西南科技大学

计算机实验报告

课程名称：数据挖掘应用技术与实践

实验名称： 贝叶斯算法

学 号：5120180269

学生姓名：李若昊

班 级： 计科1803

指导教师：吴珏

评 分：

实验日期： 2020年9月26日

实验目的：

通过贝叶斯算法的编程实现，加深对贝叶斯算法理解，同时利用贝叶斯算法对简单应用实现预测分类。

实验内容：

1. 分析贝叶斯算法；
2. 计算条件概率；
3. 预测精度的计算与评估；
4. 编程实现贝叶斯分类算法，并对简单应用样本数据进行预测分类。

实验要求：

利用鸢尾花数据集，根据鸢尾花的花萼和花瓣数据，找出不同类别花的特点分布情况，揭示出其中隐藏的规律性。这是一个多值分类问题。

实验步骤：

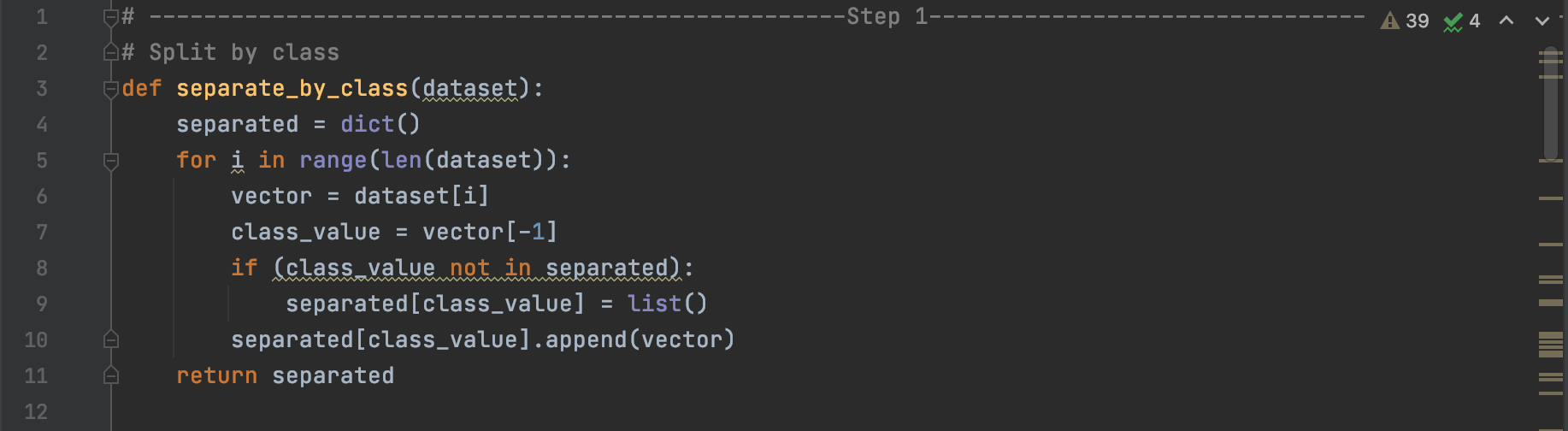
本次实验选用著名的用于模式识别的[鸢尾花数据集](http://archive.ics.uci.edu/ml/datasets/Iris)，共150个实例。其包括鸢尾花的3个种类，且每个类别各有50个实例。每个实例包含了5个属性：sepal length（花萼长）、sepal width（花萼宽）、petal length（花瓣长）、petal width（花瓣宽）、class（类别）。除最后一个为类别属性外，其余属性均以数值定义，单位是厘米。为了简化问题，假设各个属性之间不存在相关性。

以下1～3步骤为训练模型，4步骤为应用模型，5步骤为评估模型

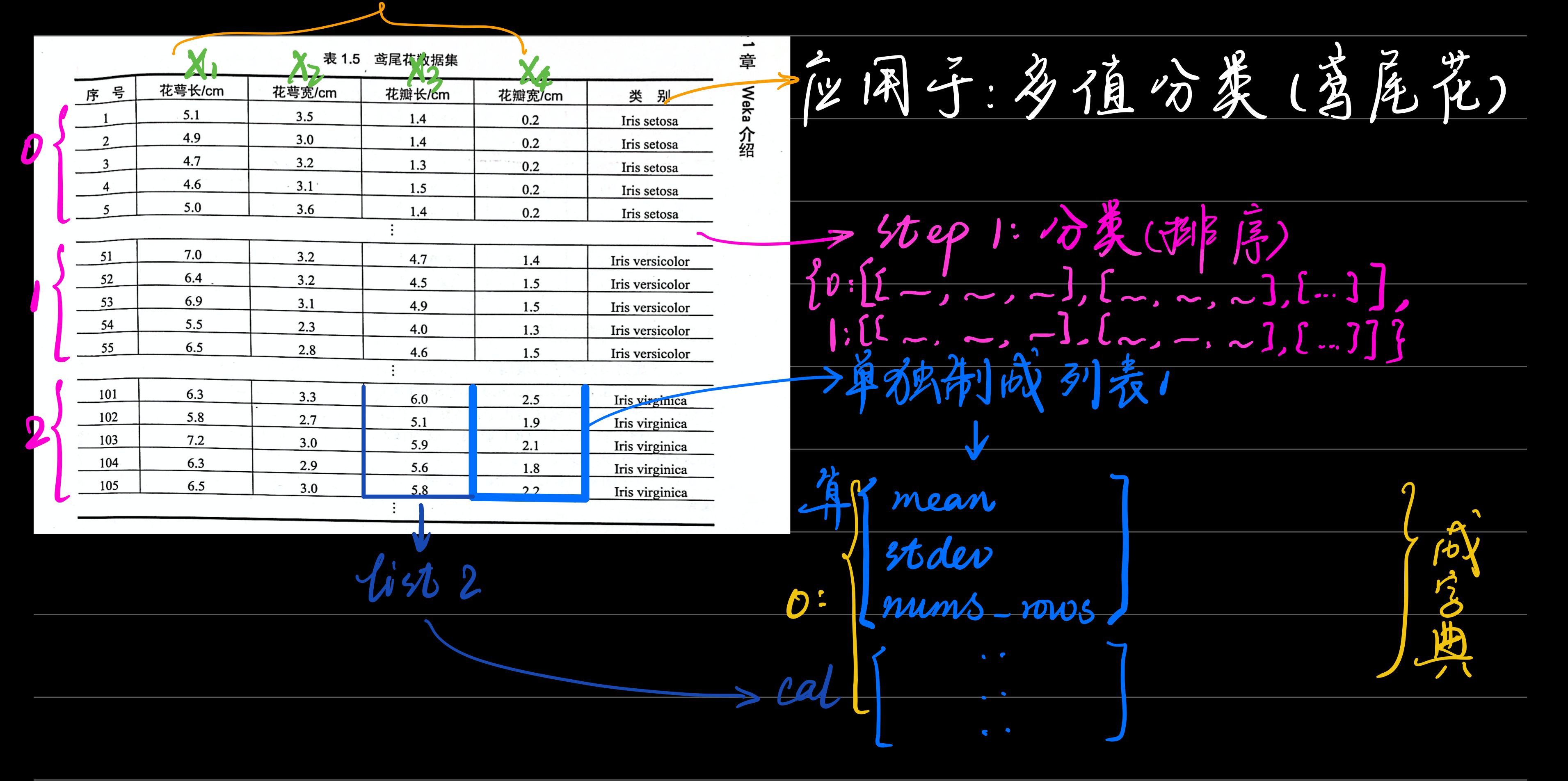
1. 首先需要知道贝叶斯理论：

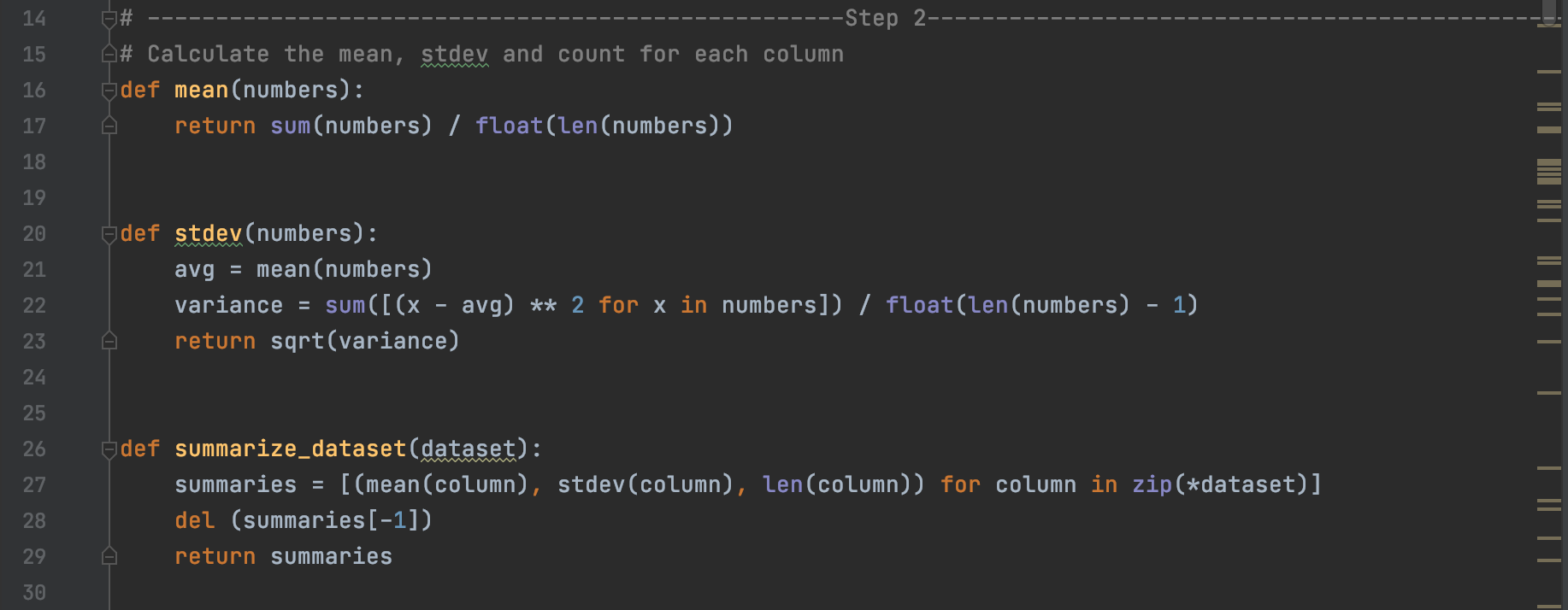
P(class|data) = (P(data|class) \* P(class)) / P(data)

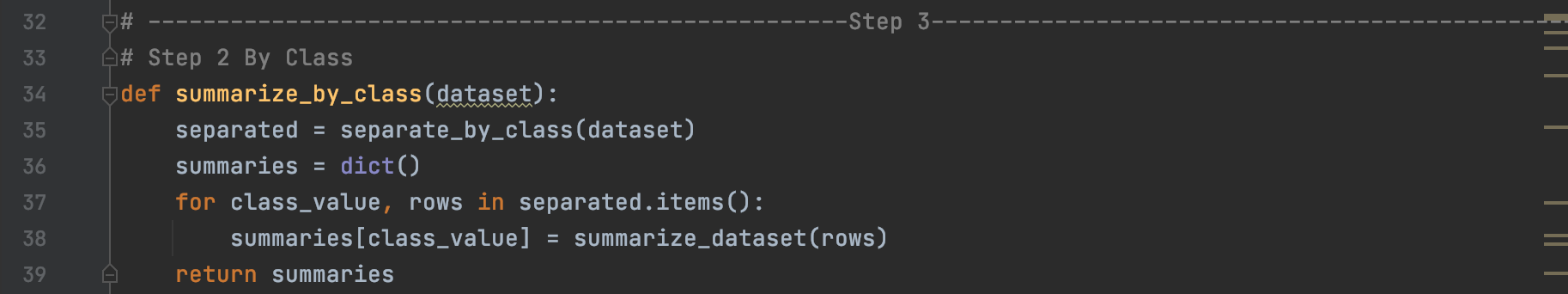
1. 先将整个数据集中的数据按鸢尾花所属的类别划分，根据上面对数据集的简介，即分出三组。为实现方便，为第一类鸢尾花设置种类标号“0”，第二类设“1”，第三类设“2”；利用Python中的字典类型，字典的键名为种类标号，键值为对应种类的所有实例（实例利用列表存放）：



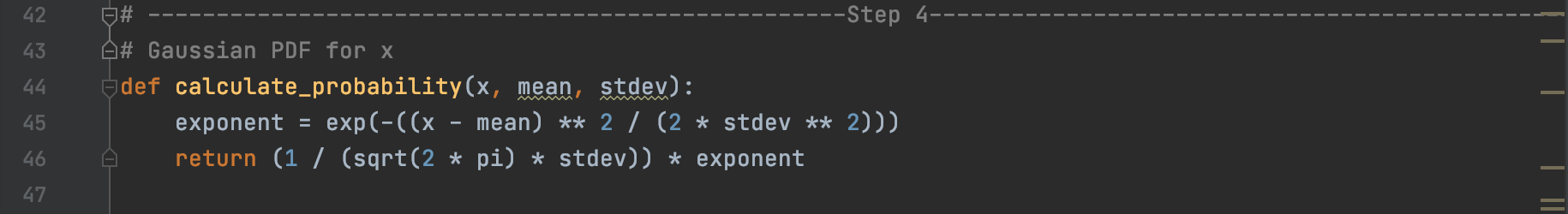
1. 接着编写用于计算所需统计量的函数，对每一个花的种类的每一个属性值，都需要计算该属性值下所有数据的平均值和标准差。对于某一种花，将它某一属性下数据的均值和标准差连同实例个数（50）存放于一个列表，每一个属性均遵此操作。之后将这些列表（每一种花应有4个列表，见下图示）作为键值，以花的种类标号作为键名，制成元组存放。如下：







1. 我们假设花的四个观测指标均服从某个分布，这里方便起见认为服从正态分布。这样我们就可以通过概率密度函数由前面算出的均值和标准差计算出它的概率：

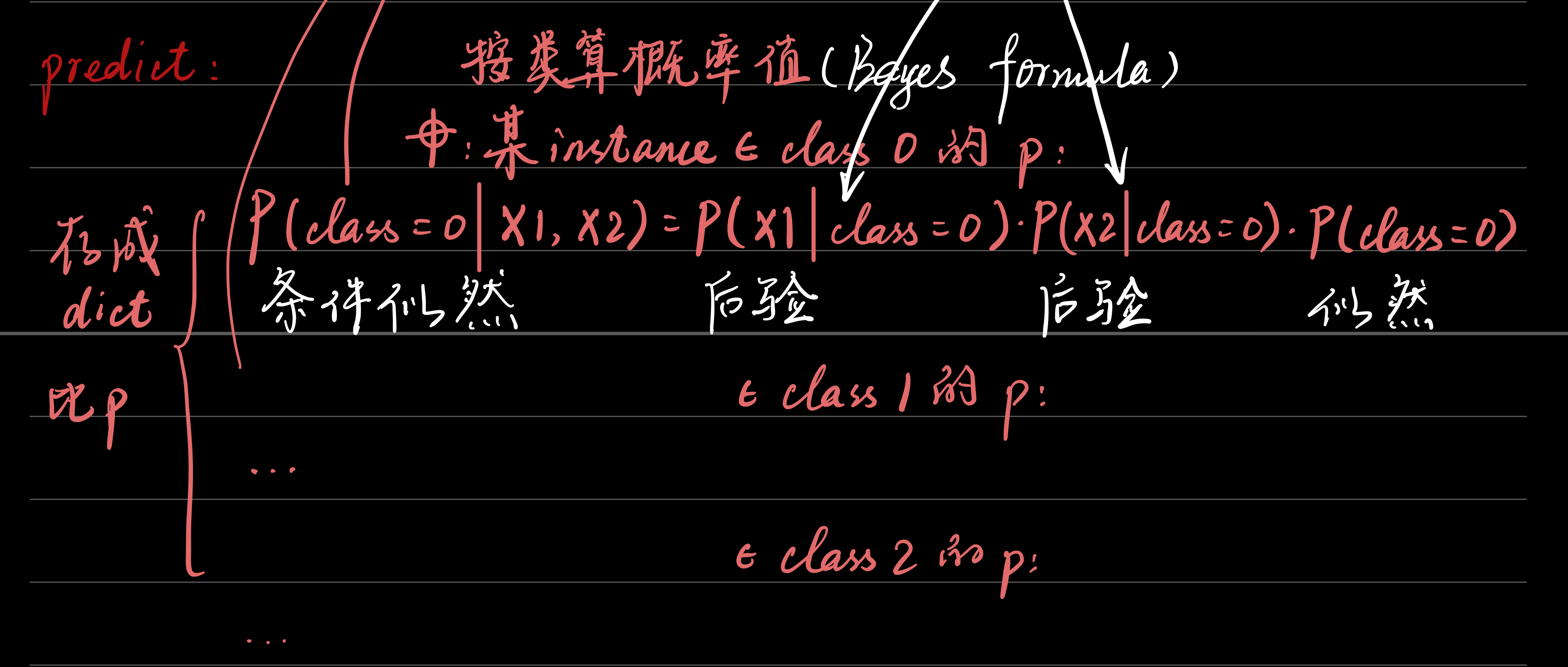


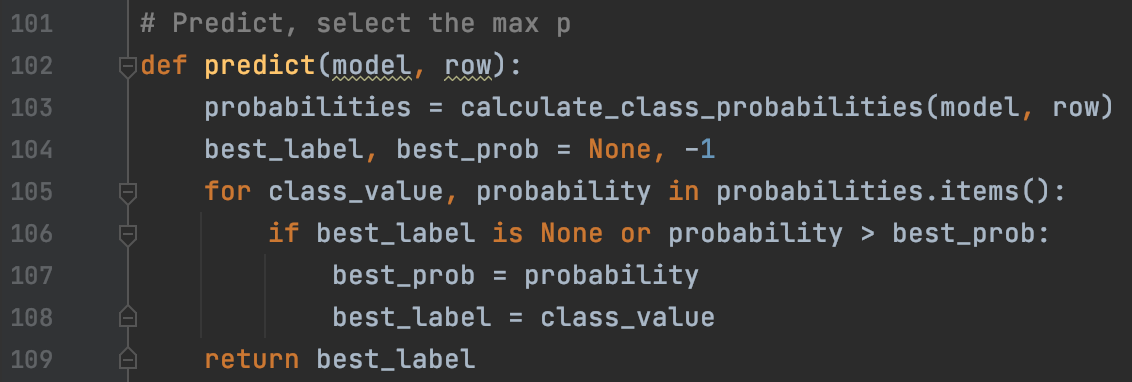
1. 接下来应用模型：对于每一个新观察的花朵，要分别计算它隶属于每一个花的种类的概率：

P(class|data) = P(X|class) \* P(class)

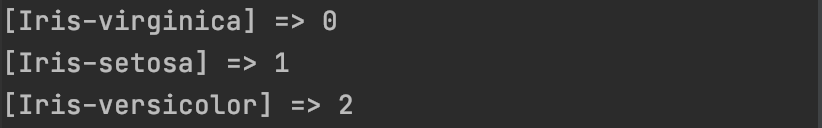
P(class=0|X1,X2) = P(X1|class=0) \* P(X2|class=0) \* P(class=0)

我们从中选取最大的概率值所对应的种类作为这朵新花的预测种类，如下：



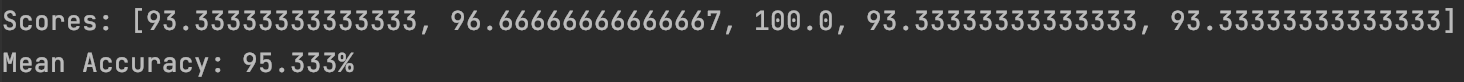








1. 最后应该利用多折交叉验证（这里选取5折，也即1折有150/5=30个观测值）来评估算法质量：



此为每一折上的准确率和5个准确率的均值。

实验思考：

1. 如果各属性不是条件独立的，则如何进行分类。

其实也可以依然使用朴素贝叶斯方法（Domingos and Pazzani 1997）。理由是模型足够简洁，因此相对几乎不会发生过拟合问题。

实验体会：

在上学期的机器学习课程中已经利用一些工具直接使用过该算法，现在我知道了算法背后的原理，可以帮助我在应用时更好地理解结果和参数的意义。

感谢吴老师的悉心指导。