《机器学习》读书笔记

1. 算法概述（算法名称及原理）

naïve-bayes

前三周学习的KNN算法和决策树算法，数据实例最终被明确的划分到某个分类中，而这次所学的naïve-bayes算法不能完全确定数据实例应该划分到哪个类别，只能给数据实例属于给定分类的概率。

朴素的含义就是假设属性之间相互独立

其原理主要是贝叶斯定理P(A|B)=P(A)×P(B|A)/P(B)

用文字表述：后验概率=先验概率×相似度/标准化常量，而朴素贝叶斯算法要解决的问题就是如何求出相似度即：P(B|A)的值[属性/特征]，同时它的一个局限在于它必须知道先验概率的值.

朴素贝叶斯分类器通常有三种分类方式，一种是基于伯努利模型实现，一种是基于多项式模型实现，还有一种是基于高斯模型实现

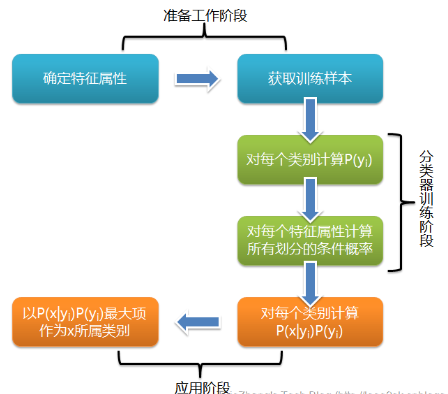
如果样本特征的分布大部分是连续值，使用GaussianNB会比较好。

如果样本特征的分大部分是多元离散值，使用MultinomialNB比较合适。例如文本分类单词统计，以出现的次数作为特征值

如果样本特征是二元离散值或者很稀疏的多元离散值，应该使用BernoulliNB

1. 算法设计（流程图及主要分段代码，附详细代码注释）

朴素贝叶斯分类的流程：

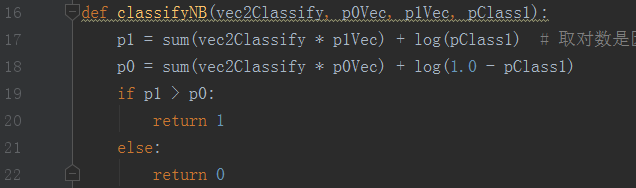


对于朴素贝叶斯来说

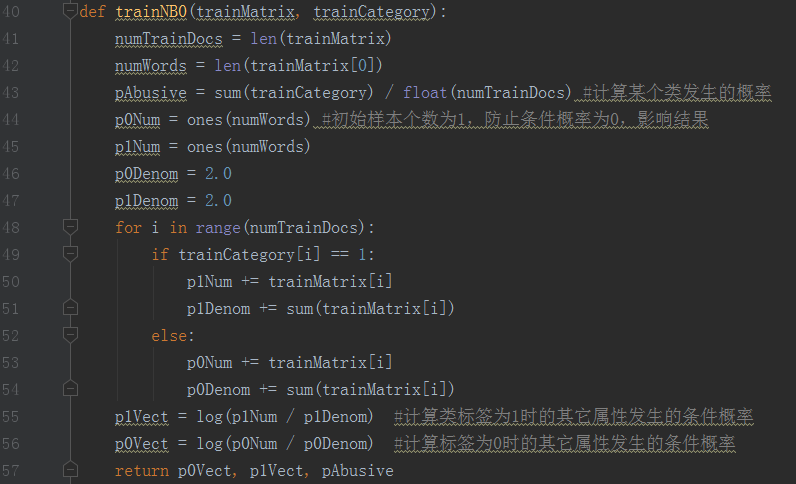
数据预处理工作

特征选择的时候不是选择所有的属性，一般选择属性之间无关的具有代表性的属性做实验，先做特征再分类

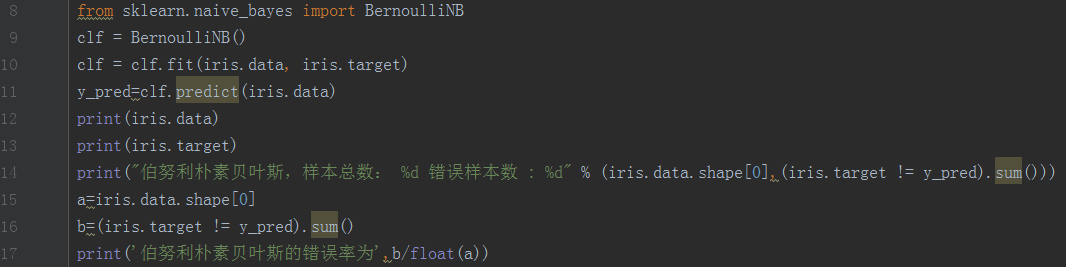
BernoulliNB相关算法代码



因为BernoulliNB是一个二类分类问题，所以可以使用互斥理论用1减去其中一个的概率即为剩下的概率



调包实现的BernoulliNB模型



1. 选用数据（数据集描述，包括来源，行数，列数，格式等）

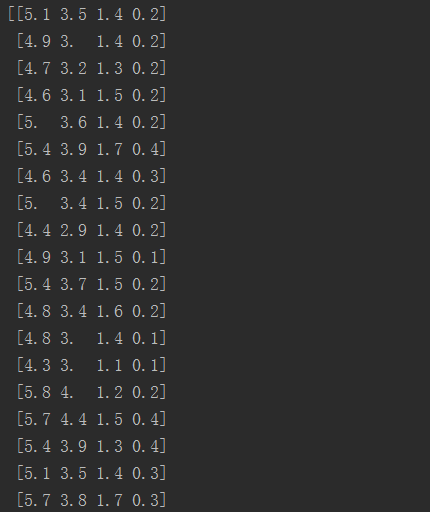
数据集1：鸢尾花数据集iris

150行表示150条数据

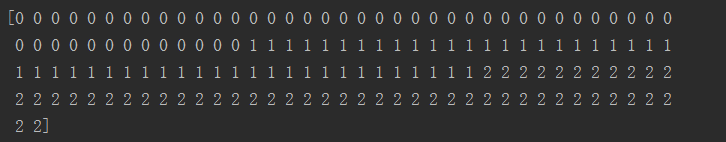
4列表示属性即特征

类别有三种，分别用0，1，2来表示setosa,versicolor和Virginia

部分数据展示



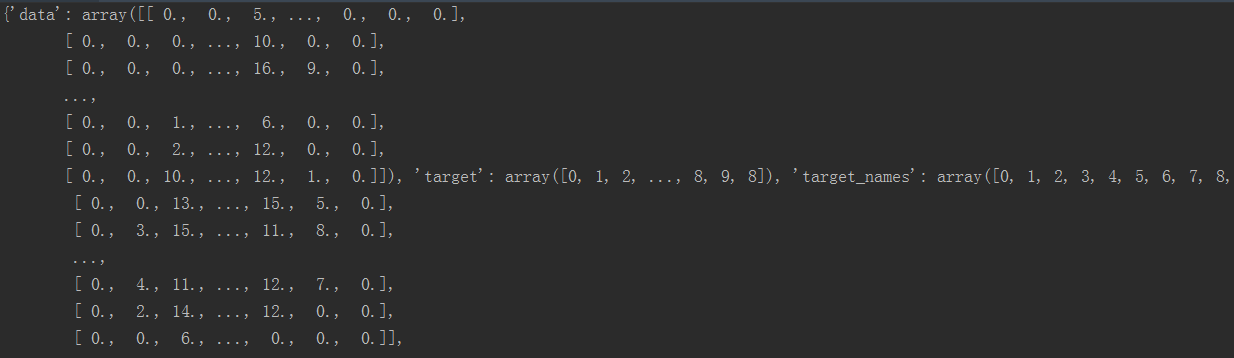
类别展示



数据集2:digits

行数：1797

列数10

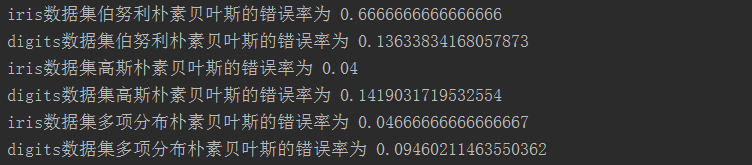


1. 评价方法（说明训练集和测试集分配方法及评价指标）

评价方法：十次十折交叉验证

评价指标：通过十次错误率求平均并比较大小值，值越小效果越好

1. 实验结果截图



1. 实验结果分析及比较

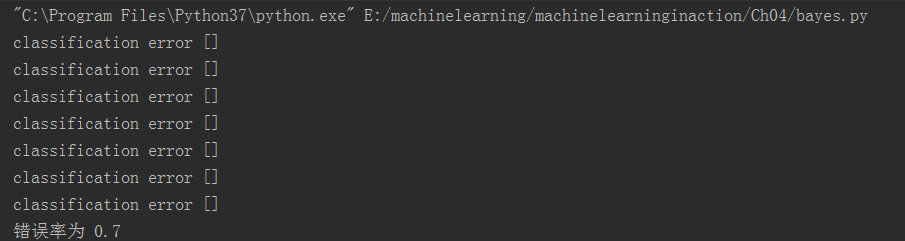
朴素贝叶斯对于数据的类型要求很高，对于先验数据为连续性的数据，一般使用高斯模型，从上述的结果也能看出高斯模型对于连续性的数据错误率相比于其他两种模型来说较小

1. 遇到的问题及解决方法，实践心得

朴素贝叶斯的缺陷之一在于，各个特征之间要相互独立，但是很多属性之间都有着或多或少的相关性，所以在实验之前要对属性进行筛选，选择独立性相对较高的并且具有代表性的属性进行实验

由上面的实验结果可以看出不同的数据集，不同的模型都会对实验结果产生很大的影响，所以使用某种算法不能一概而论，在这里我遇到的问题一是换不同的数据集会产生一系列问题，有的不适用于某种模型，本次实验只运用了两个数据集，还不能看出根本问题所在，后面我将根据新学的算法对不同的模型以及数据集进行整合比较，得出一个更加客观的结果

问题二：



为什么wordList[]列表中的元素不显示以及为什么错误率这么大？

错误原因：环境问题，python2.x的代码在python3.x上面不兼容，解决办法：可以使用python中自带的包2to3.py将2版本的代码转换为3.x的代码

编译过程中遇到的问题三：

UnicodeDecodeError: 'gbk' codec can't decode byte 0xae in position 199: illegal multibyte sequence

解决办法：可能是文档中出现了识别不了的字符，可以手动跳转到出现问题的地方修改编码

问题四：

TypeError: 'range' object doesn't support item deletion

解决办法：在python3.x中 ，range返回的是range对象，不返回数组对象，所以需要在前面加上list, 把 trainingSet = range(50) 改为 trainingSet = list(range(50))

建议：

每次使用三个数据集，iris，wine，car