2020 年 兰 州 大 学 数 学 建 模 竞 赛

题目： 混合物中特定成分的检测

Hollow Man

成员姓名 蒋嵩林 肖锦恒 白舒睿

学　　院 信息科学与工程学院

班　　级 计算机科学与技术基地班(2018 级)

日　　期 二〇二〇年七月

摘　要：

我们使用 PCA（主成分分析）算法，利用 Python 的 sklearn 库，设定主成分的方差和所占的最小比例阈值 0.95，通过程序自动判定，将 7 维的判断成分主要指标通过 PCA 降维到 3 个维度，然后将给出判定结果的“训练数据”划分为 18000 个训练集和 2000 个测试集。随后使用 17 种分类机器学习模型进行训练，选取准确率最高——94.25%的模型即随机森林分类模型，在测试集进行测试，并给出了模型对“测试数据”的预测结果。

关键词：

主成分分析算法，随机森林，机器学习分类算法

1 问题重述

1.1 问题背景

给定 25000 个混合物数据样本，并且其中 20000 个混合物数据样本给出了混合物是否包含特定成分,所有混合物数据样本都给出了 7 项指标(记为 V1,V2,…,V7)的检测值。

1.2 问题概述

问题 1 要求给出判断特定成分存在的主要指标，问题 2 则要求给出判定是否存在对于 7 项指标的模糊区域。问题 3 要求建模并给出“测试数据”前 10 个混合物的判定。

2 模型假设

为了简化模型，因而设定 PCA 主成分的方差和所占的最小比例阈值 0.95。

附件解释：

MModeling.ipynb 问题 1 中使用 PCA 将 7 维降维到 2 维的代码。

MModeling2.ipynb 问题 1 中使用 PCA 将 7 维降维到 3 维的代码。

MModeling3.ipynb 问题 3 中使用 PCA 将 7 维降维到 3 维，随后使用 17 种分类机器学习方法进行训练的代码。

Train.csv 问题提供的附件 Data.xlsx 中“训练数据”前 18000 行

Test.csv 问题提供的附件 Data.xlsx 中“训练数据”前 2000 行

Predict.csv 问题提供的附件 Data.xlsx 中“测试数据”

Data.csv 问题提供的附件 Data.xlsx 中所有 25000 个数据中的所有 7 项指标

测试数据预测结果.csv 问题提供的附件 Data.xlsx 中“测试数据”及使用随机森林方法训练得到的预测结果

result.txt 对问题提供的附件 Data.xlsx 中“测试数据”使用随机森林方法训练得到的预测结果

3　名词解释

3.1　主成分分析（PCA）算法

在多元统计分析中，主成分分析（PCA）是一种统计分析、简化数据集的方法。它利用正交变换来对一系列可能相关的变量的观测值进行线性变换，从而投影为一系列线性不相关变量的值，这些不相关变量称为主成分。具体地，主成分可以看做一个线性方程，其包含一系列线性系数来指示投影方向。PCA 对原始数据的正则化或预处理敏感（相对缩放）。

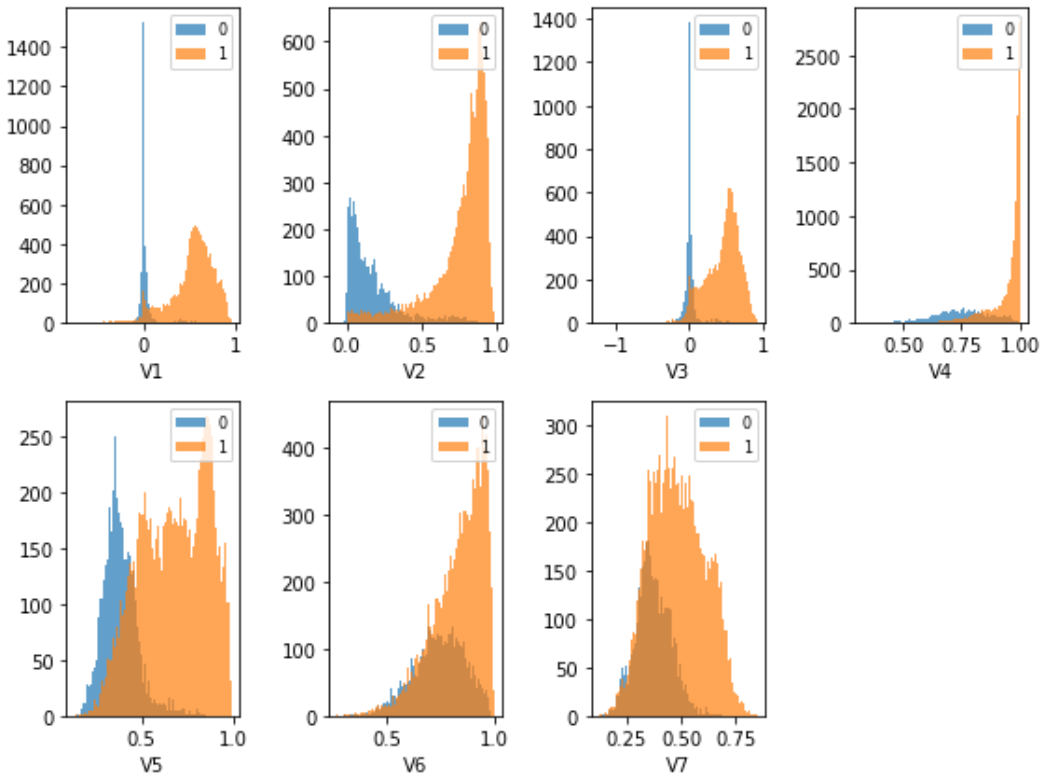
主成分分析经常用于减少数据集的维数，同时保留数据集当中对方差贡献最大的特征。这是通过保留低维主成分，忽略高维主成分做到的。这样低维成分往往能够保留住数据的最重要部分。[1]

4　问题求解

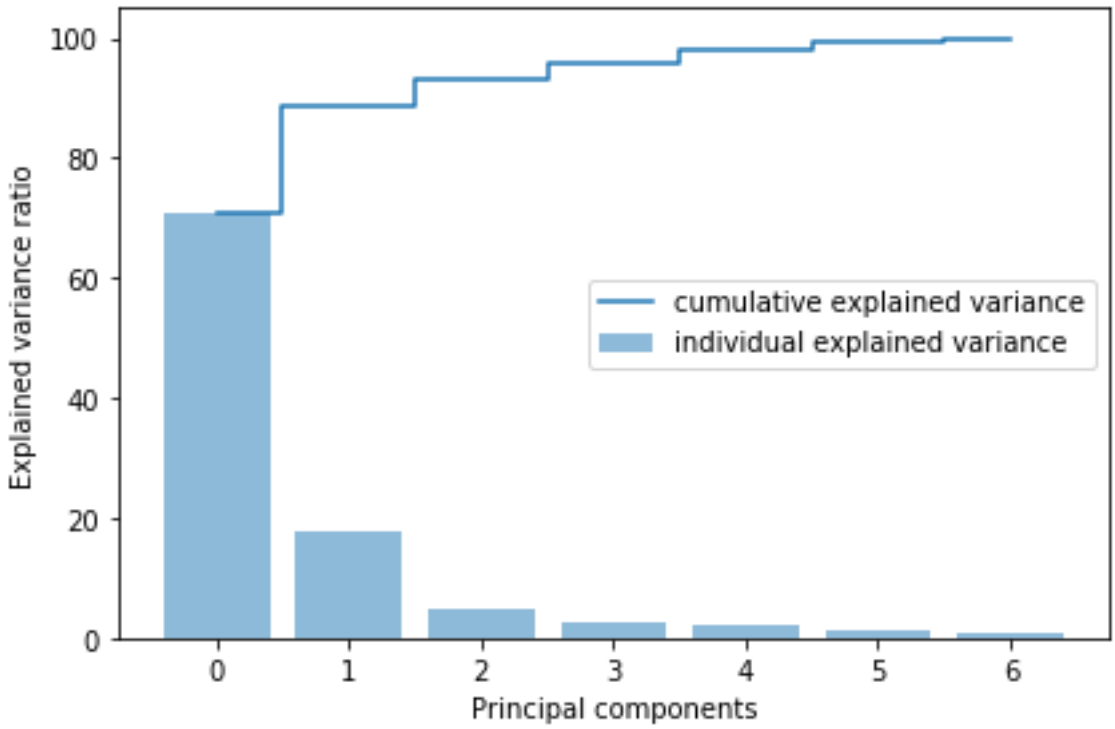
4.1　问题 1

首先将提供的 Excel 格式训练数据使用 Excel 导出到“Train.csv”文件，然后再使用 Anaconda 环境运行 Jupiter Notebook，在“Train.csv”文件存放同一目录下，之后我们的所有操作都在此环境下进行。

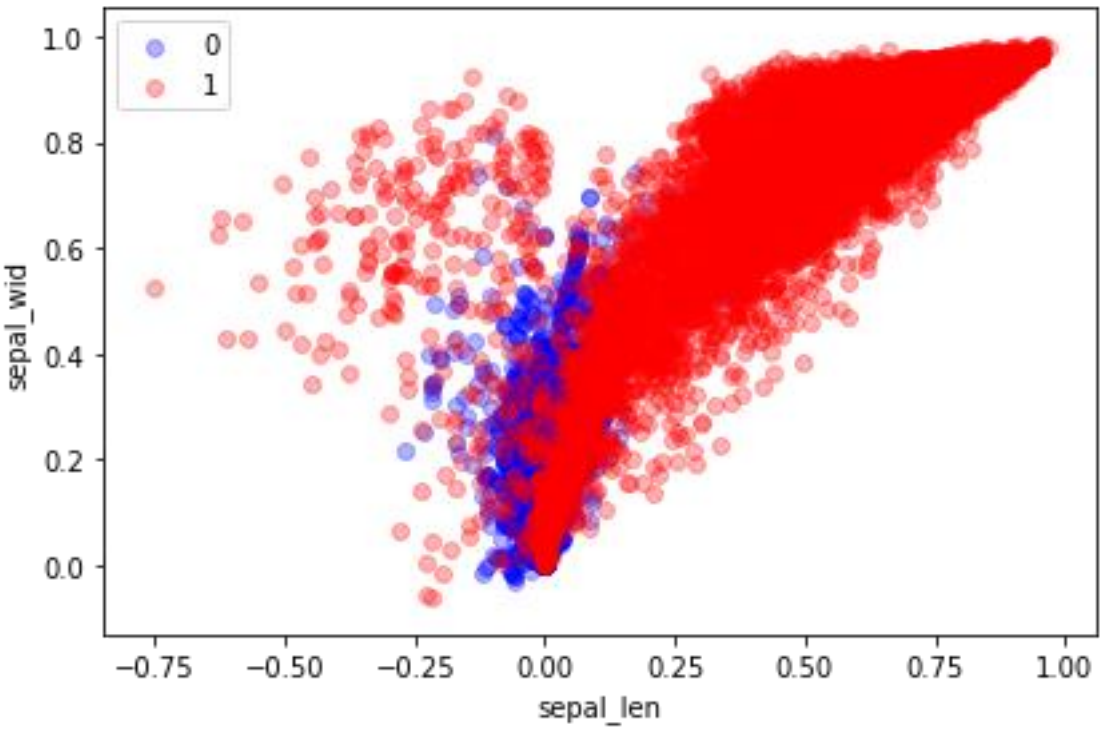
新建“MModeling.ipynb”在导入数据之后，首先我们对各个指标进行可视化分析，得出初步结论：



将数据进行标准化之后，随后得到按列计算的均值，按照公式代入得到协方差矩阵，然后转置矩阵，求得 PCA 降维的特征值和特征向量。最后求得各特征值占比，输出每一个值与前一个和值的加和，得到方差贡献率和累计贡献率，绘图，得到以下图片：

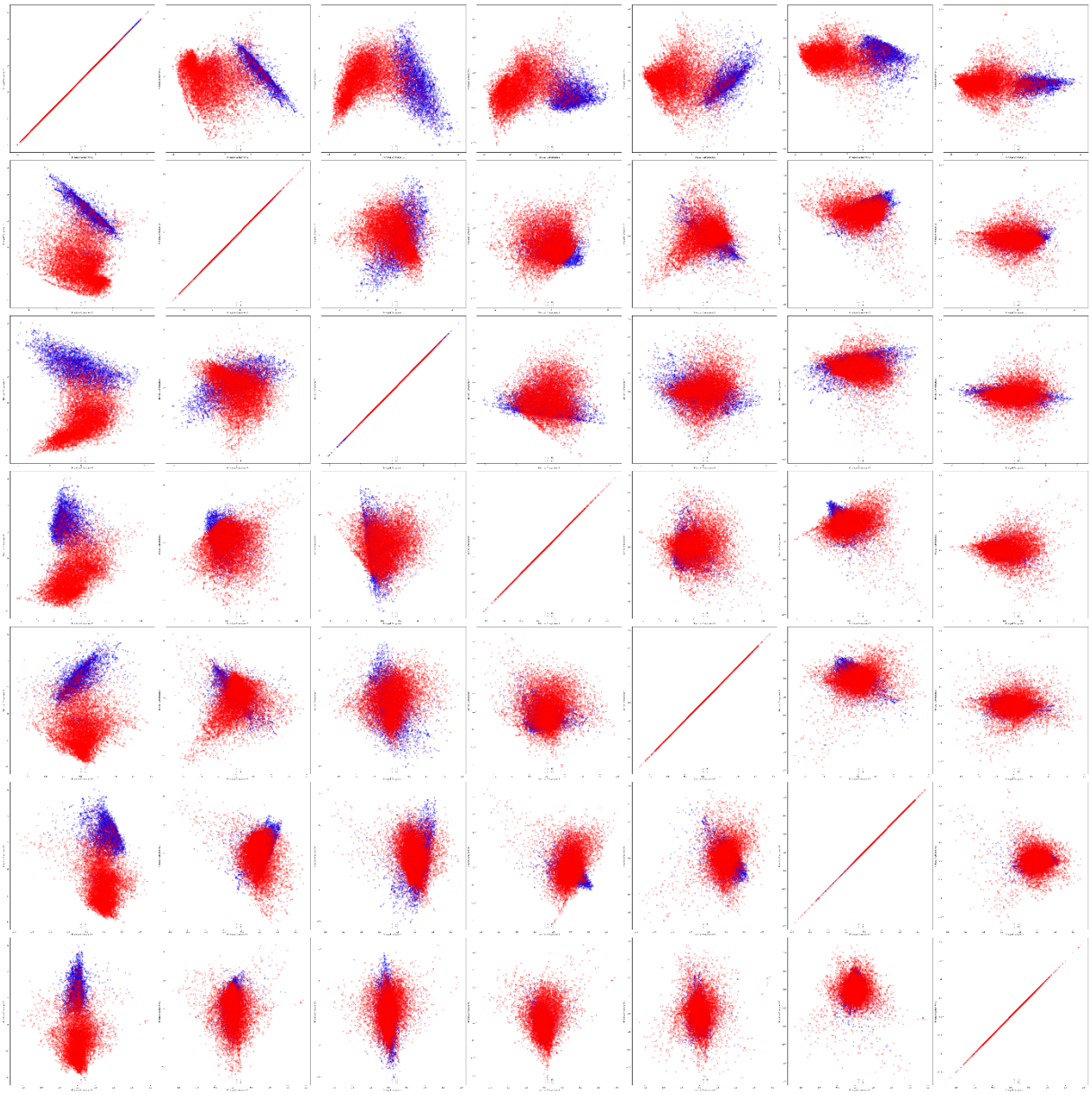


下面是目标降维数据：


0
1
1.0
0.8
0.6
0.4
0.2
0.0
sepal_wid
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
sepal_len

我们首先尝试的是 7 维降到 2 维：

按照所有可能的特征值进行组合，绘制出以下结果，图中 0 为蓝色，1 为红色：



观察图片右上角部分可见，将 7 维降维到 2 维，其区分效果并不十分明显，所以我们考虑更降维到更高维度。

为了方便，我们使用 sklearn[2]的 decomposition 模块带有的 PCA 算法计算功能。通过“from sklearn.decomposition import PCA”可以导入 PCA。sklearn.decomposition 模块包括矩阵分解算法，包括 PCA，NMF 或 ICA。 该模块的大多数算法可以被视为降维技术。

在经过调参之后，我们决定使用 PCA 主成分的方差和所占的最小比例阈值 0.95 来

进行降维，这样可以实现将 7 维降维到 3 维。

新建“MModeling2.ipynb”，使用 sklearn 的 PCA 功能，设定主成分的方差和所占的最小比例阈值 0.95，得到特征向量矩阵：

```
[[-0.53233416 -0.55912462 -0.48940805 -0.18916932 -0.32799045 -0.094590
19
  -0.11031631]
 [-0.30302547 -0.03427317 -0.22147118  0.27836892  0.3662336   0.588740
65
   0.5474665 ]
 [-0.07201895  0.72571157 -0.51590055  0.1264762  -0.42464461  0.059047
17
  -0.04686977]]
```

特征值：

```
[0.32440247 0.0289495  0.01383733]
```

上面所示特征向量矩阵为 PCA 算法生成的特征向量矩阵 V，原数据矩阵 A 点乘 V 矩阵的转置之后便可得到降维之后的数据矩阵 A1，

权重矩阵的横向 index 为：

Index[‘V1’,’ V2’,’ V3’,’ V4’,’ V5’,’ V6’,’ V7’]

即特征向量 V 矩阵中每组的七个数据都代表着 V1 到 V7 各指标的权重。

由特征向量矩阵可得主成分得分系数矩阵：

	X	Y	Z
V1	-0.53233416	-0.30302547	-0.07201895
V2	-0.55912462	-0.03427317	0.72571157
V3	-0.4894085	-0.22147118	-0.51590055
V4	-0.18916932	-0.27836892	0.1264762
V5	-0.32799045	0.3662336	-0.42464461
V6	-0.09459019	0.58874065	0.05904717
V7	-0.11031631	0.5474665	-0.04686977

因此，X，Y，Z 的因子得分模型分别为：

X=-0.53233416*V1-0.55912462*V2-0.4894085*V3-0.18916932*V4
-0.32799045*V5-0.09459019*V6-0.11031631*V7

Y=-0.30302547*V1-0.03427317*V2-0.22147118*V3-0.27836892*V4

+0.3662336*V5+0.58874065*V6+0.5474665*V7

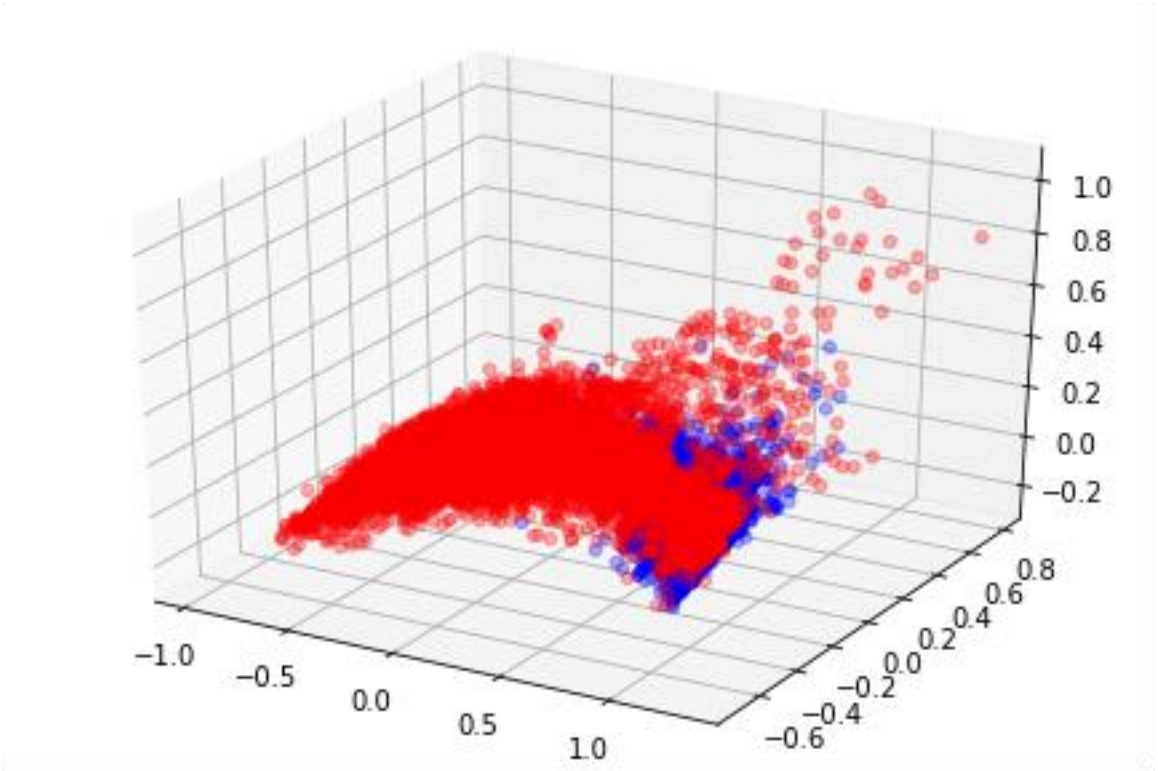
Z=-0.07201895*V1+0.72571157*V2-0.51590055*V3+0.1264762*V4
-0.42464461*V5+0.05904717*V6-0.04686977*V7

主成分贡献率与特征值如下表：

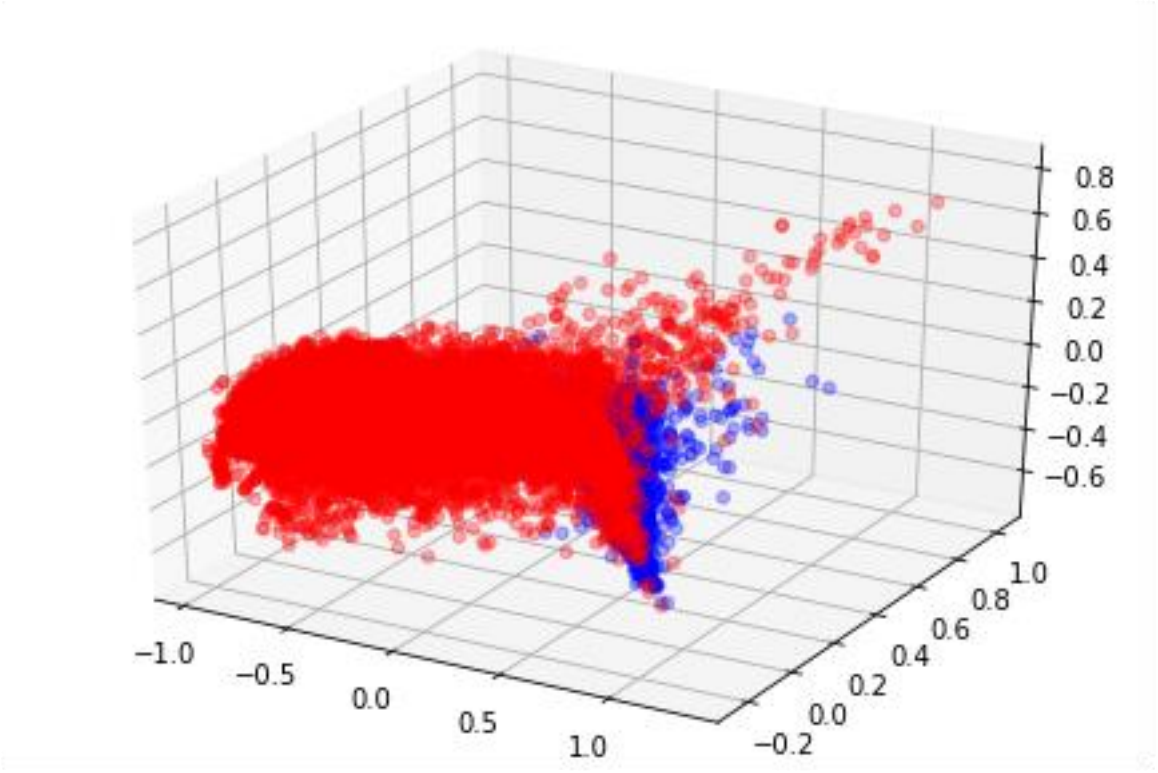
成分	特征值	方差贡献率	累计贡献率
1	0.3244	0.84126204	0.84126204
2	0.0289	0.07507376	0.9163358
3	0.0138	0.03588389	0.95221969

下面六张图是在使用 PCA 算法将 7 维数据降至 3 维之后在三维坐标系下的图形示意图（红色的点 Classification 为 1，蓝色的点 Classification 为 0）

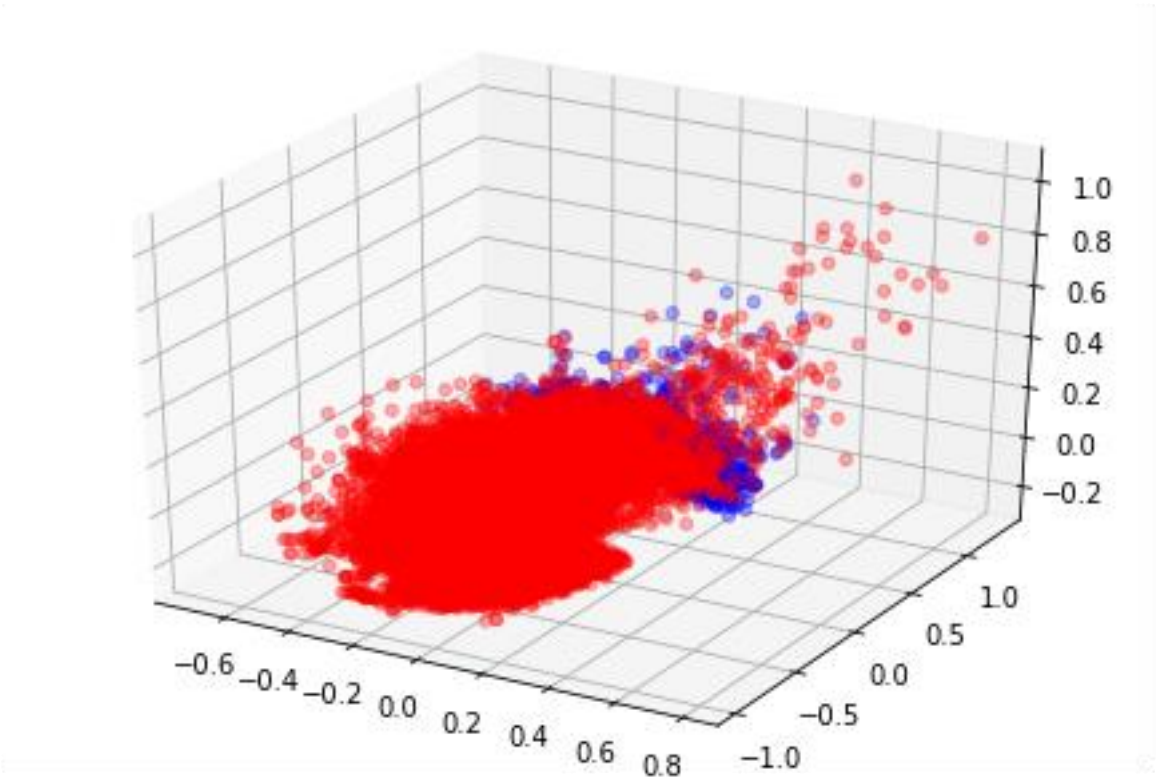
XYZ:



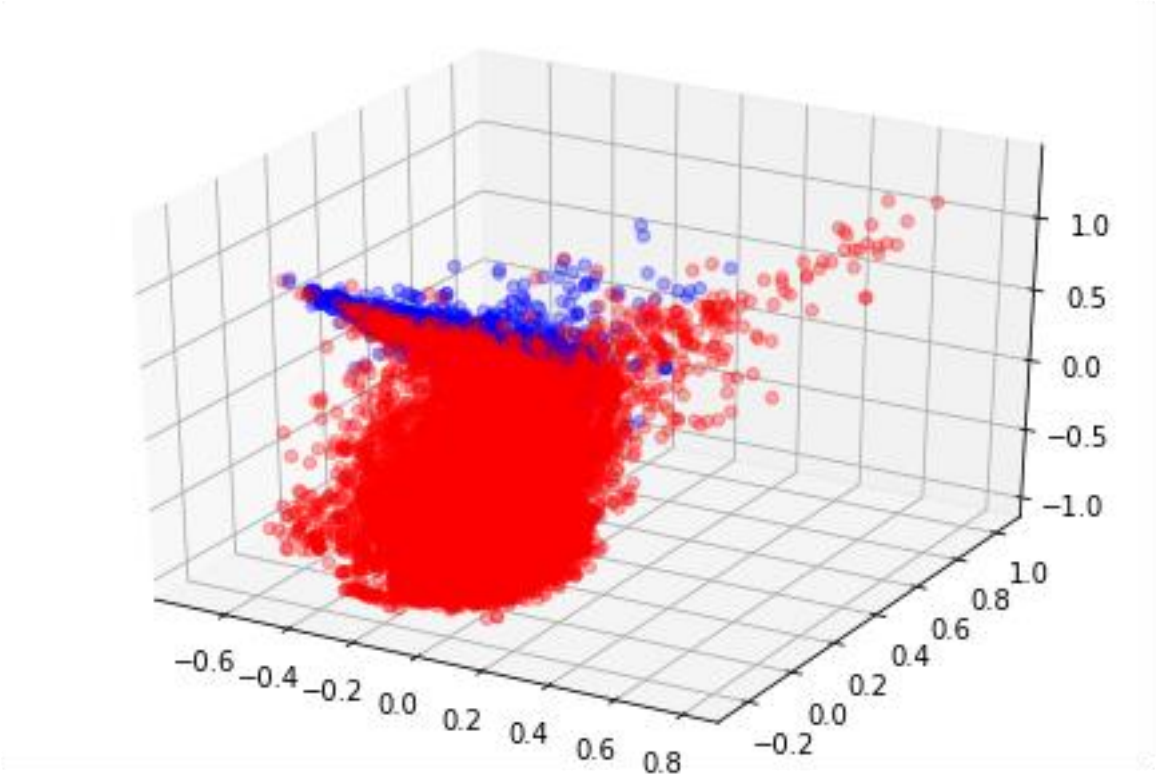
XZY:


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.0
−0.5
0.0
0.5
1.0
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

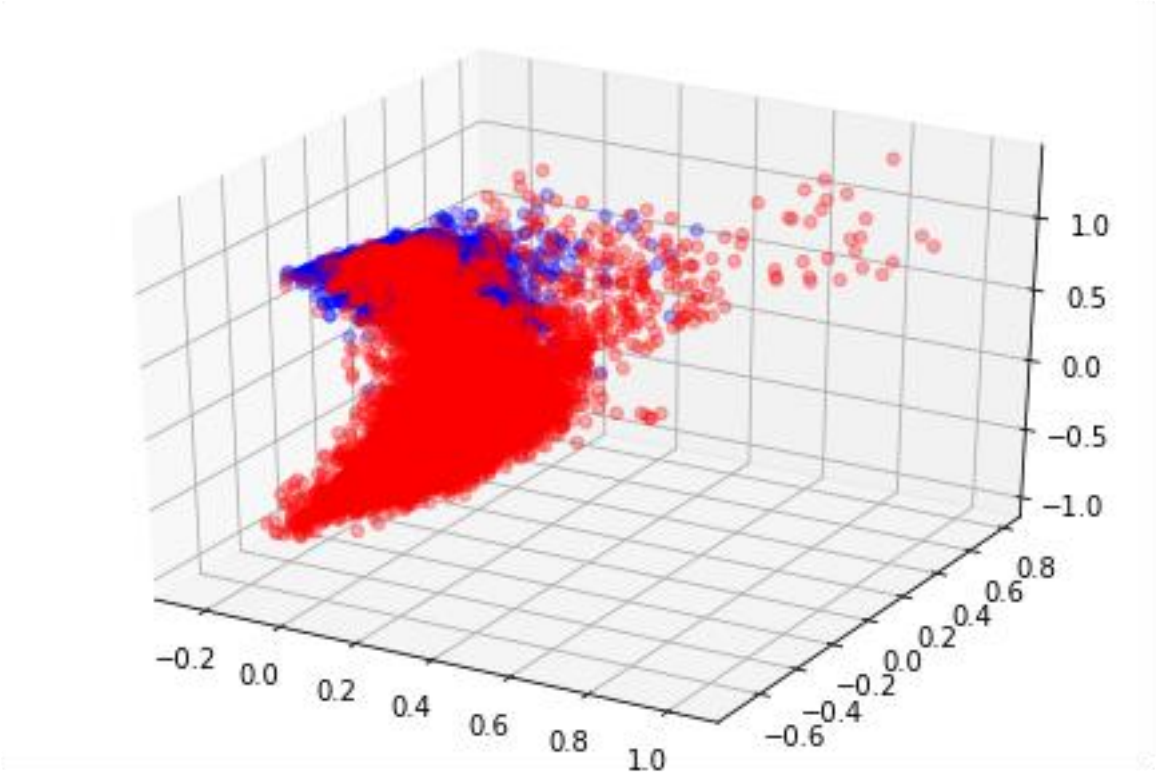
YXZ:


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.0
−0.5
−1.0

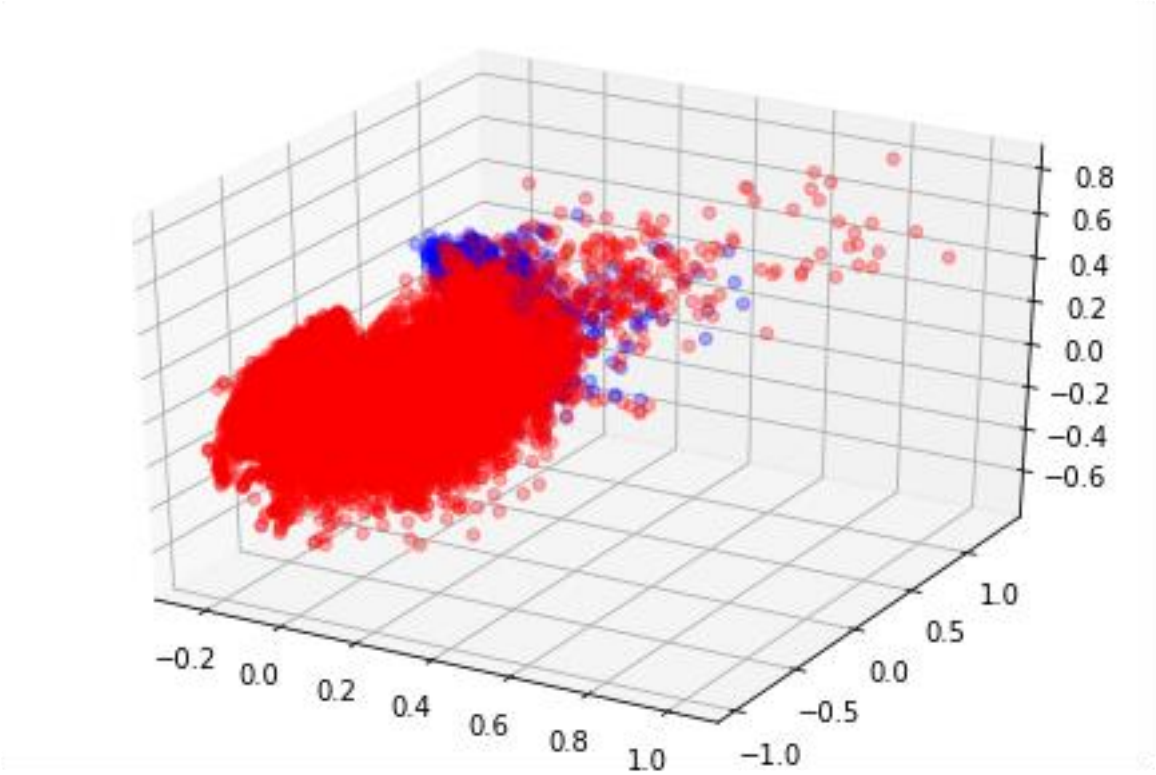
YZX:


1.0
0.5
0.0
−0.5
−1.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

ZYX:


1.0
0.5
0.0
−0.5
−1.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

ZXY:



根据主成分分析的的分析系数得知，X 主要由 V1，V2，V3，V5 因素决定，Y 主要由 V1，V3，V4，V5，V6，V7 因素决定，Z 主要由 V2，V3，V5 因素决定，由图分析可得图像中红色的主要集中在 x 轴和 y 轴的负半轴，由因子得分模型可知，V1，V3 为主要影响因素。

4.2 问题 2

模糊区域在图中可表示为蓝色与红色点交杂的区域。

由第一问的 6 张散点图可知，其大至为 x 在 0，0.5 之间，y 位于 0，0.6 之间，z 小于 0 的区域为两种特征的点交杂的模糊区域，X 中 V4，V6，V7 的影响较小，Y 中 V2 的影响较小，Z 中，V1，V4，V6，V7 的影响较小，综上判断 V6，V7 很难判断出成分是否存在。

4.3 问题 3

我们选取了 sklearn 中的 17 种分类机器学习方法，将 20000 个训练集划分为 18000 个训练数据和 2000 个验证数据在通过 PCA 降维之后进行训练和测试，代码见“MModeling3.ipynb”。最终得到以下分类模型预测的准确率：

	Model	Score
1	Random Forest	0.9425
3	K-Nearest Neighbours	0.9400
5	SVM	0.9390
6	Nu-Support Vector Classification	0.9390
14	Gradient Boosting Classifier	0.9380
12	Bagging classifier	0.9335
4	Naive Bayes	0.9290
13	AdaBoost classifier	0.9290
2	LogisticRegression	0.9265
7	Linear Support Vector Classification	0.9260
8	Radius Neighbors Classifier	0.9250
15	Linear Discriminant Analysis	0.9230
16	Quadratic Discriminant Analysis	0.9230
9	Passive Aggressive Classifier	0.9115
0	Decision Tree	0.9025
11	ExtraTreeClassifier	0.8880
10	BernoulliNB	0.8450

同时，为了补充验证，我们又将未经过 PCA 降维的数据进行训练测试，可得到以下准确率：

	Model	Score
1	Random Forest	0.9460
3	K-Nearest Neighbours	0.9440
14	Gradient Boosting Classifier	0.9425
13	AdaBoost classifier	0.9405
12	Bagging classifier	0.9395
5	SVM	0.9395
6	Nu-Support Vector Classification	0.9395
2	LogisticRegression	0.9300
7	Linear Support Vector Classification	0.9265
8	Radius Neighbors Classifier	0.9260
15	Linear Discriminant Analysis	0.9255
9	Passive Aggressive Classifier	0.9235
16	Quadratic Discriminant Analysis	0.9215
0	Decision Tree	0.9155
11	ExtraTreeClassifier	0.9135
4	Naive Bayes	0.9120
10	BernoulliNB	0.8440

可见经过降维的数据进行训练和未经过降维的数据进行训练，其准确率未有明显变化，也从侧面证明了我们第一问和第二问使用 PCA 进行降维的正确性。

因而，我们最终选取准确率最高的随机森林方法作为我们选用的数学模型。

测试数据预测结果已经附在附件中，前 10 个混合物的判定为：均含有特定成分。

5　模型的评价

5.1　模型的优势

通过尝试 17 种机器学习方法，对每种方法进行调参来确保获得最优效果，最后选取其中表现最佳的方法——随机森林 来生成测试数据预测结果。

5.2　可以改进的地方

这次因为建模时间紧张，只运用了传统的 17 种分类机器方法，未运用深度学习——神经网络的方法进行建模。后续如果有时间，在获得更大数据量的前提下，可以考虑使用神经网络进行训练得出模型，以期获得更高的准确率。

6　参考文档

[1] 李春春_，主成分分析（principal components analysis，PCA）——无监督学习 ，https://blog.csdn.net/zhongkelee/article/details/44064401，2020 年 7 月 18 日

[2] 未知作者，sklearn 官方文档，https://scikit-learn.org/stable/modules/classes.html，2020 年 7 月 19 日