**本 科 实 验 报 告**

实验名称： 贝叶斯分类器实现

学 员： 孙源辰 学 号： 201802001067

培养类型： 无军籍本科学员 年 级： 大三

专 业：人工智能与大数据 所属学院： 计算机学院

指导教员： 刘新旺 职 称： 副教授

实 验 室： 305-205 实验日期： 2021春季学期

国防科学技术大学训练部制

《本科实验报告》填写说明

1．学员完成人才培养方案和课程标准要所要求的每个实验后，均须提交实验报告。

2．实验报告封面必须打印，报告内容可以手写或打印。

3．实验报告内容编排及打印应符合以下要求：

（1）采用A4（21cm×29.7cm）白色复印纸，单面黑字打印。上下左右各侧的页边距均为3cm；缺省文档网格：字号为小4号，中文为宋体，英文和阿拉伯数字为Times New Roman，每页30行，每行36字；页脚距边界为2.5cm，页码置于页脚、居中，采用小5号阿拉伯数字从1开始连续编排，封面不编页码。

（2）报告正文最多可设四级标题，字体均为黑体，第一级标题字号为4号，其余各级标题为小4号；标题序号第一级用“一、”、“二、”……，第二级用“（一）”、“（二）” ……，第三级用“1.”、“2.” ……，第四级用“（1）”、“（2）” ……，分别按序连续编排。

（3）正文插图、表格中的文字字号均为5号。

一、实验目的和要求

通过编程实现贝叶斯分类器的几个模型，理解并掌握相关知识

二、实验内容和原理

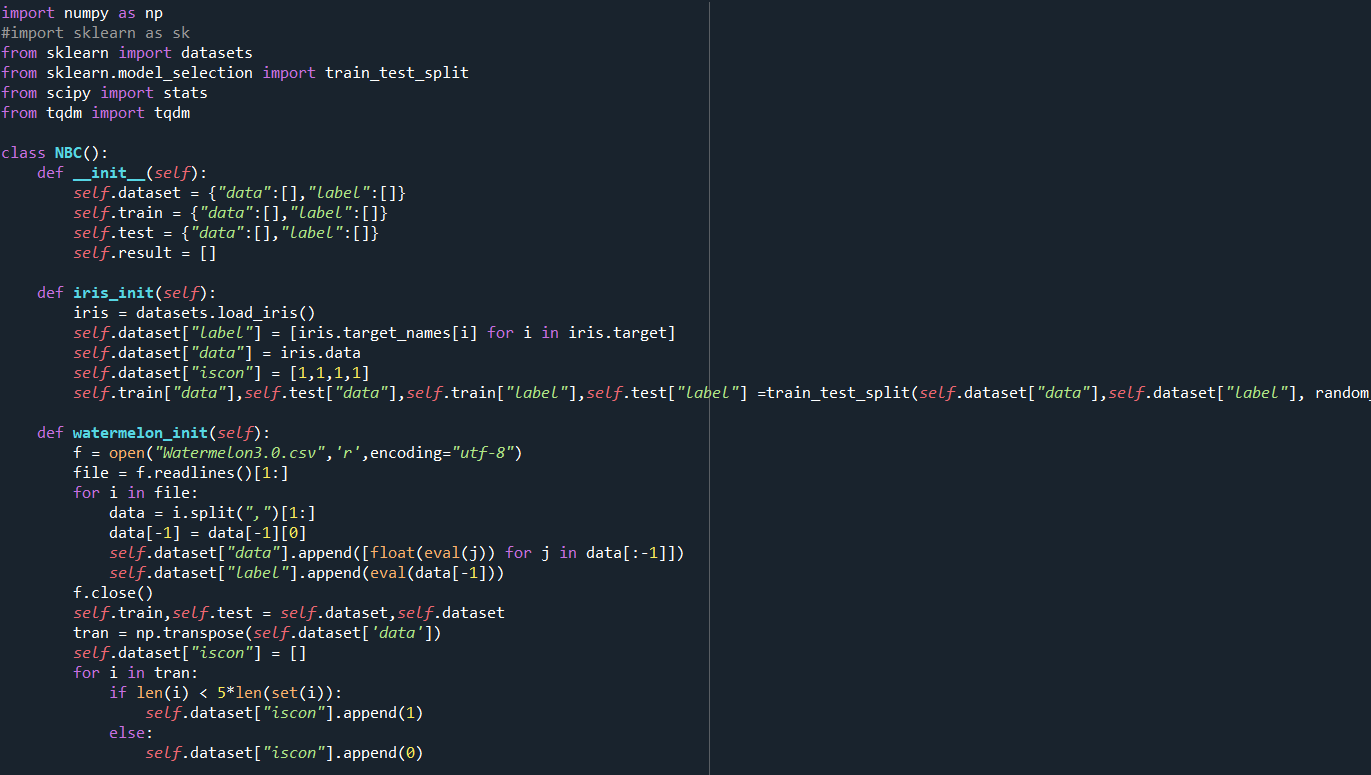
用python实现朴素贝叶斯和一种半朴素贝叶斯分类器，同时使用EM算法实现贝叶斯网。

三、操作方法与实验步骤

(一).编程思路

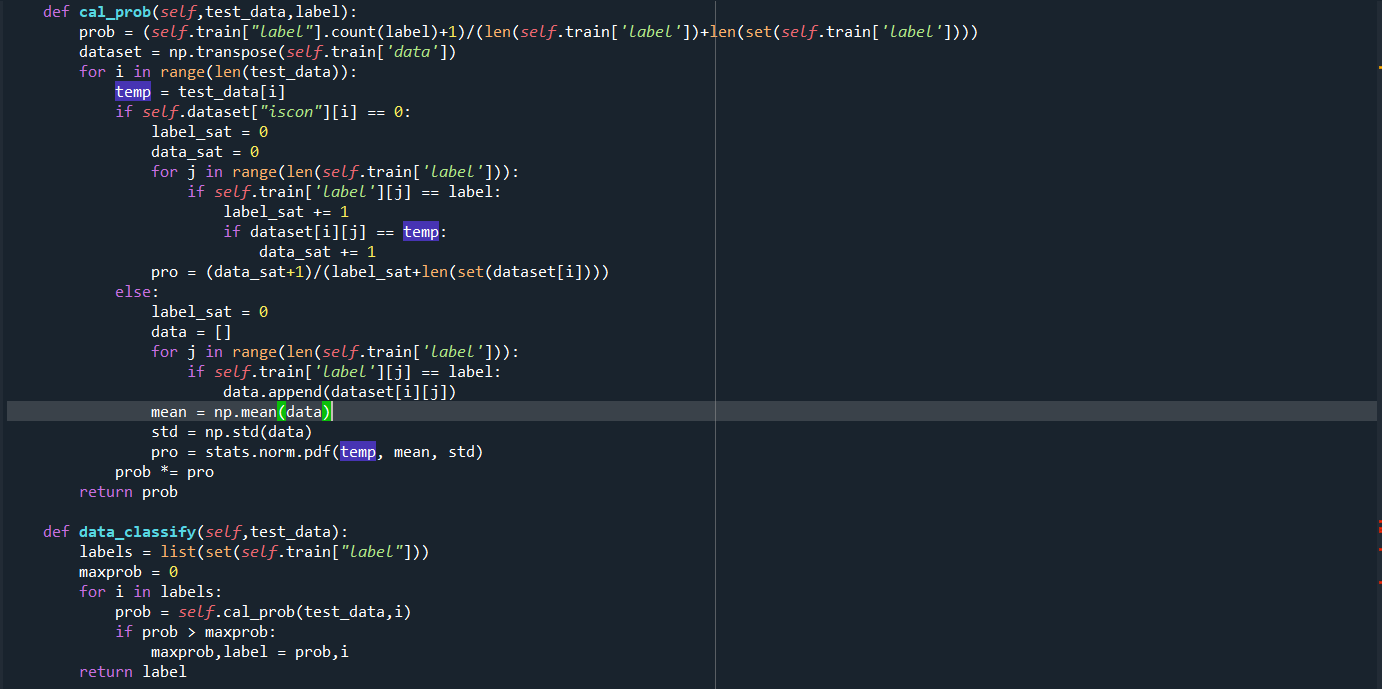
1.朴素贝叶斯

(1)类初始化



初始化NBC类，两个init分别对应两种数据集的初始化，模拟式根据需要选择。Self.dataset为字典，keys分别为data和label，data对应列表为数据矩阵，label为对应标签数组。iscon为某一属性是否连续的标签。

(2)概率计算与分类



这部分由两个函数组成，cal\_prob输入一条数据以及与其分类，计算带有拉普拉斯修正的分类后验概率，data\_classify则是根据前者返回的结果，将数据以最大概率分类。

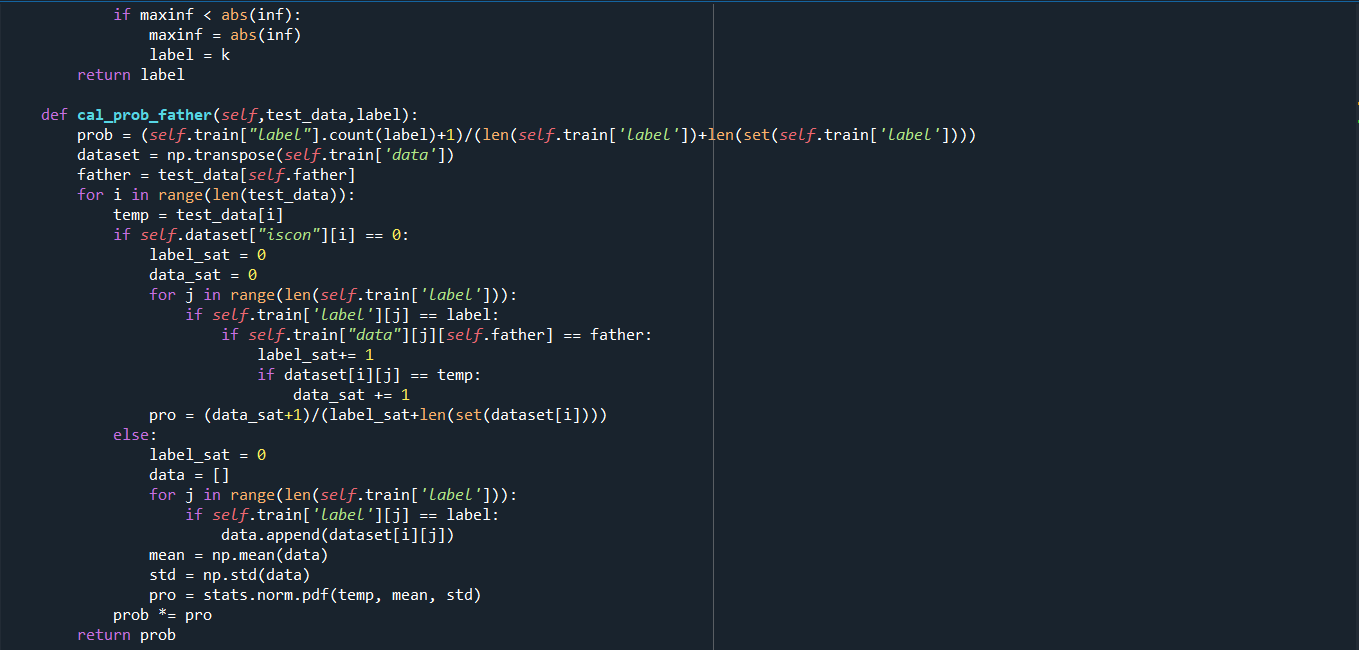
(3)计算准确率



该函数遍历data，对每一条data分类，并和自身标签对比，最后输出总的分类准确率。

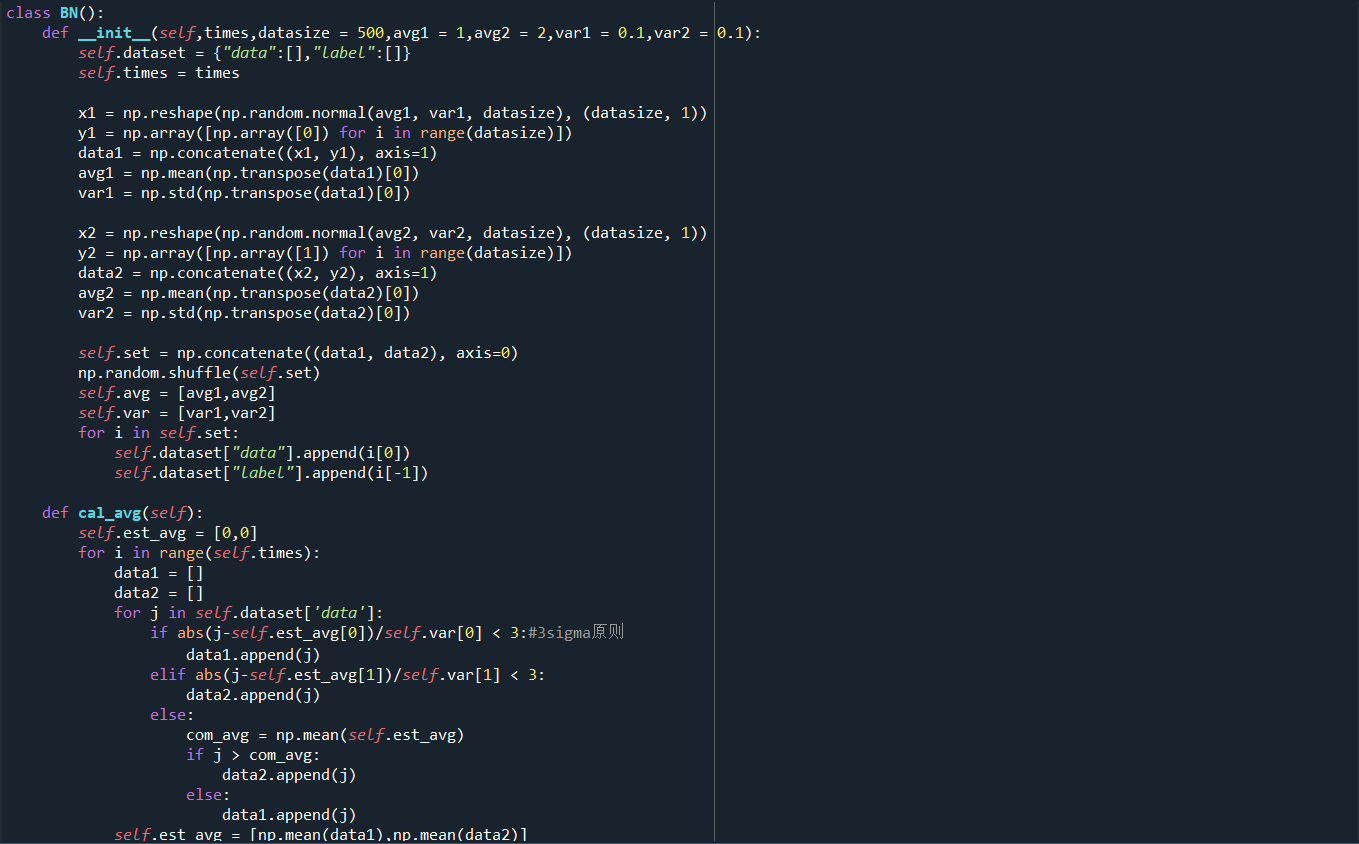
2.SPODE

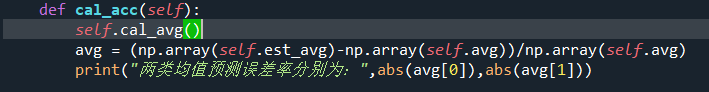




SPODE类继承NBC，编写了选择超父属性的函数，并根据模型改写了概率计算部分，将后验概率里Dc等进行了相应的调换。超父属性由互信息决定。

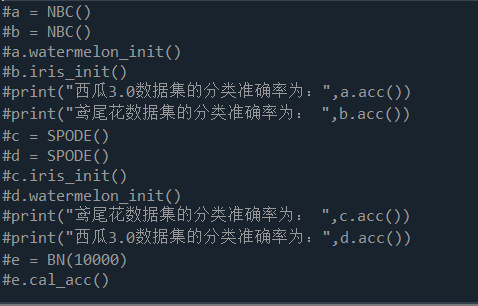
3.估算均值





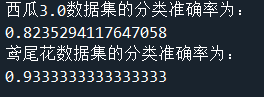
BN类首先初始化了两组同方差但不同均值的数据并混合，为了准确性计算了实际均值与实际方差，之后是用EM算法，E步计算均值，M步通过3σ原则进行类别划分，这里默认是两个不同均值的二分类任务。最后是计算预测均值与实际均值的准确率。

4.调试部分



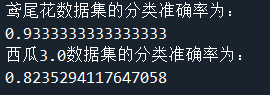
四、实验结果及分析

(一)朴素贝叶斯



因为西瓜3.0数据集数据量太少，所以测试集数据集均为本身，最后结果准确率还可以接受。鸢尾花我做了7：3的数据集分隔(那行太长了截不下来)，训练分类效果很好。

(二)SPODE



很惊奇的发现，和朴素贝叶斯准确率相同，我预测是超父属性的分类型较强。

(三)估算均值

均值result

均值result2

均值result3

对于3σ不交叉(标准差为0.1)的数据训练效果非常好，经10次训练（调试过程中发现第3次分类就已经定型），分类误差率最大在10^-16数量级，而对小有交叉(标准差为0.2)经过1000次训练也可以达到百分之一误差，但交叉很多(标准差0.4)则很难提升性能(10000次训练)，错误率在0.4多。