** **

**2020年兰州大学数学建模竞赛**

论文题目： 混合物中特定成分的检测

团队队长： 张鑫

所在学院： 信息科学与工程学院

团队成员： 张子涵、钟永玮

比赛时间： 2020.7.18 至 2020.7.20

摘要

针对竞赛题目，在团队成员详细研究后，明确了解题的思路。混合物中特定成分的检测，给出了大量的带标签的训练数据，以及无标签的测试数据，则可以明确，对未知混合物中是否含有特定成分的判断，是机器学习中有监督学习的二分类问题，本文就从这一方面着手解决问题。

对于任务一，判定特定成分存在的主要指标，可以从不同特征与标签的相关性入手，根据各个特征和判定标签(是否含有特定成分(1为含有，0为不含)）)的相关性，去判断特征的重要程度，选择出主要指标。用python的sklearn科学计算库的特征选择算法，对主要特征进行选择，验证判断。

对于任务二，指标中模糊区域的判断，可以分别绘制出label=0 和 label=1对应的单个特征变量的dist(密度直方图)和kde(核密度估计图-平滑后的密度直方图)），查看两种标签下数据分布情况，如果在某一数据区间上，两种不同标签的数据密度分布接近，则此区域的数据会影响分类结果，即为模糊区域。在对模糊区域初步观察后，将每个特征的连续数据离散化，用整数代表一个个划分出来的区间，计算同一特征的不同标签（0，1）下对应的整数出现密度，然后用两者中的较小者和较大者做比（0<比值<1），如果比值超过某个规定的阈值，则此整数对应的区域为模糊区域。

对于任务三，前10个混合物的成分判定，可以用测试数据训练sklearn科学计算库中的分类模型，预测前10条测试数据标签，具体做法为，对比使用主成分分析PCA降维后的数据训练模型,和使用原始数据训练模型的效果,决定是否采用PCA降维技术对特征进行降维，用不同分类器,在对数据标准化后，进行模型训练,即模型的参数择优以及参数评估(检验模型的泛化能力)，用多个分类器对测试数据进行预测,对比结果,综合后得到测试集前10条数据的标签(是否含有特定成分(1为含有，0为不含))。

运用以上方法，很好的解决了给出的任务，本文会对任务流程进行详细介绍。

关键词：混合物成分检测；数据分析；决策树；SVM; KNN;

混合物中特定成分的检测

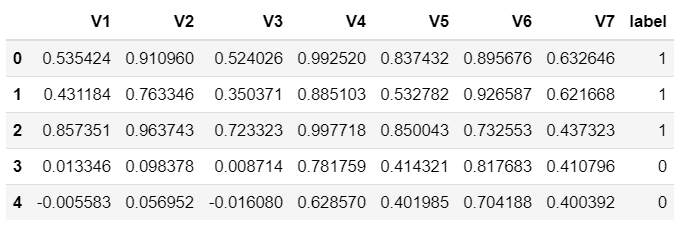
**引言：**

混合物中特定成分的检测，是典型的二分类问题，特征数据的不同，往往决定着标签的类别，主要指标的判定则和不同特征与标签的相关程度有关，而数据中，不同标签交叉的地方，往往就是难以判定类别的模糊区域，本文运用了数据探索性分析，机器学习中决策树、K近邻算法、支持向量机等方法，对以上问题进行了分析和解决，并得出最终结论。

1.数据获取及分析

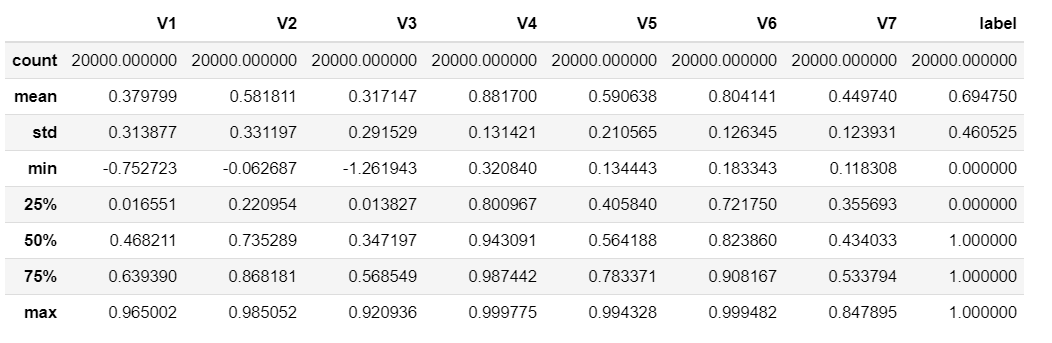
利用python工具库pandas从Data文件读取训练数据

①训练数据前5行展示



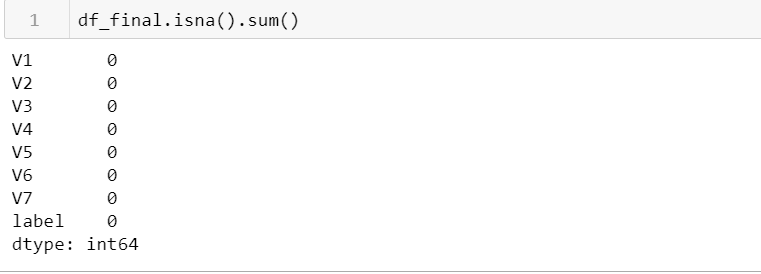
其中label列为：是否含有特定成分(1为含有，0为不含)

②查看数据描述



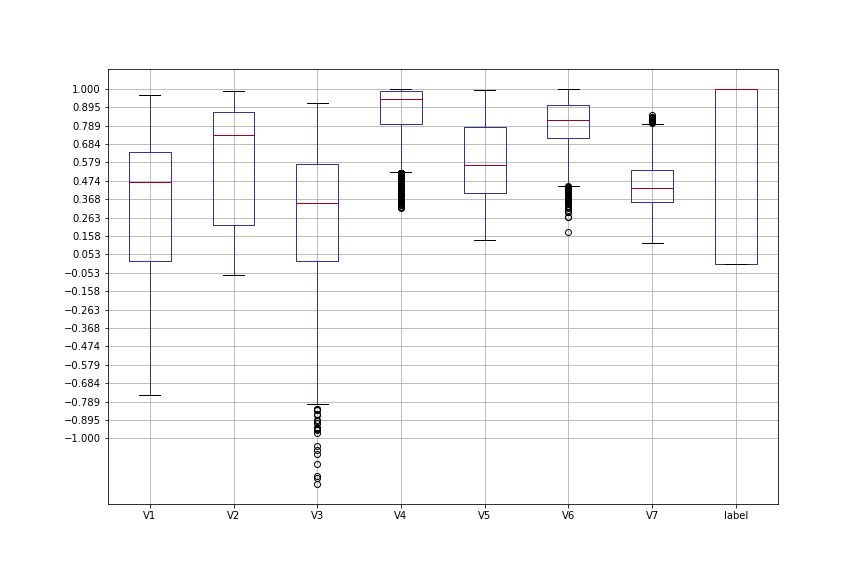
从label列的均值（mean）=0.69可以看出，样本中标签为1的占到70%，即含有特定成分的样本接近有14000条，不含特定成分的样本约为6000条，从样本的比例，可以推测混合物中大部分含有特定成分。还可以观察到特征数据几乎都分布在（-1，1）的区间，较为规范。

③查看训练数据是否有空值和缺失值（对每一列特征的空值进行求和）



发现数据中空值和缺失值数量为0，即数据具有较高完整性。

④绘制箱形图查看异常值



箱形图包含了六个数据节点，会将一组数据按照从小到大的顺序排列，分别计算出数据的上边缘、上四分位数、中位数、下四分位数、下边缘、异常值。可以观察到，特征V3、V4、V6、V7都有不同程度的异常值，留待训练分类模型时处理。

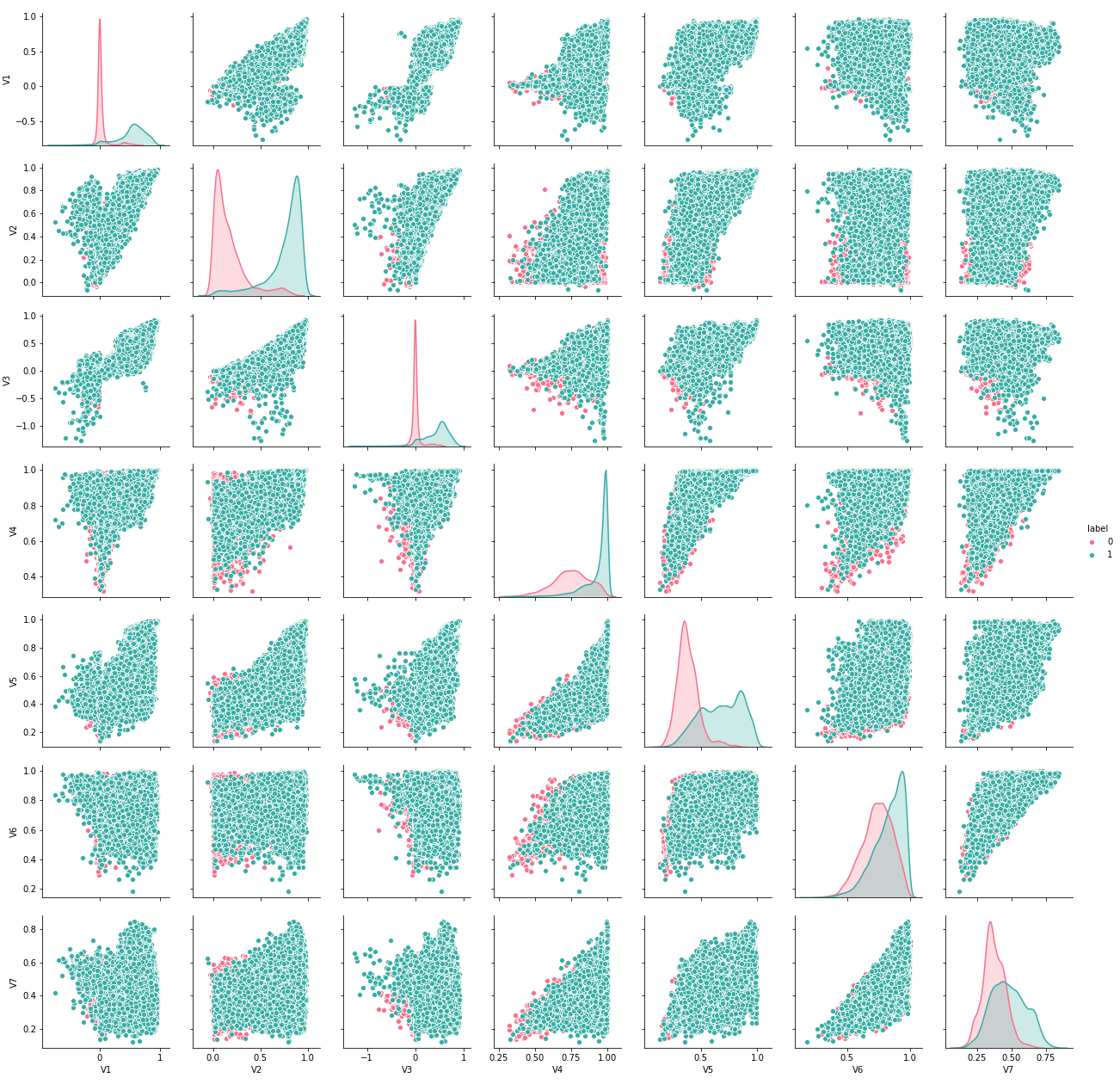
2. 判定特定成分存在的主要指标

思路：

①根据各个特征和判定标签(是否含有特定成分(1为含有，0为不含)）)的相关性，去判断特征的重要程度，选择出主要指标。

②用sklearn科学计算库特征选择(feature\_selection)库，对主要特征进行选择，验证判断。

（1）针对7项指标数据，绘制散点图矩阵（pairplot）可视化探索数据特征间的关系

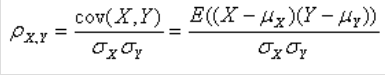


从图中可以看出，红点为label=0、绿点为label=1的数据，但由于数据量过大，变量间的散点分布不明显，但从单个变量自身角度来看，V1、V2、V3、V4、V5等指标的不同标签数据区分较明显，即对混合物是否含有特定成分具有较好的区分能力，而V6、V7指标区分度则偏低，初步推测V1、V2、V3、V4、V5为主要判定指标。

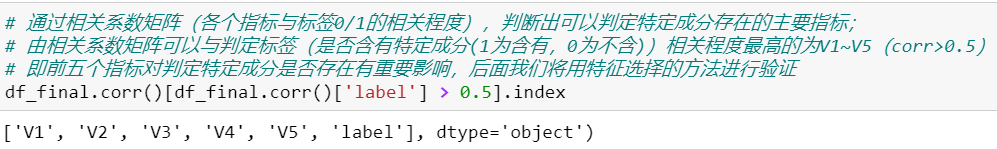
（2）计算相关系数矩阵，查看具体指标间的相关系数

相关系数用于检查两个变量之间变化趋势的方向以及程度，值范围-1到+1，0表示两个变量不相关，正值表示正相关，负值表示负相关，值越大相关性越强。

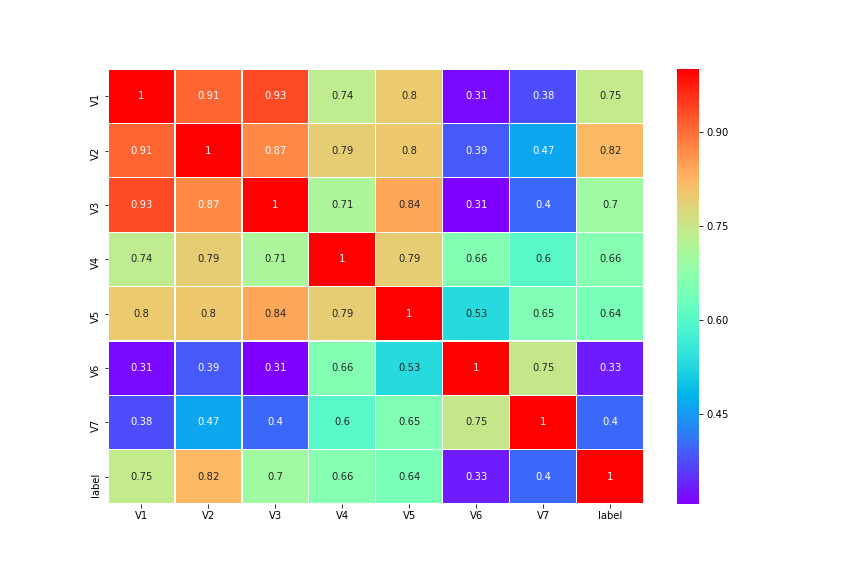
pearson correlation coefficient(皮尔森相关性系数）[1]计算公式：



则通过计算各个特征与label即判定标签间的相关关系，判断出哪些指标对特定成分的判定具有重要作用，pandas的corr()方法，提供了这一功能：



热度图绘制：

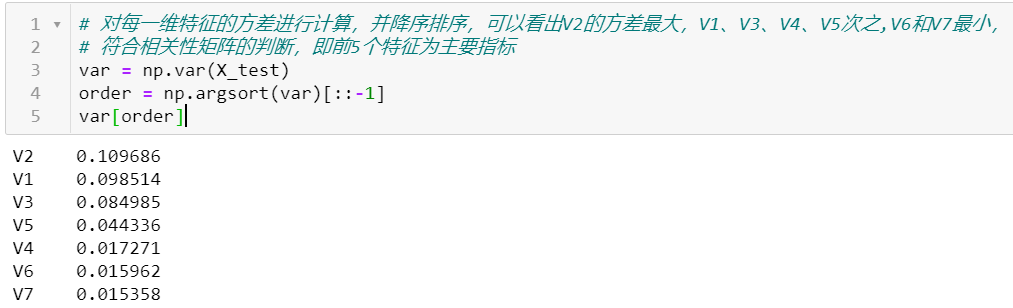


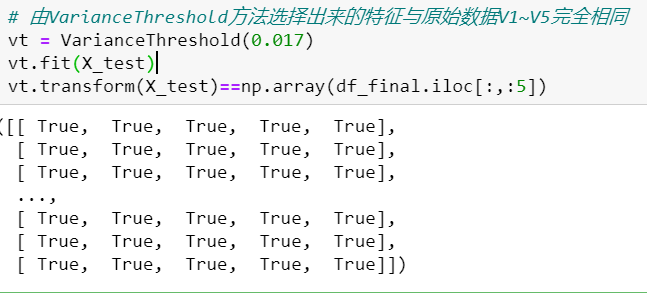
由相关系数矩阵可以与判定label（是否含有特定成分(1为含有，0为不含)）相关程度最高的为指标V1~V5（相关系数>0.5）,且V2与label的相关程度最高，而V6、V7指标的相关程度则不超过0.4，相关性很弱，与从散点图矩阵得到的结论一致，下面再用python科学计算库sklearn的feature\_selection库中的几种特征选择方法，验证特定成分存在的主要指标是否为V1~V5；

（3）特征选择（四种方法）

①移除低方差特征[2]：

VarianceThreshold 是特征选择的一个简单基本方法，它会移除所有那些方差不满足一些阈值的特征。默认情况下，它将会移除所有的零方差特征，即那些在所有的样本上的取值均不变的特征。

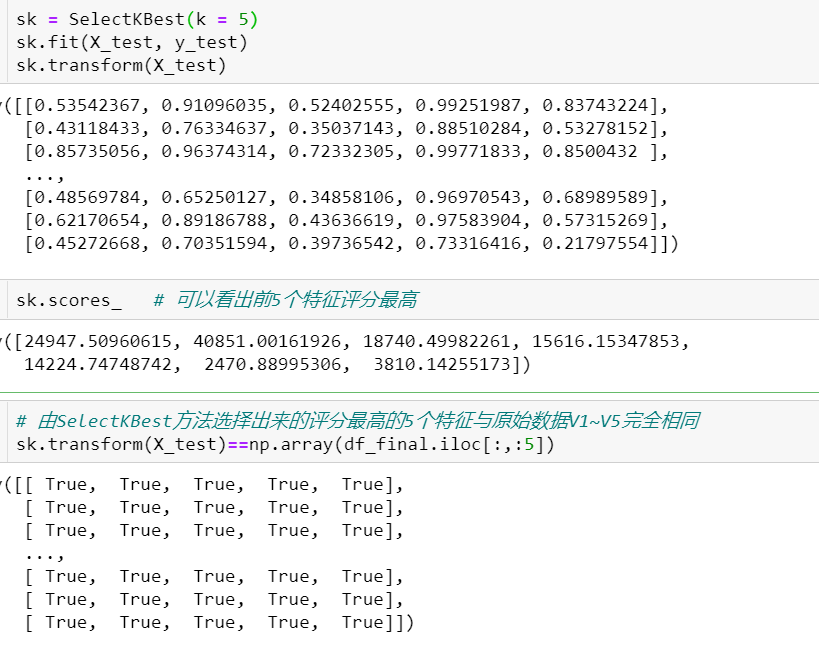




计算出指标V2的方差最大,V1、V3、V4、V5次之，且由VarianceThreshold方法选择出来的特征与原始数据V1~V5列数据完全相同，验证了V1~V5为主要指标。

②单变量的特征选择[2]：

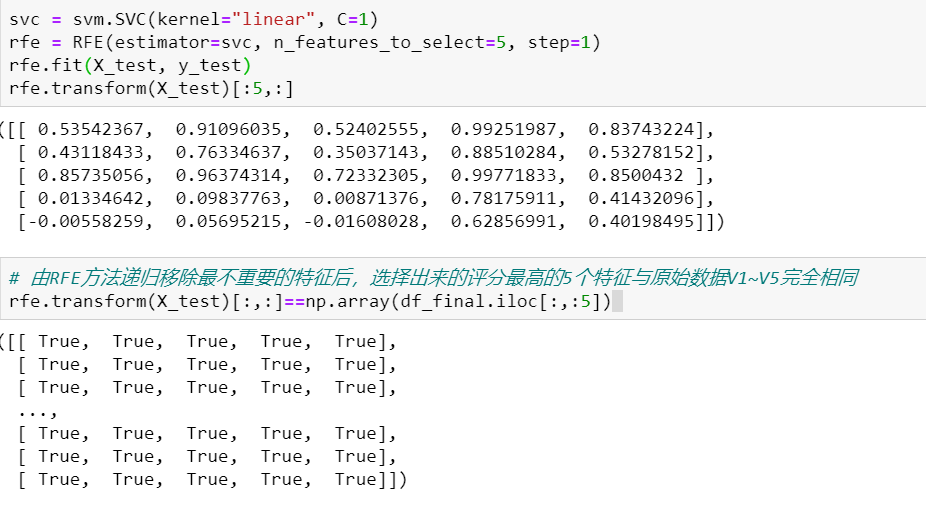
通过基于单变量的统计测试来选择最好的特征。它可以当做是评估器的预处理步骤。Scikit-learn 将特征选择的内容作为实现了 transform 方法的对，SelectKBest 移除那些除了评分最高的 K 个特征之外的所有特征。



由SelectKBest方法计算出来的特征评分，前V1~V5远大于V6和V7，且选择出来的评分最高的5个特征与原始数据V1~V5完全相同，验证了V1~V5为主要指标。

③递归式特征消除[2]：

给定一个外部的估计器，可以对特征赋予一定的权重（比如，线性模型的相关系数），recursive feature elimination( RFE )通过考虑越来越小的特征集合来递归的选择特征。首先，评估器在初始的特征集合上面训练并且每一个特征的重要程度是通过一个coef\_属性或者feature\_importances\_属性来获得。然后，从当前的特征集合中移除最不重要的特征。在特征集合上不断的重复递归这个步骤，直到最终达到所需要的特征数量为止。



验证验证了V1~V5为主要指标。

④基于Tree（树）的特征选取[2]：

基于树的estimators（sklearn.tree模块和树的森林在sklearn.ensemble模块）可以用来计算特征的重要性，然后可以消除不相关的特征（当与sklearn.feature\_selection.SelectFromModel等元转换器一同使用时）



特征重要程度最高的仍然是前5个特征远大于V6、V7，其中V2特征最为重要。

（4）结果

综上:使用四种feature\_selection的方法后，得到的结果都验证了特征V1、V2、V3、V4、V5为判定特定成分存在的主要指标！且V2的重要程度最高，V1、V2、V3间有较高的相关性。

3. 模糊区域估计

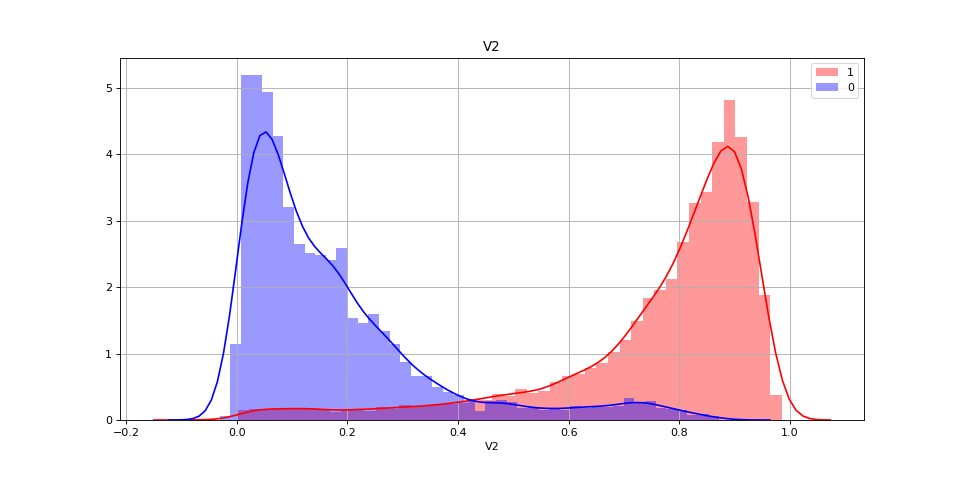
思路：

①分别画出label=0 和 label=1对应的单个特征变量的dist(密度直方图)和kde(核密度估计图-平滑后的密度直方图)），查看两种标签下数据分布情况，如果在某一数据区间上，两种不同标签的数据密度分布接近，则此区域的数据会影响分类结果，即为模糊区域。

②对模糊区域初步观察后，将每个特征的连续数据离散化，用整数代表一个个划分出来的区间，计算同一特征的不同标签（0，1）对应的整数出现密度，然后用两者中的较小者和较大者做比（0<比值<1），如果比值超过某个规定的阈值，则此整数对应的区域为模糊区域。

③计算出模糊区域后，将模糊区域可视化，即绘制到密度直方图中。

（1）**首先绘制出一个指标（如V2）的密度直方图，分析模糊区域**



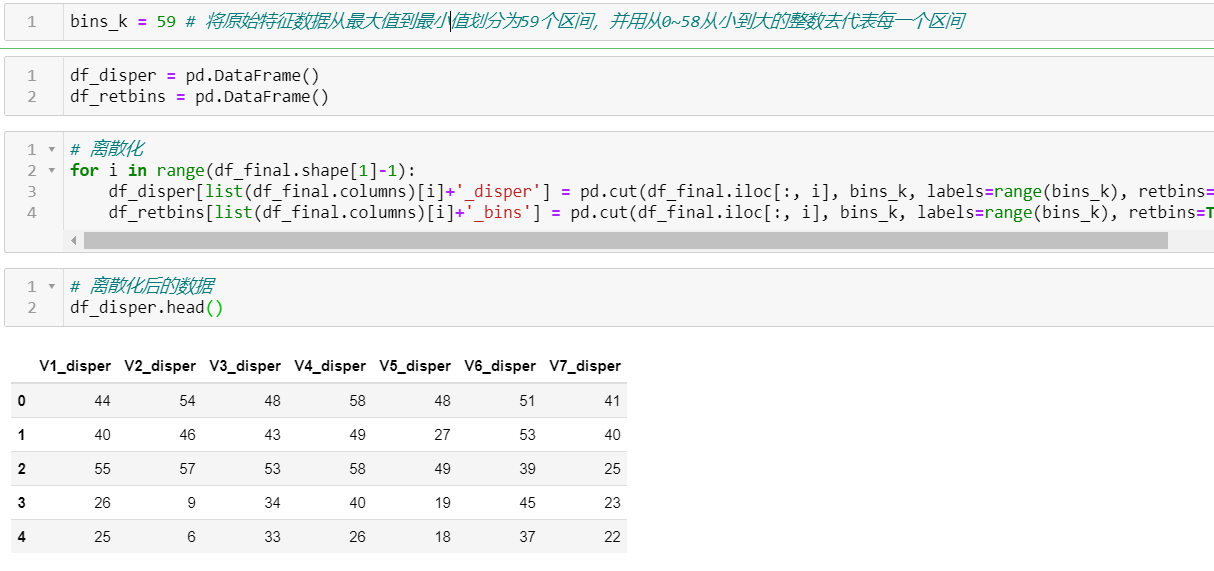
直方图横坐标为数据取值，被划分为一个个区间，纵坐标为数据密度，即该区间对应数据的数量在对应标签下数据总量的占比。

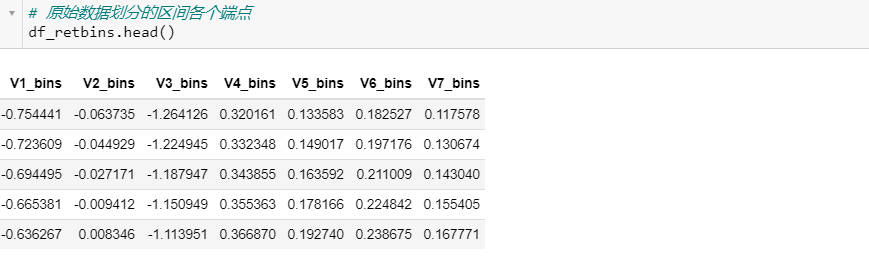
由直方图可以看出，在数据取值为0.4附近，两种标签（是否含有特定成分(1为含有，0为不含)）的对应的数据密度接近，即在这一区间附近，根据数据很难判断成分是否存在，故可认为这一区间为模糊区域。

（2）给出模糊区域的计算

1）、连续值离散化，计算不同标签（0，1）的数据密度比值

将每一个原始特征数据从最大值到最小值划分为59个区间，并用从0~58从小到大的整数去代表每一个区间。





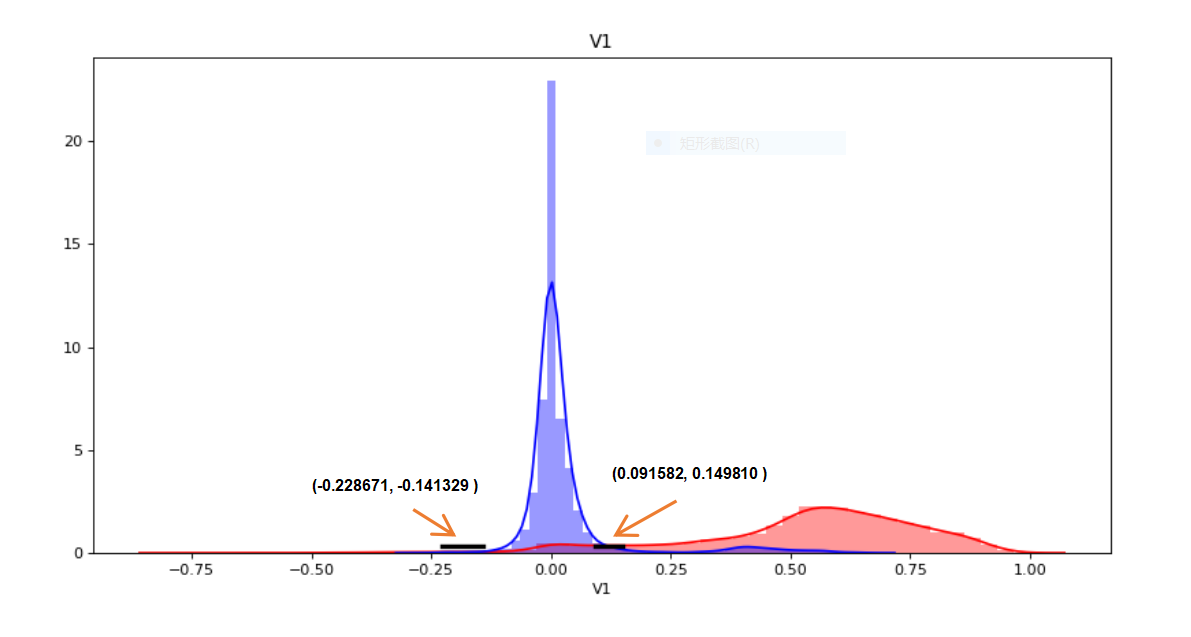
2）、计算同一特征的不同标签（0，1）对应的整数出现密度，然后用两者中的较小者和较大者做比（0<比值<1），如果比值超过某个规定的阈值(这里我们设置为0.5，因为超过0.5以后，对分类器的影响就比较大了)则此整数对应的区域为模糊区域。

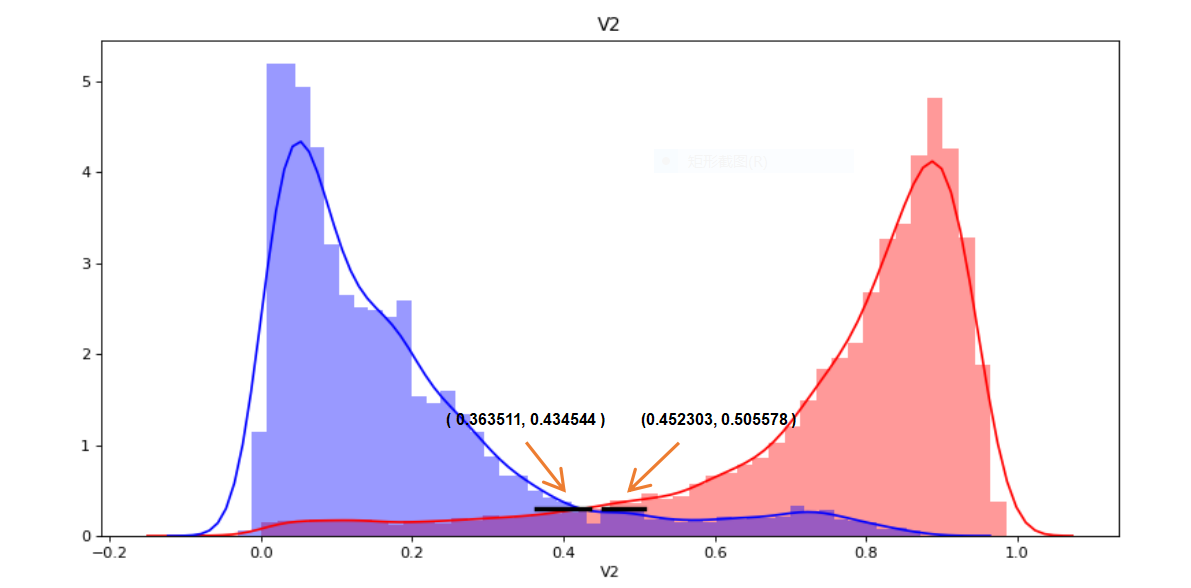
（计算结果（列表中每两个数代表一个模糊区间）：

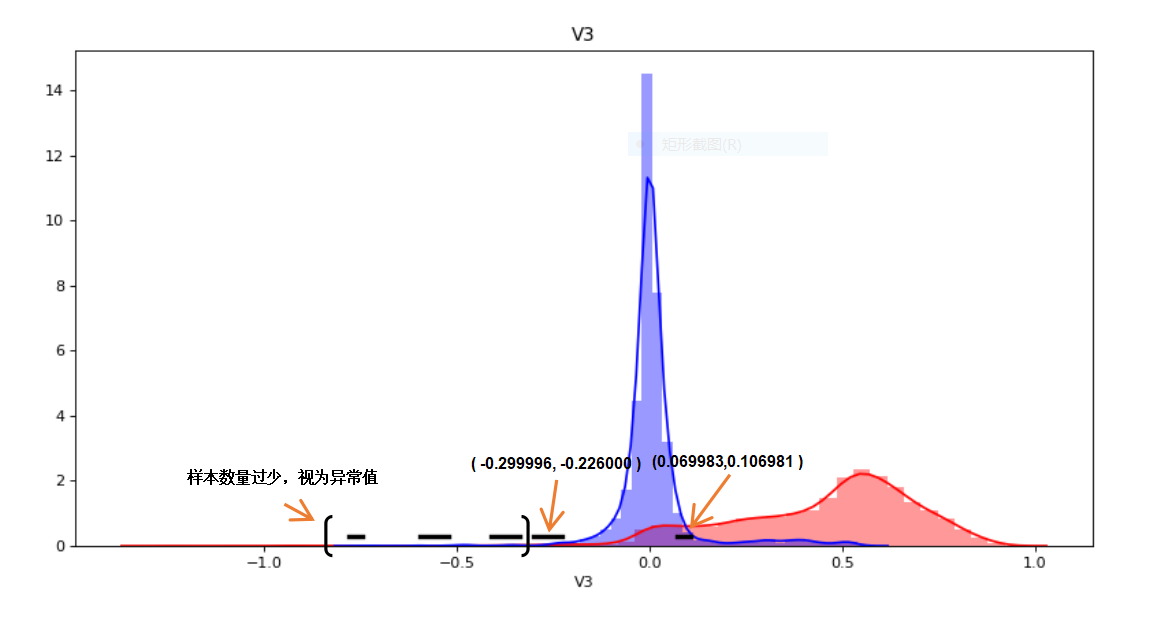


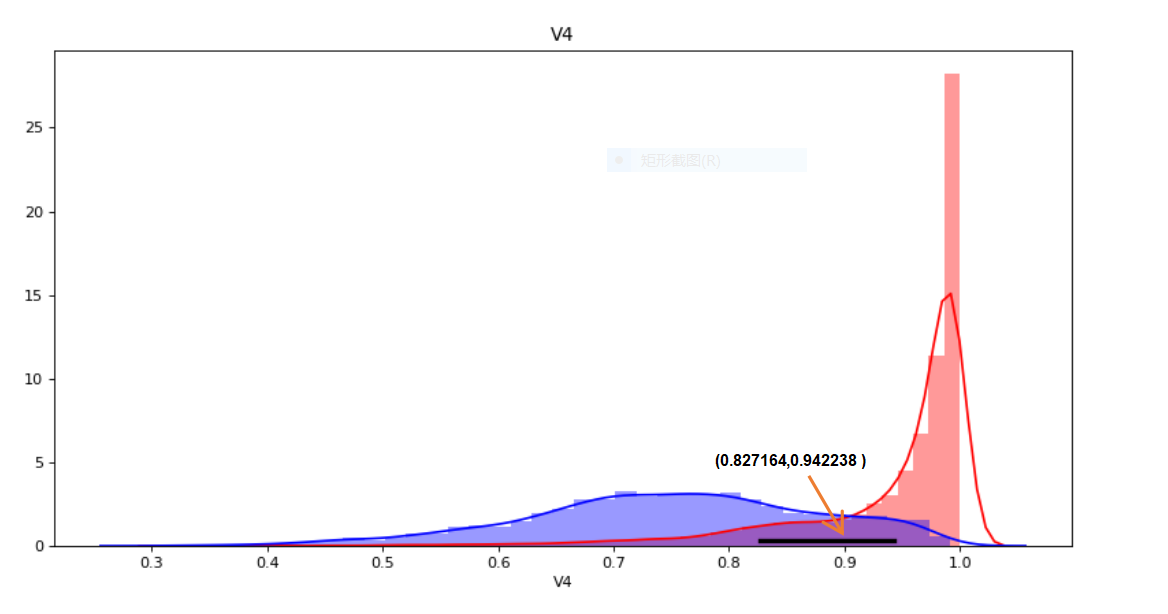
3）、计算出模糊区域后，将模糊区域可视化，即绘制到密度直方图中：

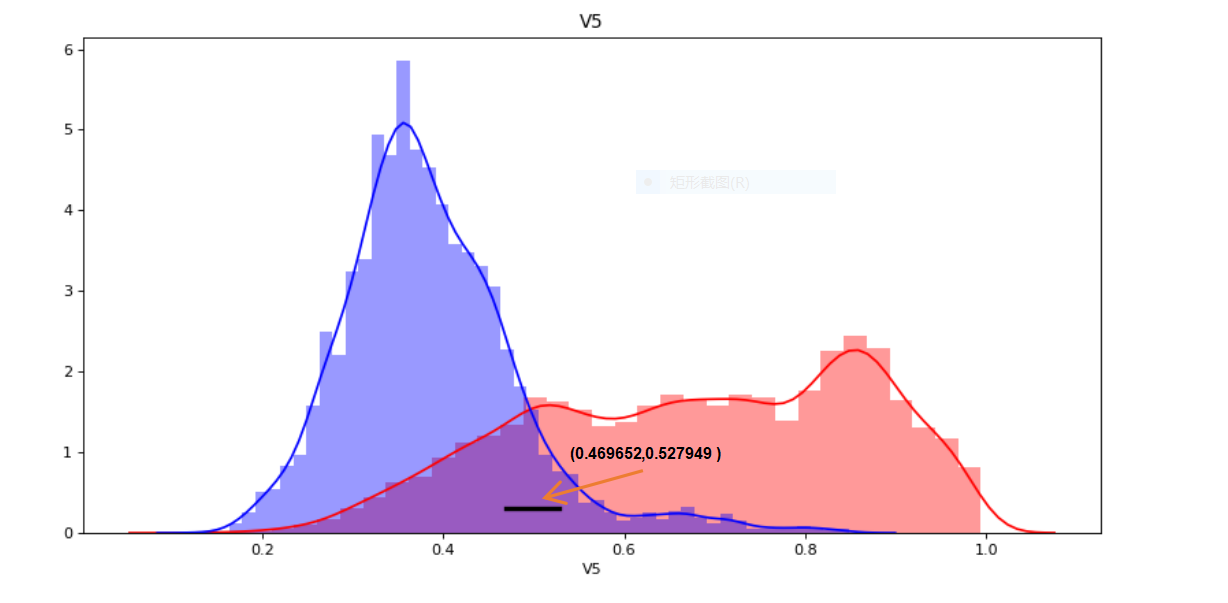
其中加粗黑线画出来的横坐标的区域为模糊区域。

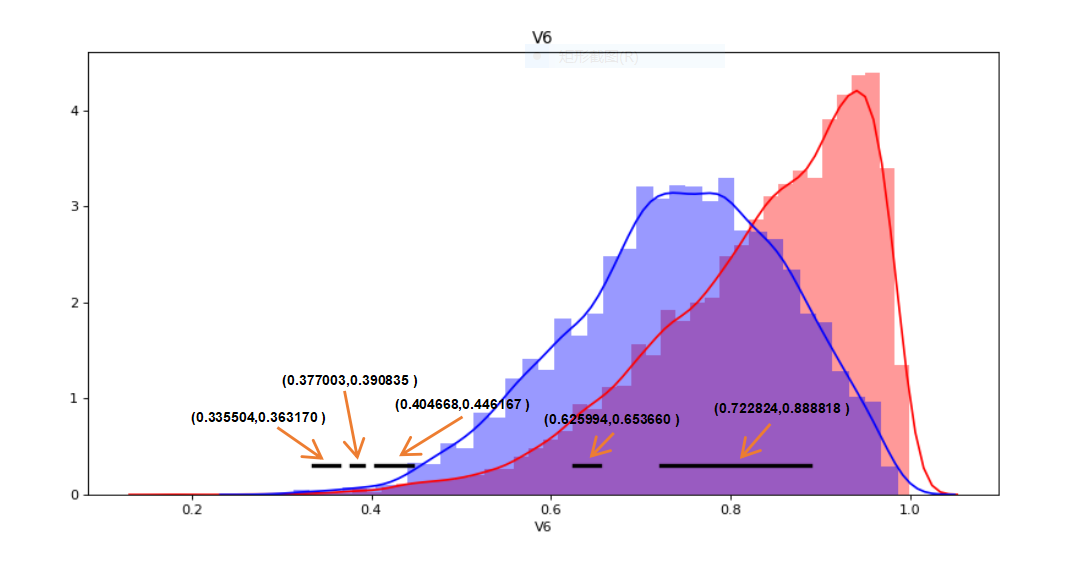


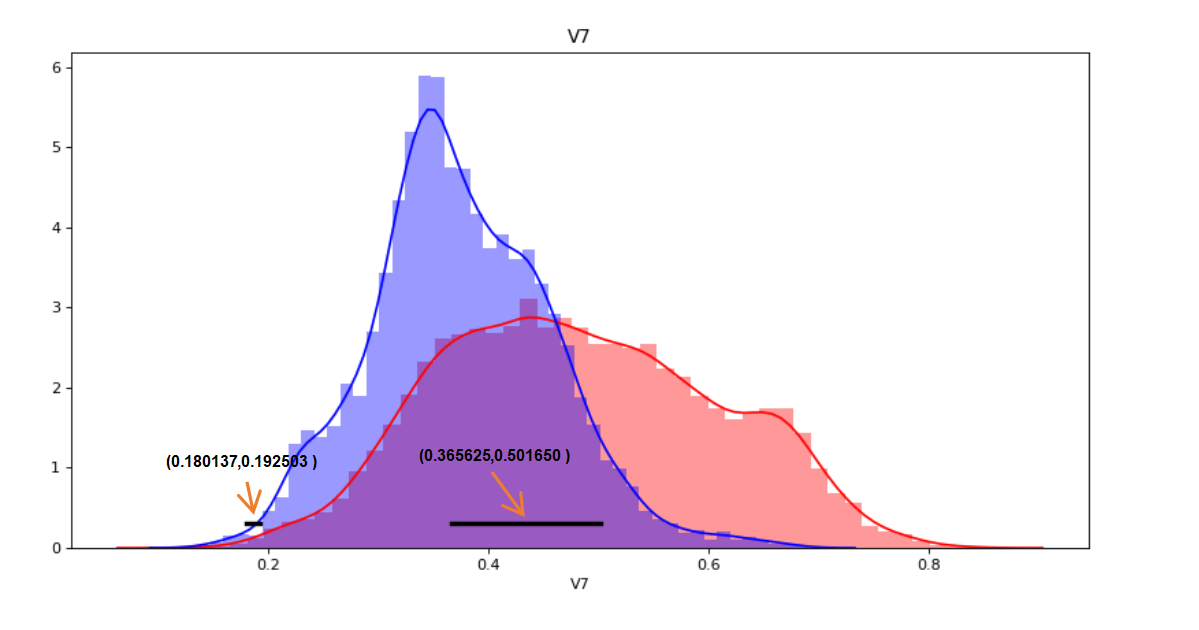












（3）结果

综上：过滤掉异常值，各个指标对应的模糊区域分别为：

V1：（-0.228671, -0.141329）,（0.091582, 0.149810）

V2：（0.363511, 0.434544）,（0.452303, 0.505578）

V3:（-0.299996, -0.226000）,（0.069983, 0.106981）

V4:（0.827164, 0.942238）

V5:（0.469652, 0.527949）

V6: （0.335504, 0.363170）, （0.377003, 0.390835）, （0.404668, 0.446167）, （0.625994, 0.653660,（0.722824, 0.888818）

V7:（0.180137, 0.192503）,（0.365625, 0.501650）

**4.** **前10个混合物的成分判定**

思路:

训练模型，预测前10条测试数据标签

①PCA降维: 由于特征矩阵过大，可能导致计算量大，训练时间长的问题，因此考虑是否需要降低特征矩阵维度

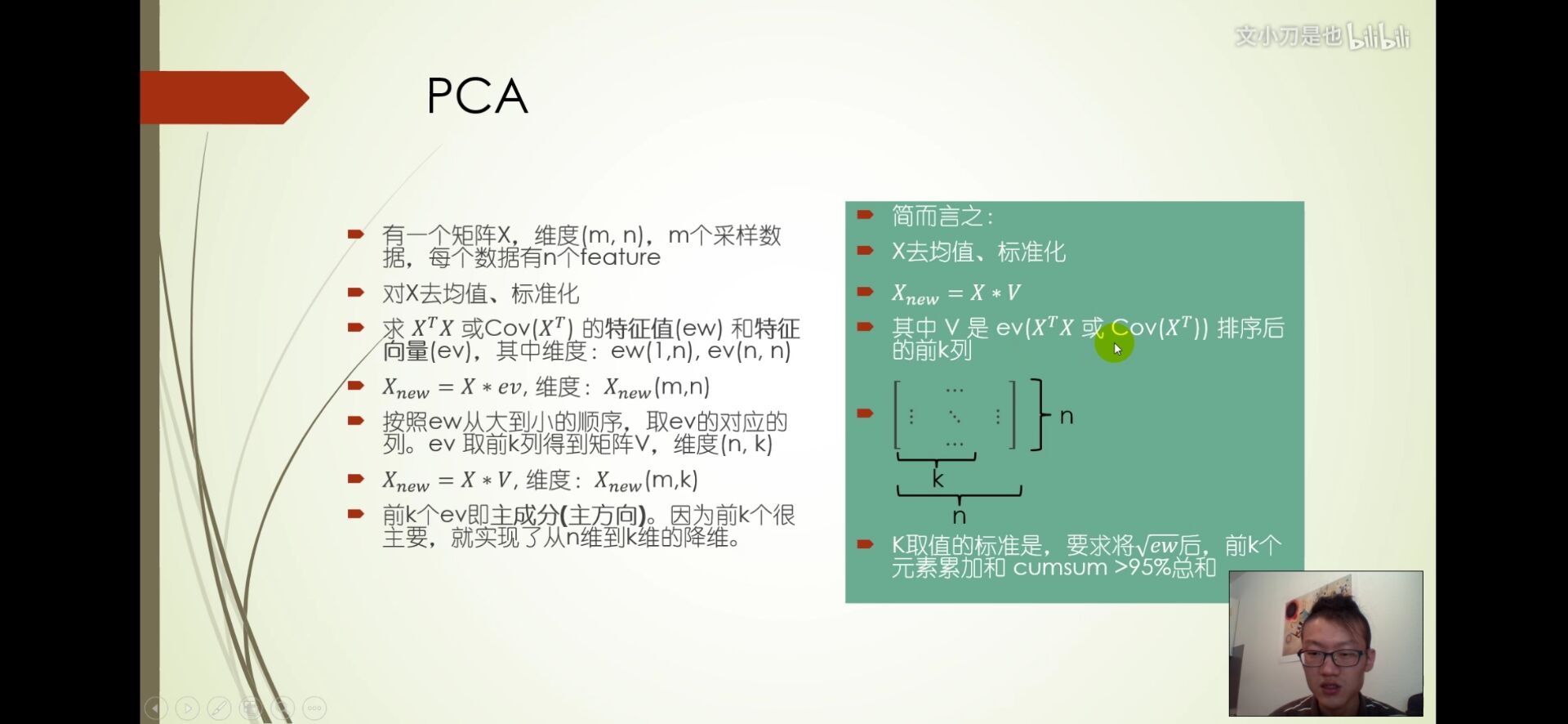
②对比使用主成分分析PCA降维后的数据训练模型,和使用原始数据训练模型的效果,决定是否采用PCA降维

③用不同分类器,在对数据标准化后，进行模型训练,即模型的参数择优以及参数评估(检验模型的泛化能力)

④用多个分类器对测试数据进行预测,对比结果,综合后得到测试集前10条数据得标签(是否含有特定成分(1为含有，0为不含))

（1）分析是否需要用到PCA降维

PCA实现原理：



作用：降维、去噪、提速、加深对特征的理解

1）、用PCA降维后的数据训练模型，预测准确率：



2）、用原始测试数据训练模型，预测准确率：



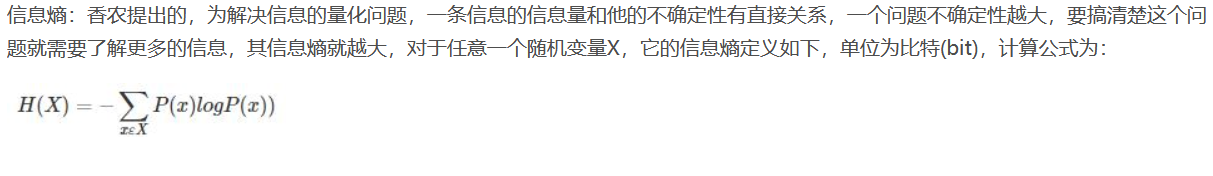
可见，在同一模型上，PCA降维的数据表现不如原始数据，且原始数据只有7维，降维不能带来较好的速度提升，故不采用PCA降维。、

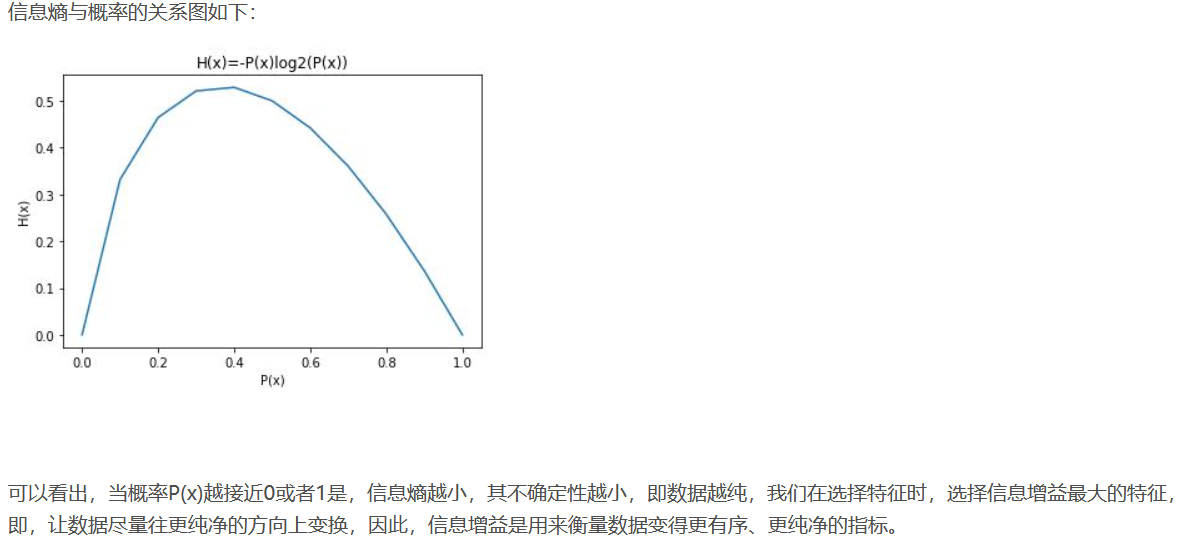
（2）模型训练

①决策树（DecisionTree）分类器

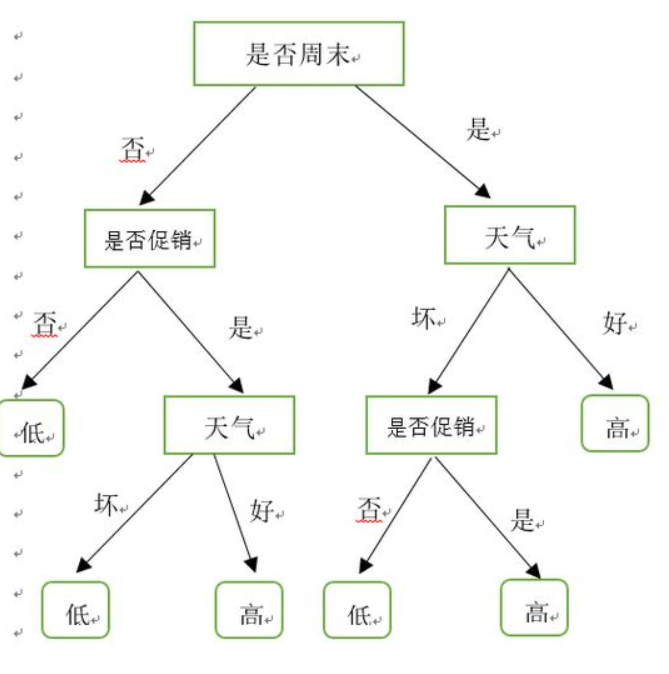
决策树实现原理[3]：

该算法是以信息论为基础，以信息熵和信息增益为衡量标准，从而实现对数据的归纳分类。根据信息论理论，采用划分后样本集的不确定性作为衡量划分好坏的标准，用信息增益度量不确定性：信息增益越大，不确定性越小。





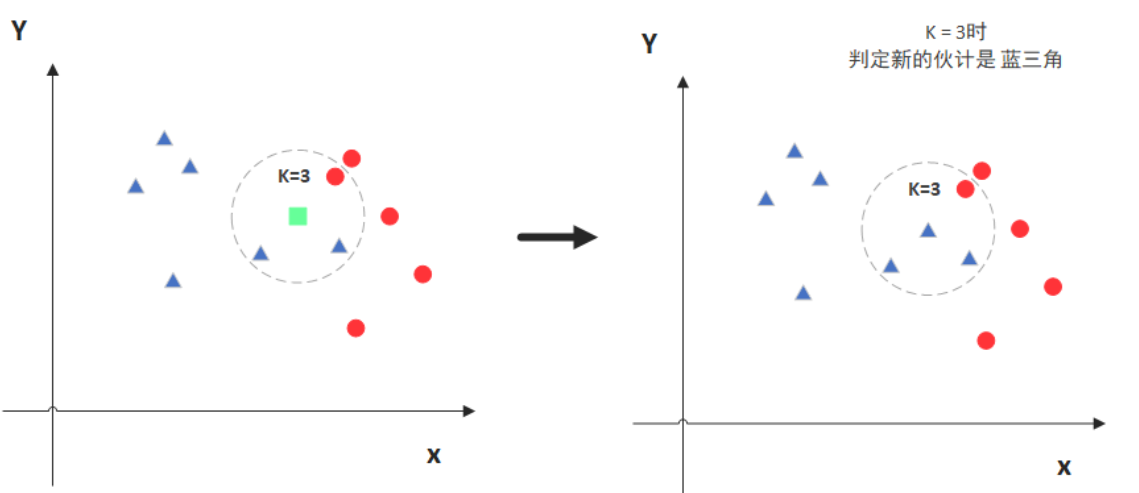
可视化分类过程：

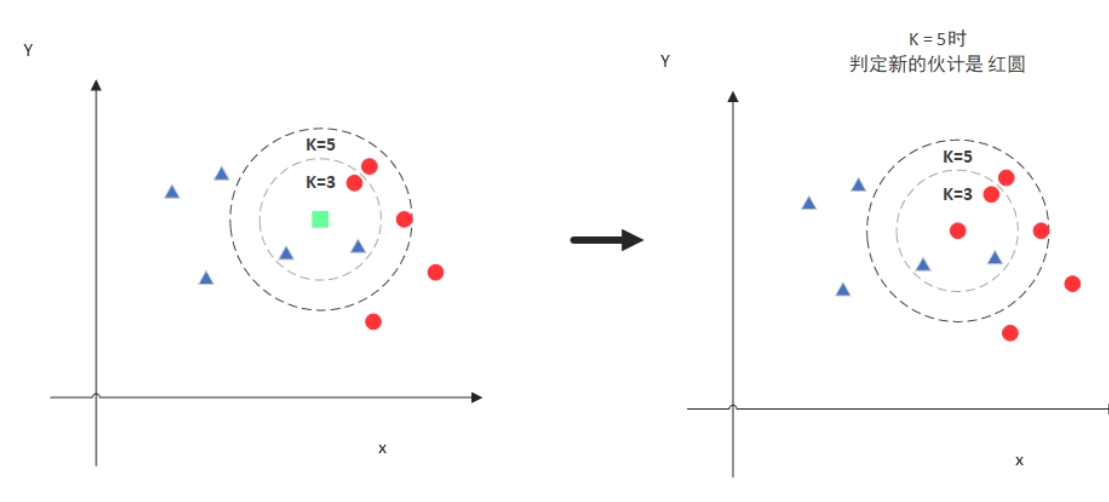


②K近邻算法（KNN）分类器

K近邻实现原理[4]：

KNN的全称是K Nearest Neighbors，意思是K个最近的邻居， KNN的原理就是当预测一个新的值x的时候，根据它距离最近的K个点是什么类别来判断x属于哪个类别。





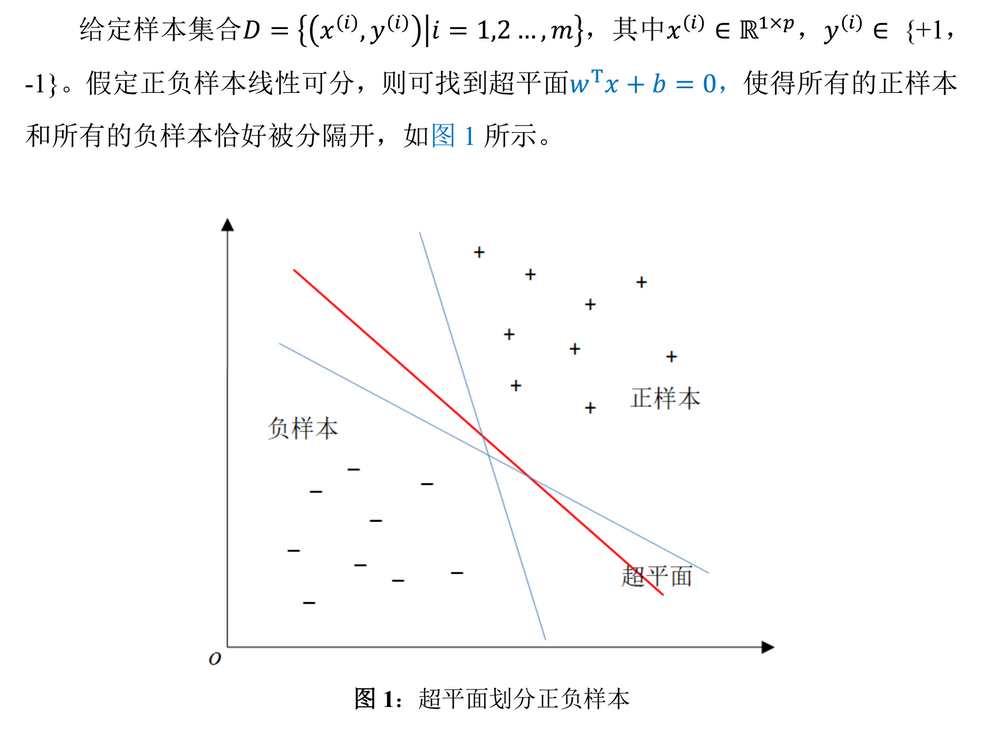
要度量空间中点距离的话，有好几种度量方式，比如常见的曼哈顿距离计算，欧式距离计算等等。不过通常KNN算法中使用的是欧式距离。

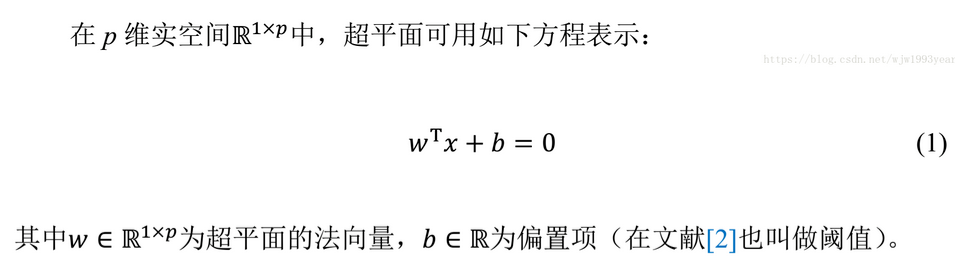
欧氏距离计算公式:

二维空间欧式距离

③支持向量机（SVM）分类器

SVM实现原理[5]：

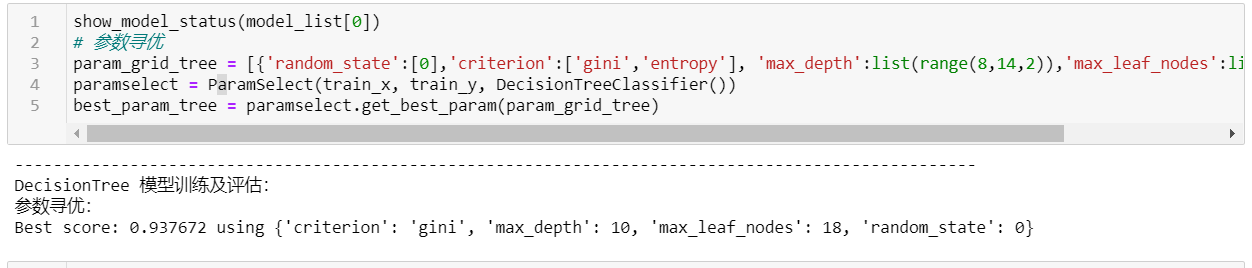




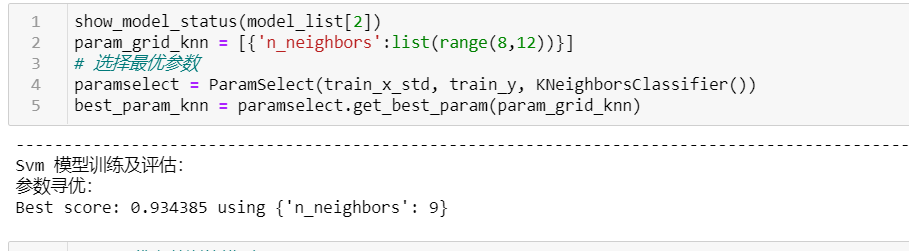
（3）参数择优

利用十折交叉验证选择模型的最优参数

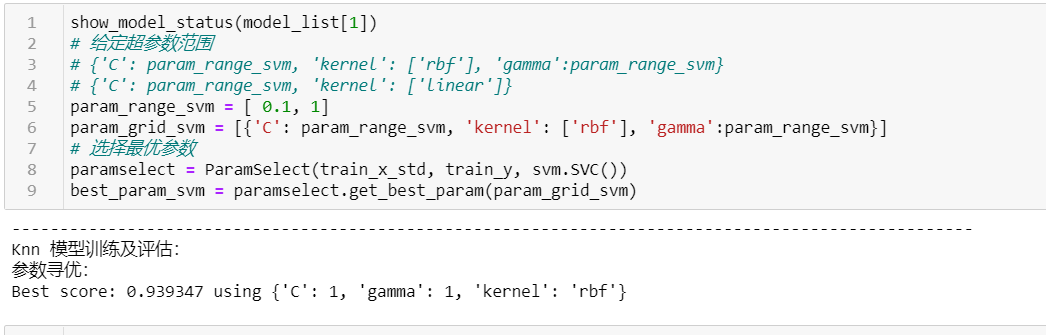
①决策树（DecisionTree）分类器



②K近邻算法（KNN）分类器



③支持向量机（SVM）分类器



（4）模型评估

1）、混淆矩阵[6]:

我们通过样本的采集，能够直接知道真实情况下，哪些数据结果是positive，哪些结果是negative。同时，我们通过用样本数据跑出分类型模型的结果，也可以知道模型认为这些数据哪些是positive，哪些是negative。

因此，我们就能得到这样四个基础指标，我称他们是一级指标（最底层的）：

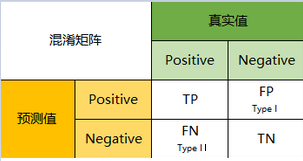
真实值是positive，模型认为是positive的数量（True Positive=TP）

真实值是positive，模型认为是negative的数量（False Negative=FN）：这就是统计学上的第二类错误（Type II Error）

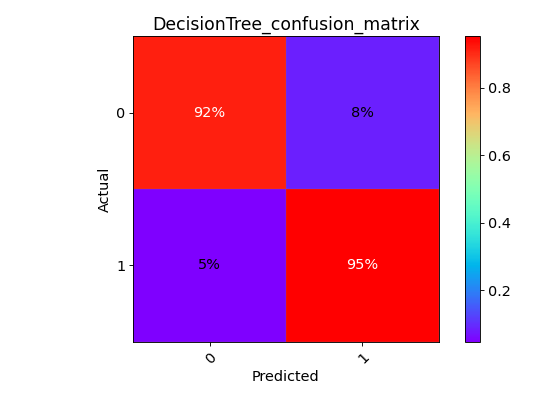
真实值是negative，模型认为是positive的数量（False Positive=FP）：这就是统计学上的第一类错误（Type I Error）

真实值是negative，模型认为是negative的数量（True Negative=TN）

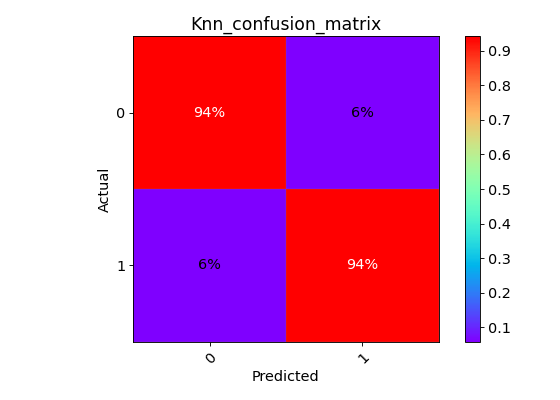
将这四个指标一起呈现在表格中，就能得到如下这样一个矩阵，我们称它为混淆矩阵（Confusion Matrix）



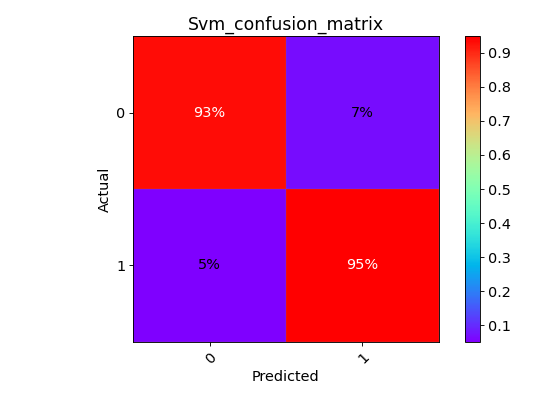
①决策树（DecisionTree）分类器



②K近邻算法（KNN）分类器



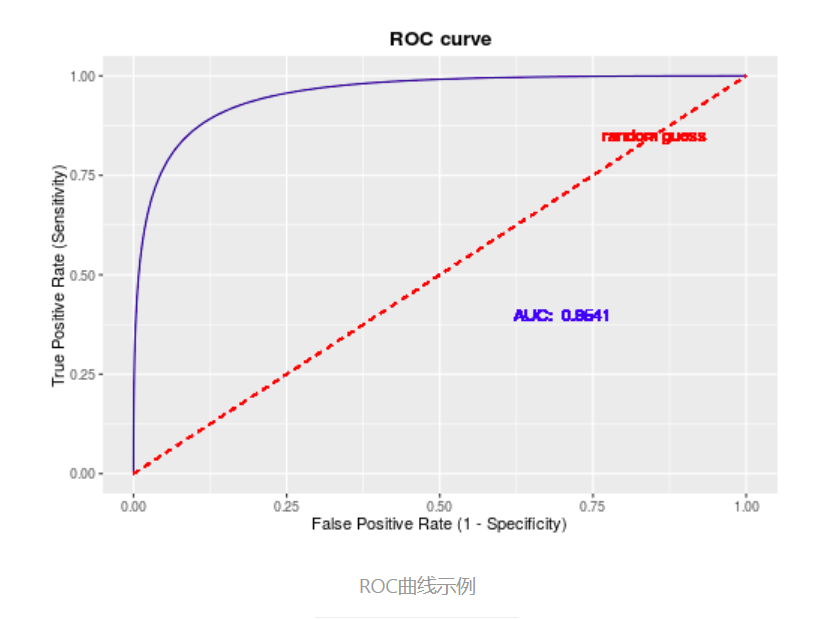
③支持向量机（SVM）分类器



可以看出三个模型的预测准确率都很高。

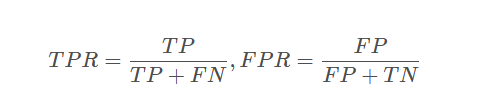
2）ROC曲线[7]：

ROC的全称是Receiver Operating Characteristic Curve，中文名字叫“受试者工作特征曲线”



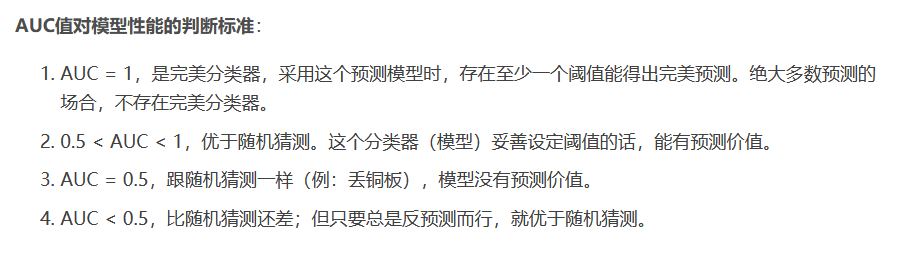
该曲线的横坐标为假阳性率（False Positive Rate, FPR），N是真实负样本的个数，FP是N个负样本中被分类器预测为正样本的个数。

纵坐标为真阳性率（True Positive Rate, TPR）P是真实正样本的个数，TP是P个正样本中被分类器预测为正样本的个数。

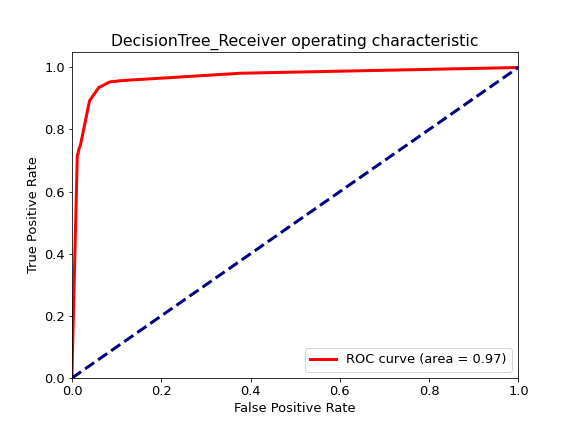


ROC曲线的应用场景有很多，根据上述的定义，其最直观的应用就是能反映模型在选取不同阈值的时候其敏感性（sensitivity, FPR）和其精确性（specificity, TPR）的趋势走向。

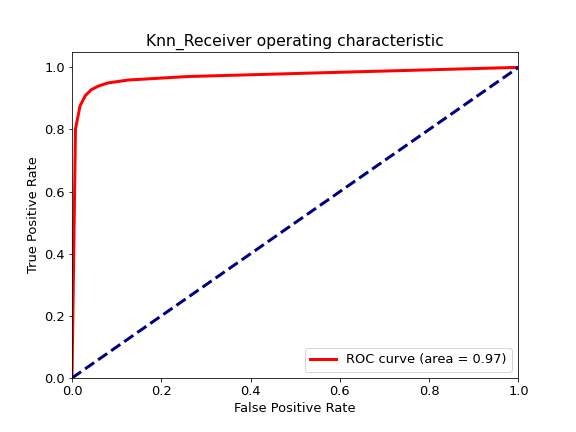
AUC表示ROC曲线下的面积，主要用于衡量模型的泛化性能，即分类效果的好坏。



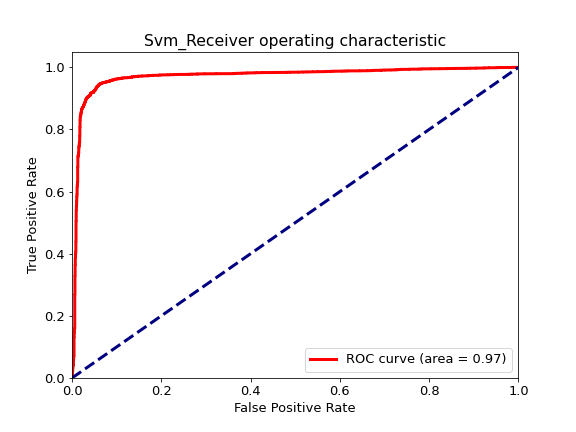
①决策树（DecisionTree）分类器



②K近邻算法（KNN）分类器



③支持向量机（SVM）分类器

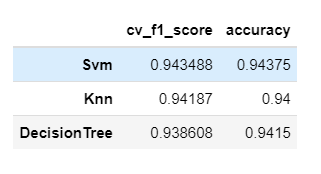


可以看出三个模型的AUC即ROC曲线的面积都达到了0.97以上，非常接近1，证明训练的模型都非常合格。

**（5）最终结果**

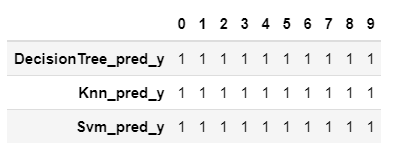
对三个分类器，用测试数据进行训练，并对它们进行参数择优，交叉验证后，得到他们的在训练数据上的预测得分和在测试数据集的前十条的预测结果。

预测得分：



SVM分类器的预测准确率最高，达到94.4%。

三种模型的对测试数据前10条的预测结果：



分类模型预测测试数据前10条的结果，三个分类器均为1。

**综上，**测试数据中前10个混合物中都含有特定成分。

5.任务结论

（1）任务1——判定特定成分存在的主要指标

V1、V2、V3、V4、V5为判定特定成分存在的主要指标！且V2的重要程度最高，V1、V2、V3间有较高的相关性。

（2）任务2——模糊区域估计

各个指标对应的模糊区域分别为：

V1：（-0.228671, -0.141329）,（0.091582, 0.149810）

V2：（0.363511, 0.434544）,（0.452303, 0.505578）

V3:（-0.299996, -0.226000）,（0.069983, 0.106981）

V4:（0.827164, 0.942238）

V5:（0.469652, 0.527949）

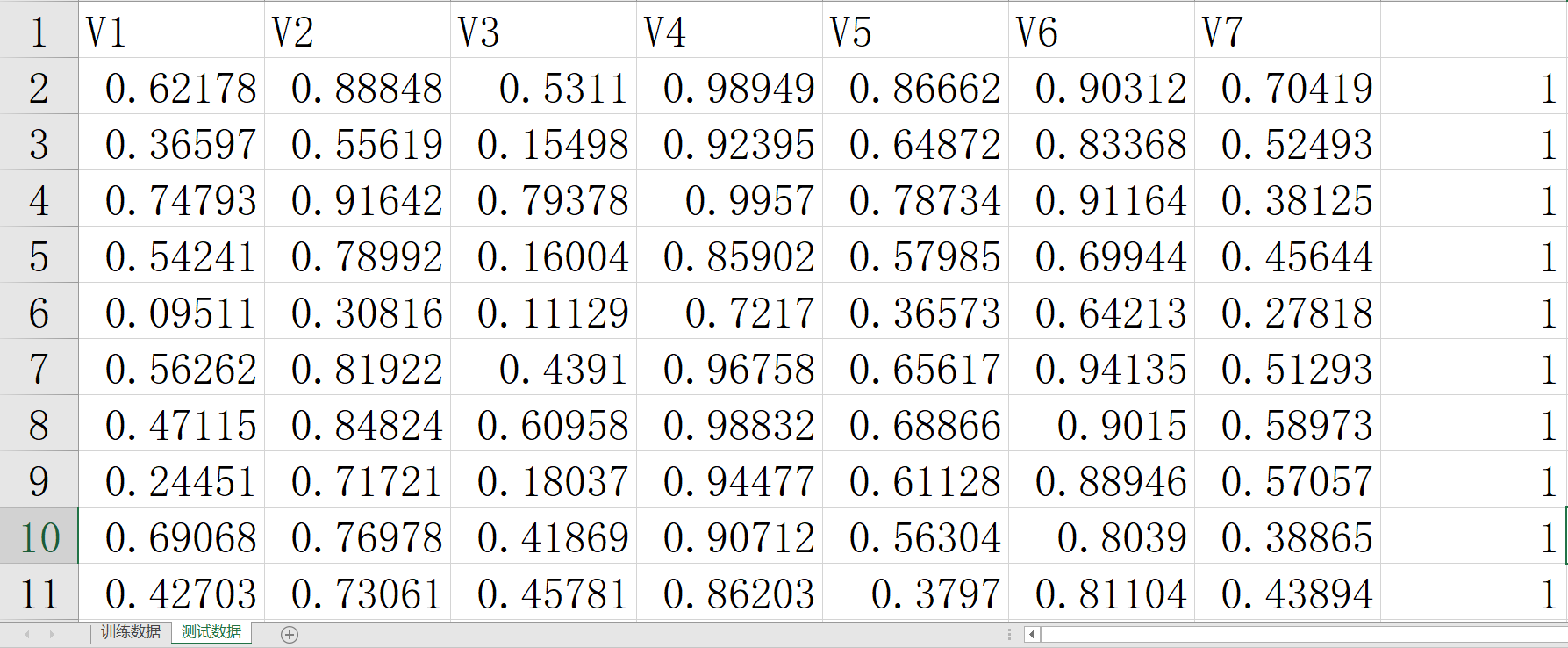
V6: （0.335504, 0.363170）, （0.377003, 0.390835）, （0.404668, 0.446167）, （0.625994, 0.653660,（0.722824, 0.888818）

V7:（0.180137, 0.192503）,（0.365625, 0.501650）

（3）任务3——前10个混合物的成分判定

前10个混合物中都含有特定成分，取值均为1。

即前十条测试数据对应的混合物是否含有特定成分的的判定结果为：



参考文献

[1]顶尖菜鸟.pandas——相关系数函数corr().<https://blog.csdn.net/qq_40946639/article/details/102984166.2020-07-19>.

[2]sklearn官方文档.Feature selection.<https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection>.2020-07-19

[3]zhuimengshaonian66.决策树的数学原理. https://blog.csdn.net/zhuimengshaonian66/article/details/81740318.2020-07-19

[4]zzzzMing.深入浅出KNN算法（一）KNN算法原理. https://www.cnblogs.com/listenfwind/p/10311496.html.2020-07-19.

[5]wjw1993year.SVM的数学原理详解. https://blog.csdn.net/wjw1993year/article/details/81166198.2020-07-19.

[6]Orange\_Spotty\_Cat.分类模型评判指标（一） - 混淆矩阵(Confusion Matrix).https://blog.csdn.net/Orange\_Spotty\_Cat/article/details/80520839.2020-07-19.

[7]蘑菇轰炸机.机器学习基础（1）- ROC曲线理解. https://www.jianshu.com/p/2ca96fce7e81.2020-07-19.