**Lab02实验报告**

**项子扬 PB16001768**

1. **数据预处理**

要求为：将2个数据集分别随机平均分为5份并保存，要求每个被试人员出现在每份样本集中的数量均匀。后续使用5折交叉验证。可以对数据进行删减、降维、组合等。

1. **SVM**

要求为：使用非线性kernel的软边界SVM对2个数据集的valence\_arousal\_label进行分类，分析结果。

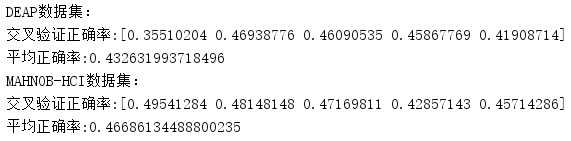
实验内容：

语言：Python3.6.2 环境：Windows10

数据预处理：采用sklearn工具包中的preprocessing工具包，将数据平均划分成5份，做5折交叉检验。

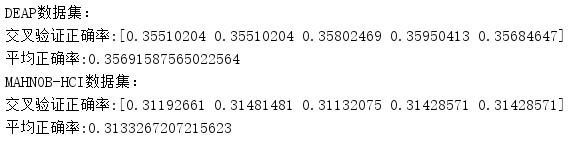
测试方面：根据训练结果对测试结果进行比对，计算得出正确率。核函数为rbf（高斯核函数）。将其他输入属性各自归一化再作为SVM的输入属性。

经过测试惩罚系数C=6时分类成功率较高：

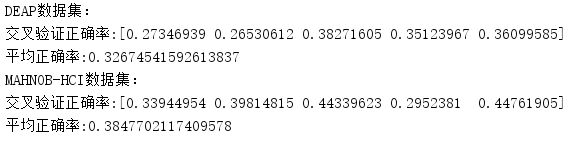


C过高或过低都会导致成功率降低：

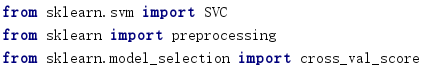
C=0.01：



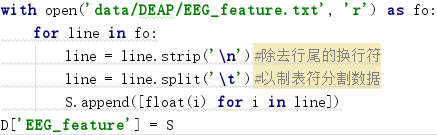
C=100：



代码说明：



此部分为导入的库函数，SVC为支持向量机库函数，cross\_val\_score用于交叉验证。



此部分为从文件中读取数据并存入字典中，利用strip和split函数将文件中的元素分割开来以便存储。



此句语句用于将属性一并存入X表，用作库函数的输入。



此语句用于将输入属性归一化，用于提高分类成功率。



此语句将待分类标签的两个分量组合成一个（取值范围为11、12、21、22）。



此语句调用库函数进行分类，并使用分类器对测试集验证得到正确率。

**3.** **Naïve Bayes（手动实现）**

要求为：使用带拉普拉斯修正的Naïve Bayes对2个数据的被试人员(subject id)，对MAHNOB-HCI数据集的情感标签进行分类，分析结果。

实验内容：

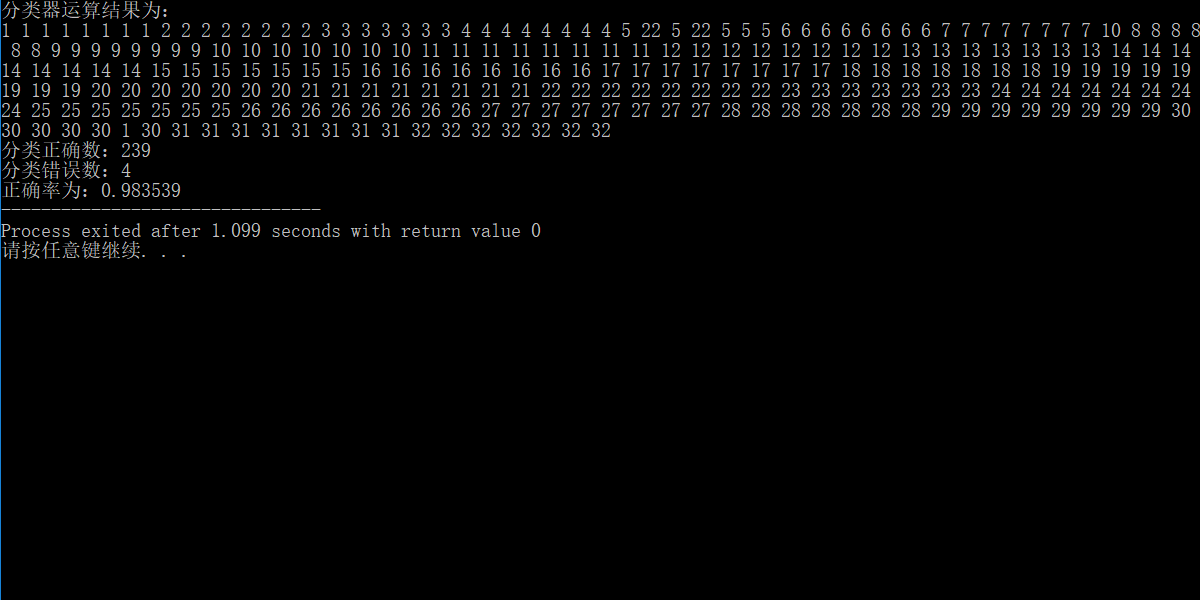
语言：C语言 环境：Windows10

数据预处理方面：经分析，在对被试人员进行分类时，视频编号标签没有太大意义，故删去该标签；在对情感标签进行分类时，被试人员和视频编号标签没有太大意义，故删去这两个标签。valence\_arousal\_label标签由于只有1和1、1和2、2和1、2和2四种取法，故组合为一个标签，取值为离散的{11,12,21,22}；DEAP数据集中共有1216个样本，MAHNOB-HCI数据集中共有533个样本，为了方便分组以及读取，故舍弃若干样本，将剩余的1215个样本和530个样本按被试人员编号(subject id)均匀分配至5个文本文件中。

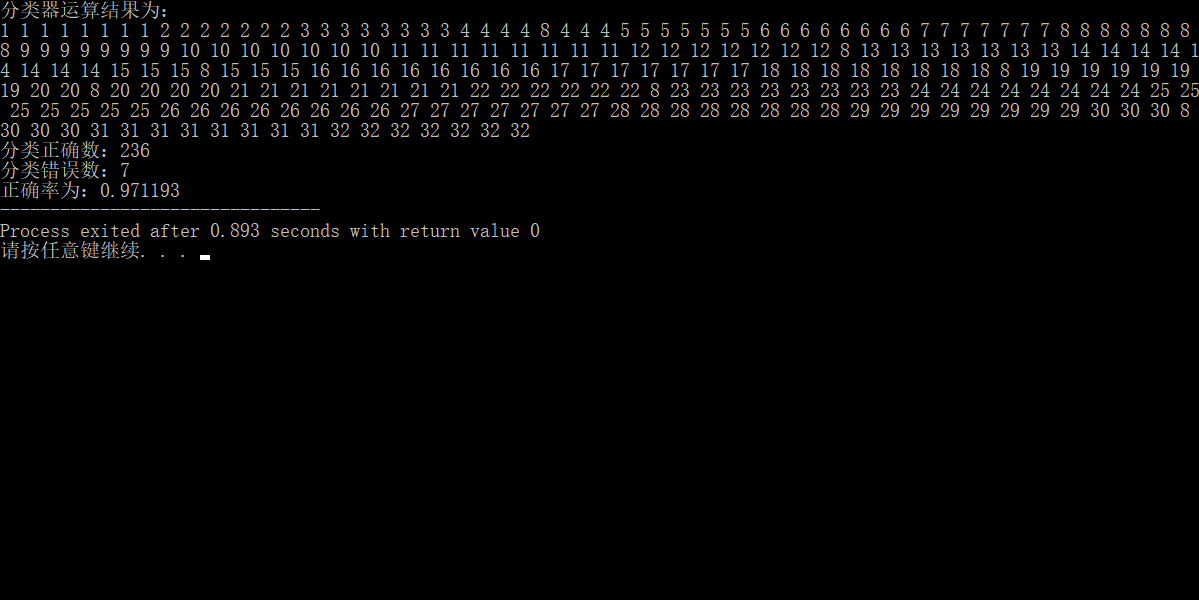
测试方面：根据训练结果对测试结果进行比对，记录正确和错误的结果数，计算得出正确率。

1.DEAP数据集的被试人员(subject id)分类：

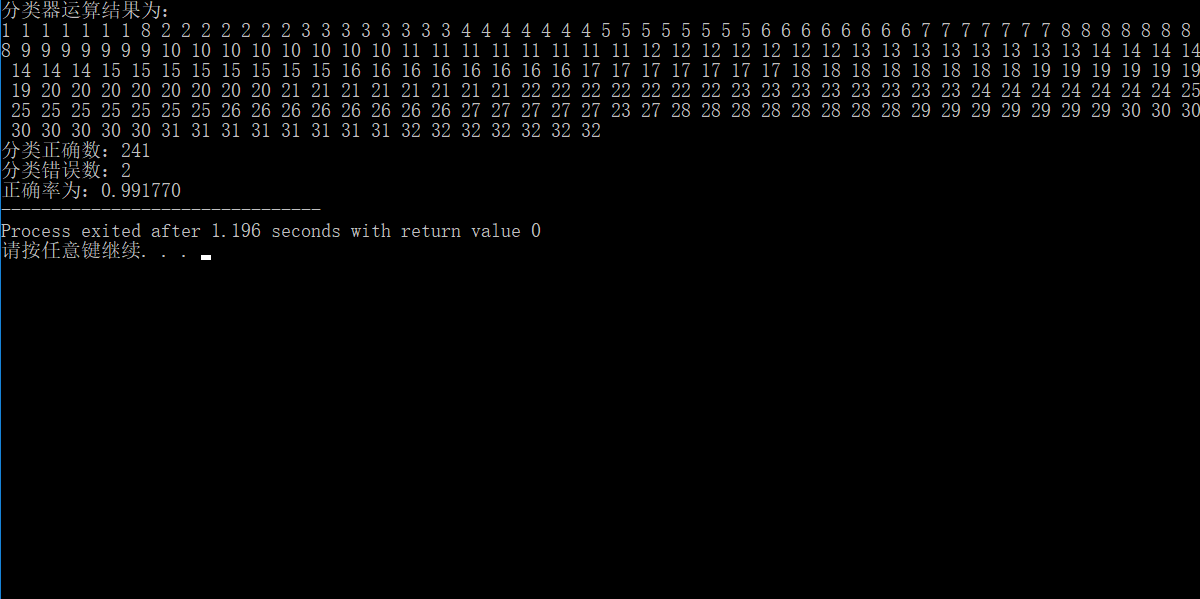
第一次：



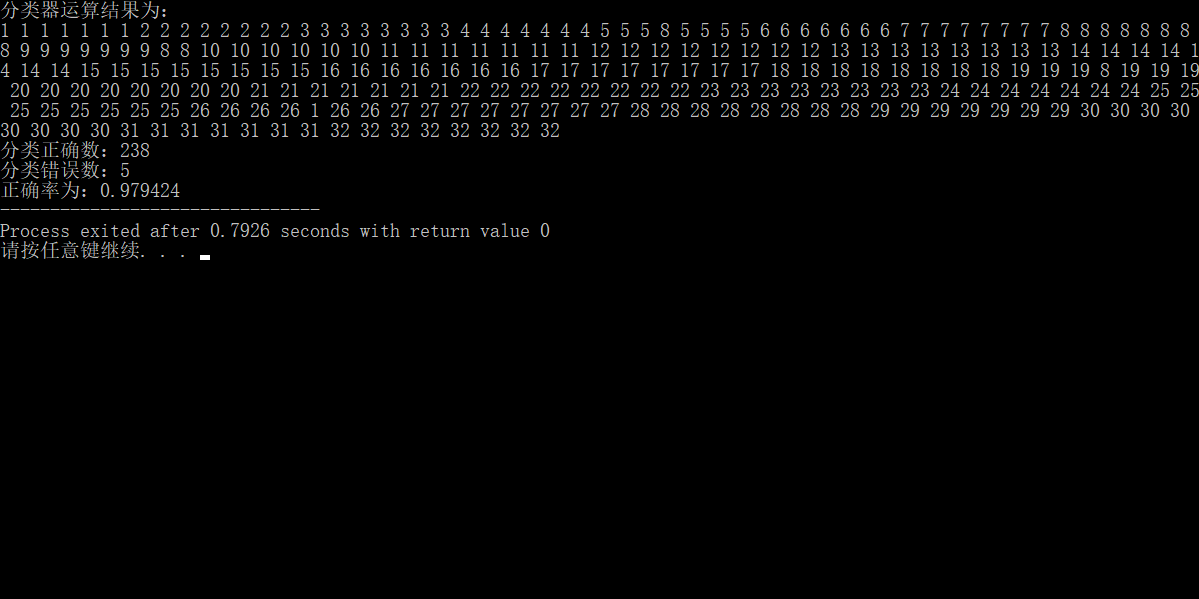
第二次：



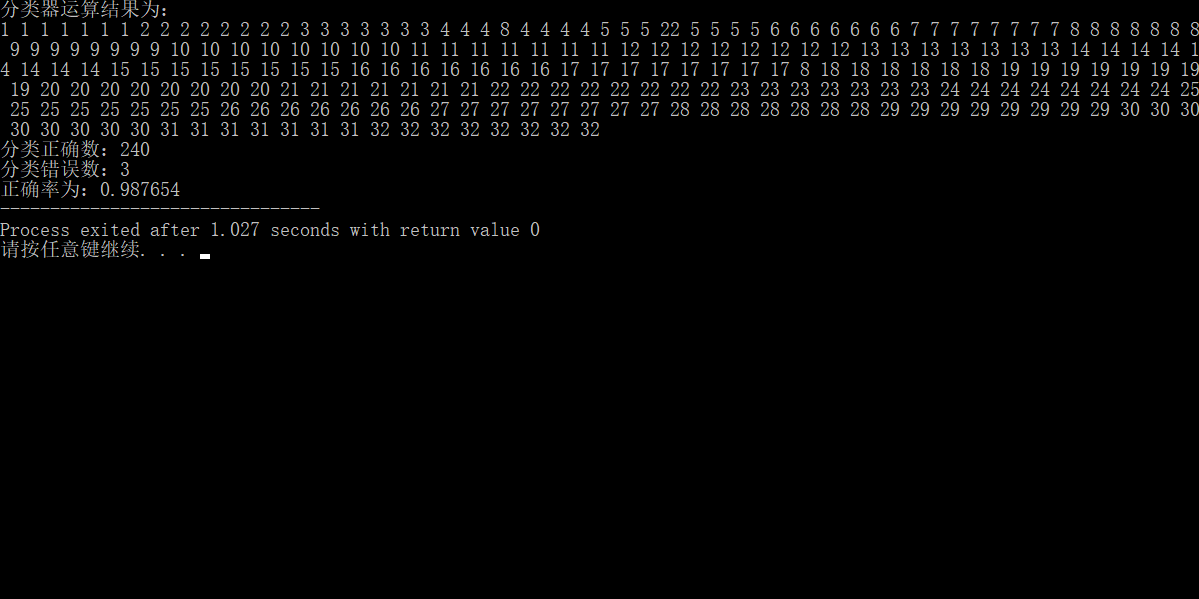
第三次：



第四次：



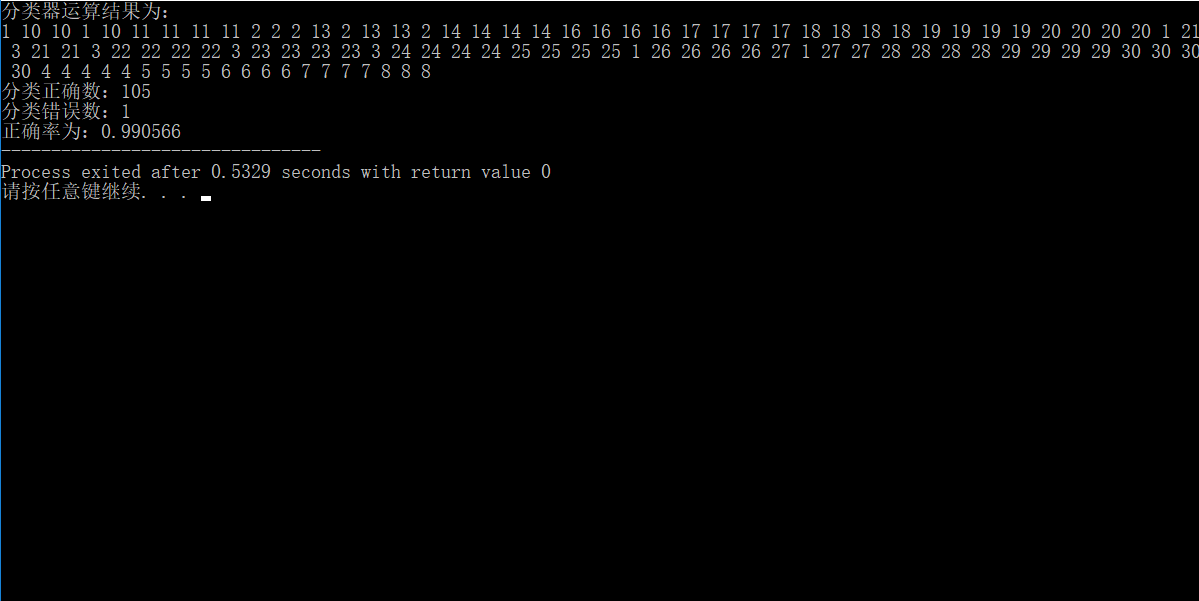
第五次：



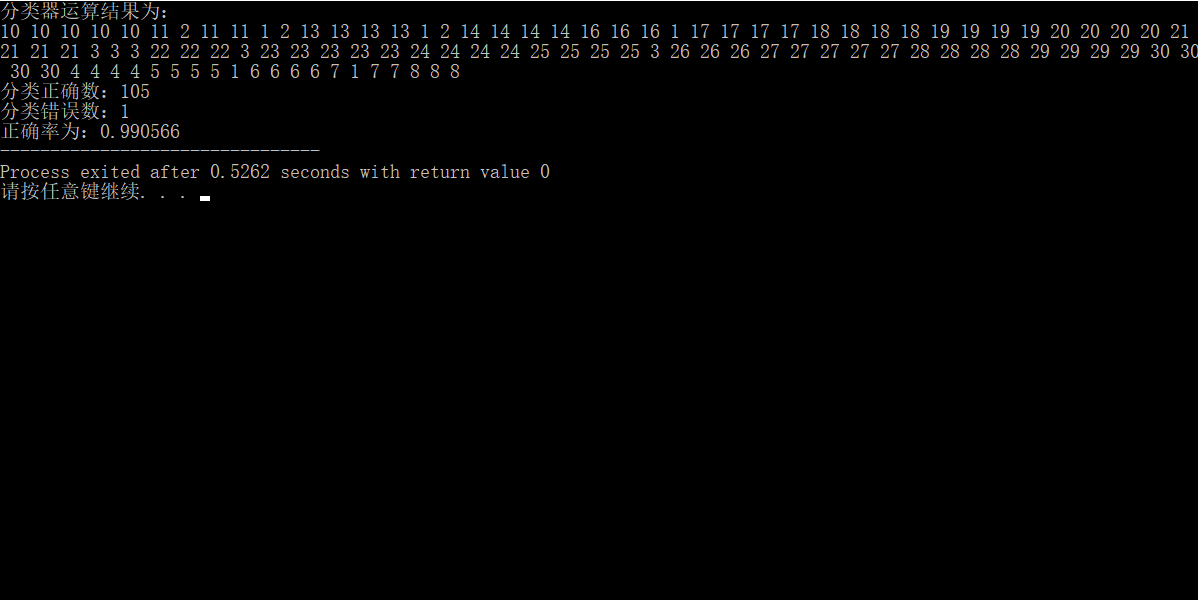
5次平均正确率为98.27%。

2.MAHNOB-HCI数据集的被试人员(subject id)分类：

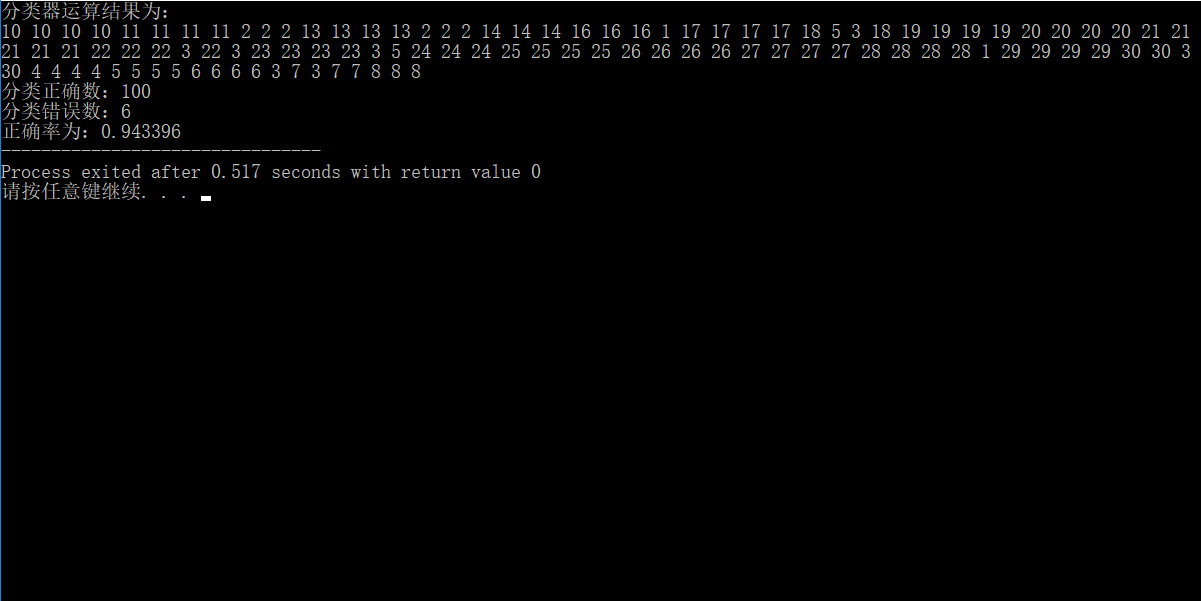
第一次：



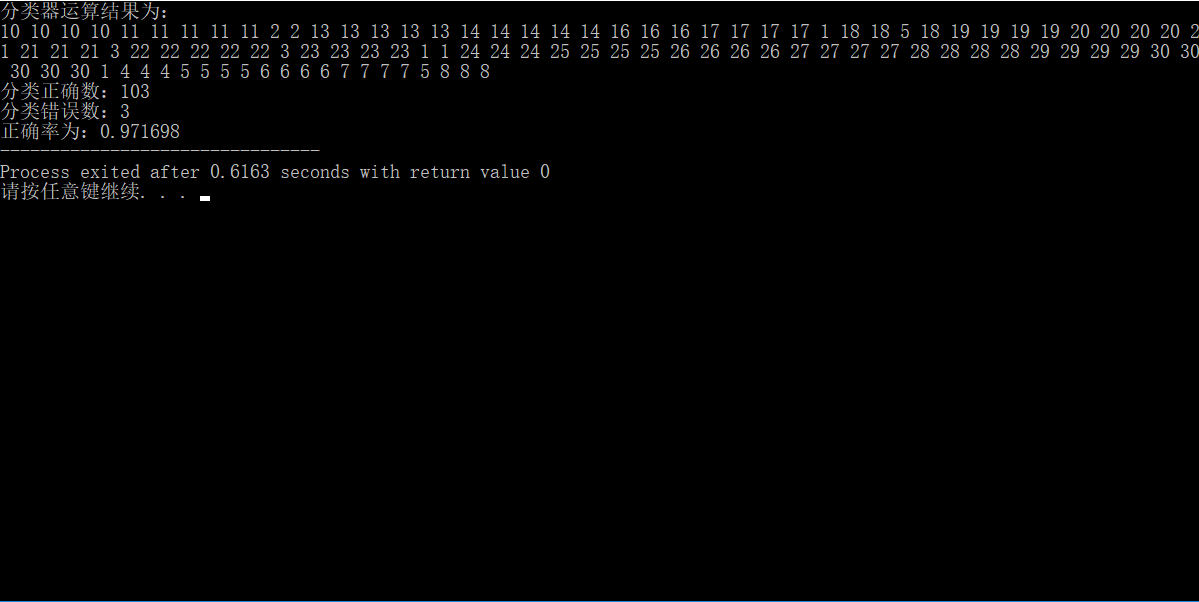
第二次：



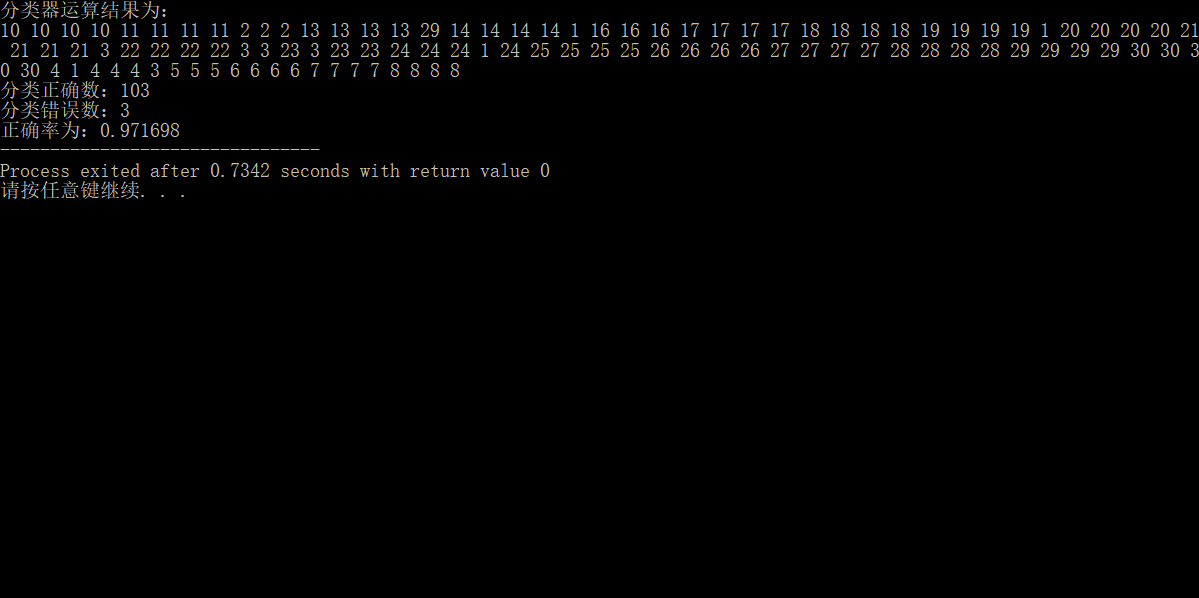
第三次：



第四次：



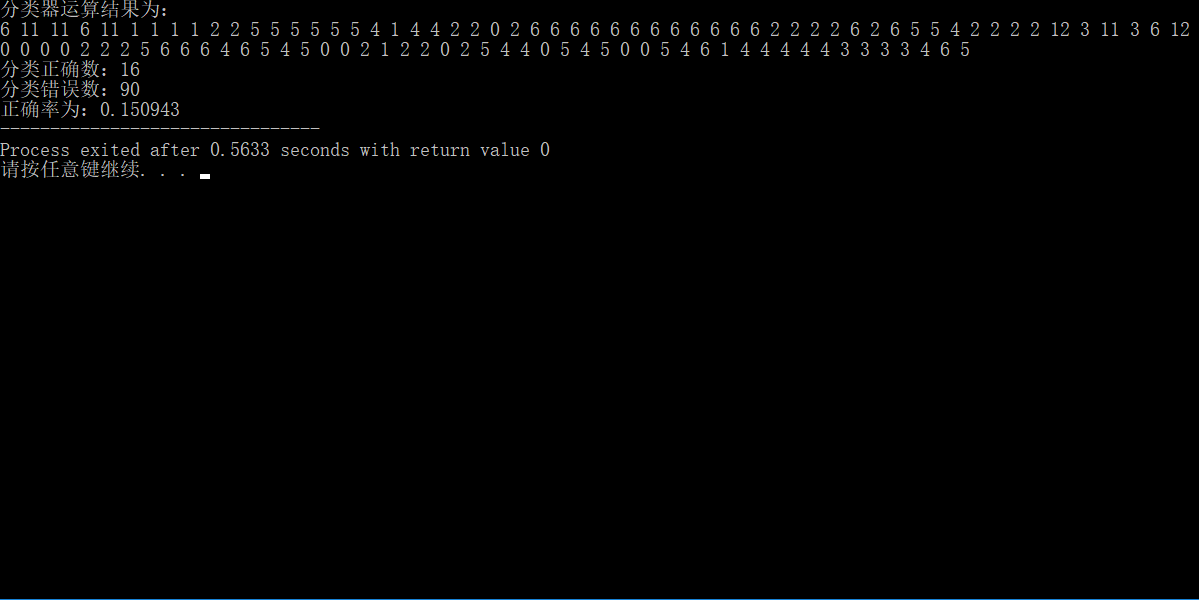
第五次：



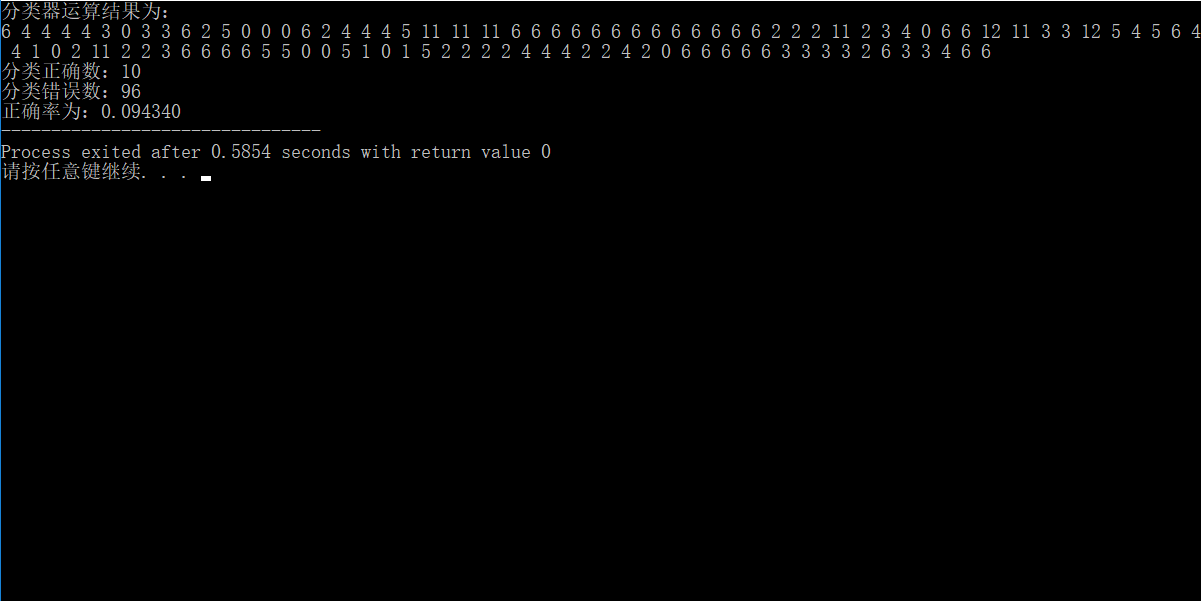
5次平均正确率为97.36%。

3.MAHNOB-HCI数据集的valence\_arousal\_label标签分类：

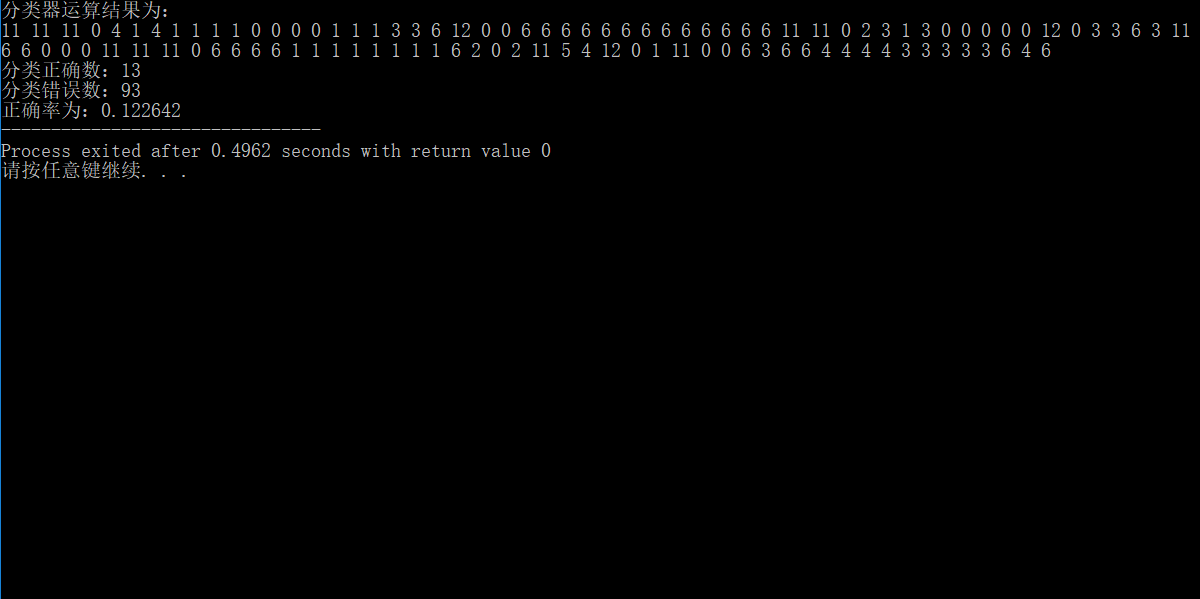
第一次：



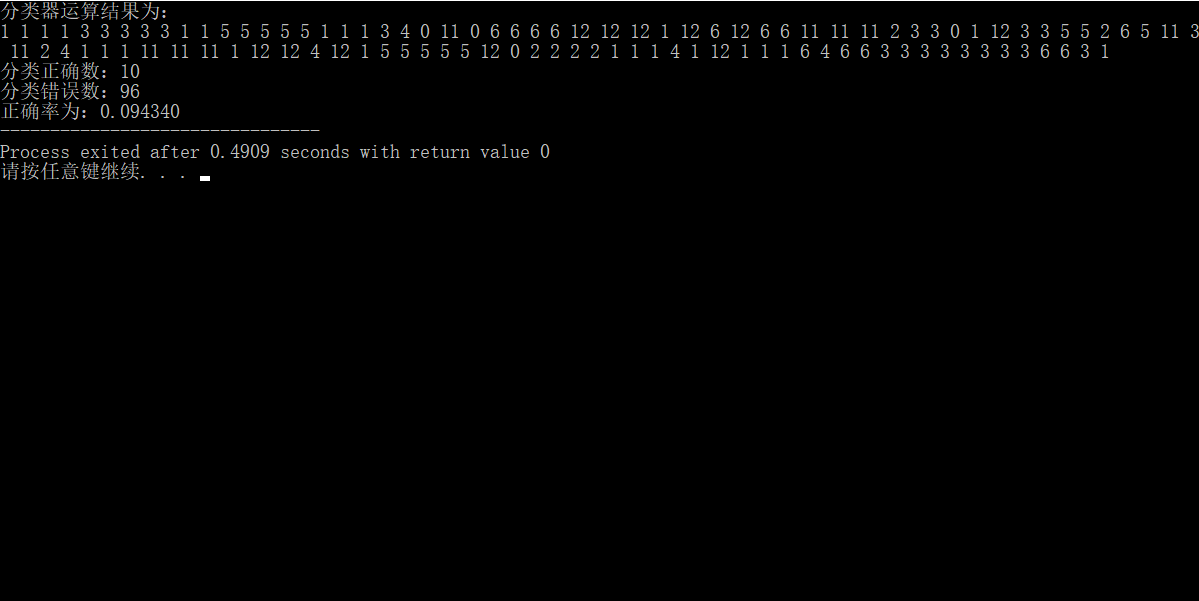
第二次：



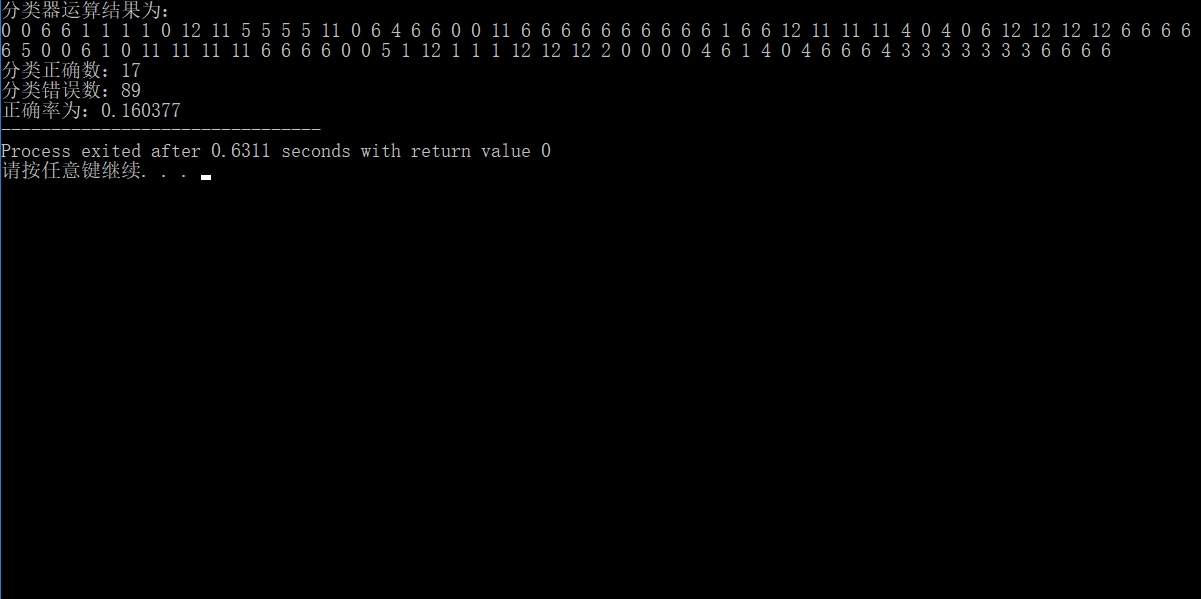
第三次：



第四次：



第五次：

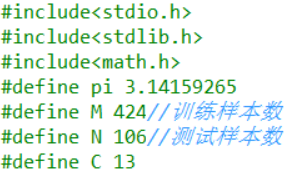


5次平均正确率为12.45%

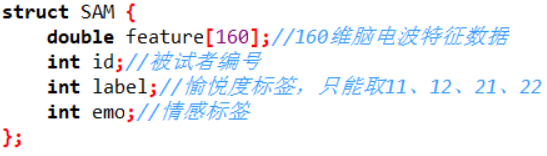
结果分析：

在对两个数据集的被试人员进行分类时，正确率较高（97%），而对MAHNOB-HCI数据集的情感标签进行分类时，正确率较低（12%），可能是由于不同情绪在本实验中各属性上的影响有些类似，导致分类失误；也有可能是删去的被试人员和视频编号对实验有较大影响。

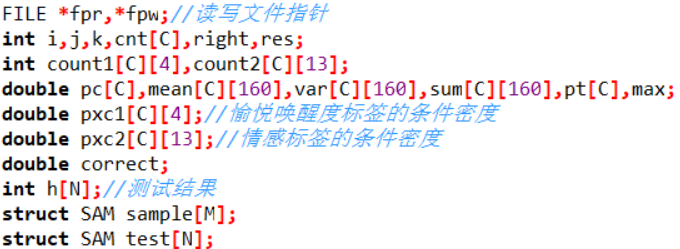
代码说明：



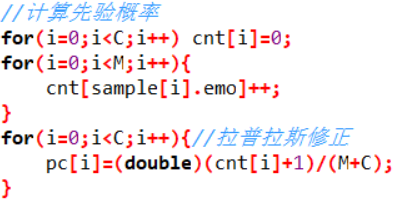
此部分为宏定义，M是训练集的大小，N是测试集的大小，C是待分类标签的取值范围大小。



此为存储样本各个属性的结构体，feature是脑电波特征，id是被试人员编号，label是经过合并的愉悦度-唤醒度标签，取值为{11,12,21,22}，emo是情绪标签（只会在MAHNOB-HCI数据中出现）。



此为程序变量定义。fpr、fpw为读写文件指针，cnt用于统计待分类标签的数量，count1、count2用于统计离散标签的数量；pc为先验概率，mean为平均值，var为方差，sum用于计算平均值，pt为测试样例计算的概率，pxc1、px2为离散的条件密度，sample、test为训练集数组、测试集数组。



此部分为先验概率的计算，cnt自增时，若对id分类，下标为sample[i].id-1，若对emo分类，下标为sample[i].emo；由于是离散型的统计，需要进行拉普拉斯修正。



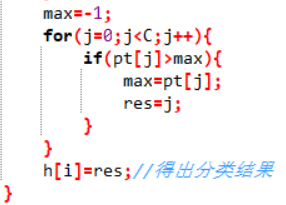
此部分计算的是160维连续变量的平均值和方差，用于之后计算条件概率。由于是连续变量，不需要进行拉普拉斯修正。



此部分计算离散变量的条件概率，若对id进行分类，需要计算label和emo，若对emo进行分类，需要计算label。需要进行拉普拉斯修正。



此处对测试样例的各个分类的概率进行计算，在连续变量的计算时采用的是正态分布概率密度分布函数。



此处是选出概率最大者，作为函数的分类结果。之后经过比对计算正确率。

**4. LR**

此部分实验的内容要求为：使用逻辑回归对2个数据集的valence\_arousal\_label进行分类，分析结果。

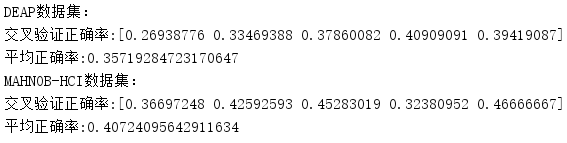
实验内容：

语言：Python3.6.2 环境：Windows10

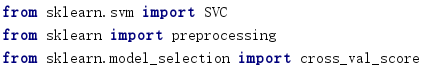
数据预处理：采用sklearn工具包中的preprocessing工具包，将数据平均划分成5份，做5折交叉检验。

测试方面：根据训练结果对测试结果进行比对，计算得出正确率。将其他输入属性各自归一化再作为LR的输入属性。

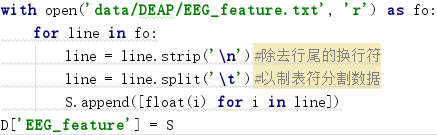
结果：



代码说明：



此部分为导入的库函数，SVC为支持向量机库函数，cross\_val\_score用于交叉验证。



此部分为从文件中读取数据并存入字典中，利用strip和split函数将文件中的元素分割开来以便存储。



此句语句用于将属性一并存入X表，用作库函数的输入。



此语句用于将输入属性归一化，用于提高分类成功率。



此语句将待分类标签的两个分量组合成一个（取值范围为11、12、21、22）。



此语句调用库函数进行分类，并使用分类器对测试集验证得到正确率。