对每个模型用一个或两个不同的数据集进行训练和预测，并计算各个评价指标，评价指标可视化并做对比，看看哪个模型更适用于哪个数据集。

以上指标可视化的结果中，不仅可以根据它评定模型的好坏，也验证了

一个结论：关于多分类问题，它有四个指标相同：accuracy = precision\_micor = recall\_micor = F1\_measure\_micor.

# 参考文献

[1]欧高炎,朱占星,董彬,鄂维南.数据科学导引[M].高等教育出版社, 70-74. [2]欧高炎,朱占星,董彬,鄂维南.数据科学导引[M].高等教育出版社, 79-88.

[3]欧高炎,朱占星,董彬,鄂维南.数据科学导引[M].高等教育出版社, 93.

[4]欧高炎,朱占星,董彬,鄂维南.数据科学导引[M].高等教育出版社, 95-98.