**《大数据分析》第四次实验报告**

201250070郁博文

1. **朴素贝叶斯**
2. **朴素贝叶斯简介**

朴素贝叶斯是一种常用的分类方法。它是基于贝叶斯定理，依据条件概率公式，计算当前特征的样本属于某个分类的概率，选最大的概率分类。从名字就可以看出，朴素贝叶斯方法作了很大的条件简化，它假设各个特征之间相互独立，因此极大地简化了分类问题。

1. **个人理解**

在理解了条件概率的情况下，朴素贝叶斯算法实际上十分简单。因为它认为各个特征属性是条件独立的。所以，给定一个训练数据集，要使用贝叶斯算法进行分类，我们首先需要计算每个属性在每种类别下的条件概率。对于测试集数据来说，给定一个未知分类的数据，我们假定它是训练数据中的某个类别，在该类别下，计算在该类别下各个特征属性的条件概率估计。完成了对所有类别的概率计算后，我们比较这些条件概率估计，选取最大的作为分类结果。这就是朴素贝叶斯的算法过程，恰如它的名字一般朴素。

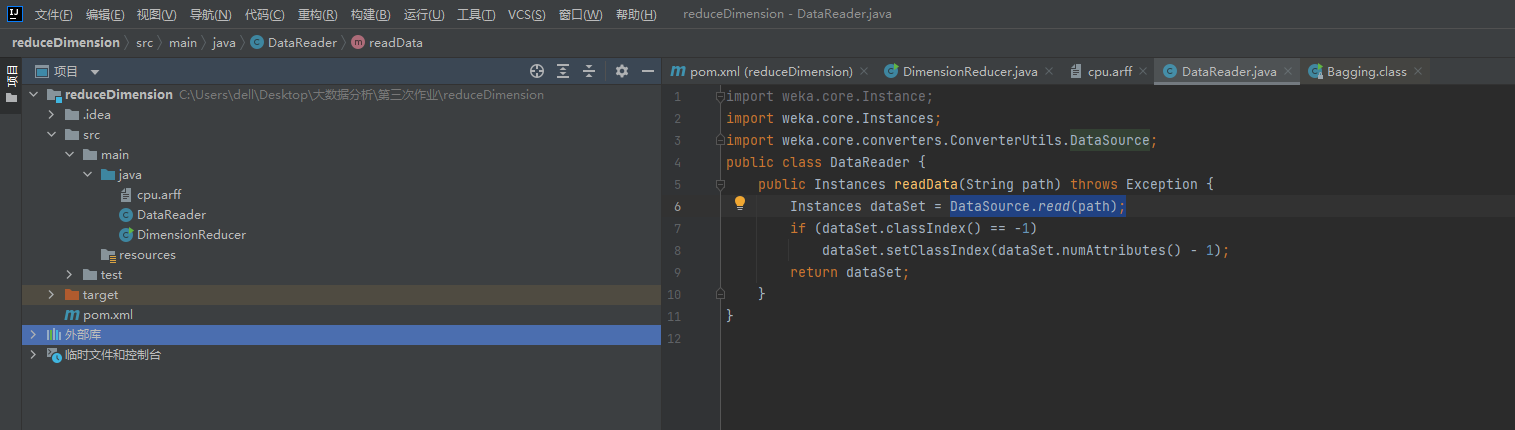
值得注意的是，朴素贝叶斯算法有一个bug。当某个特征在某个类别中没有出现时，该特征在该类别下的条件概率就等于0.这使得在预测分类时，如果某个实例中出现了这个特征，那么这个实例在这个类别下的条件概率估计就会等于0.这显然不合逻辑。因此，在进行朴素贝叶斯算法分类之前，可以引入拉普拉斯平滑方法，将每个特征的频数加一，这样就不会出现这样的bug.

1. **实验结果**

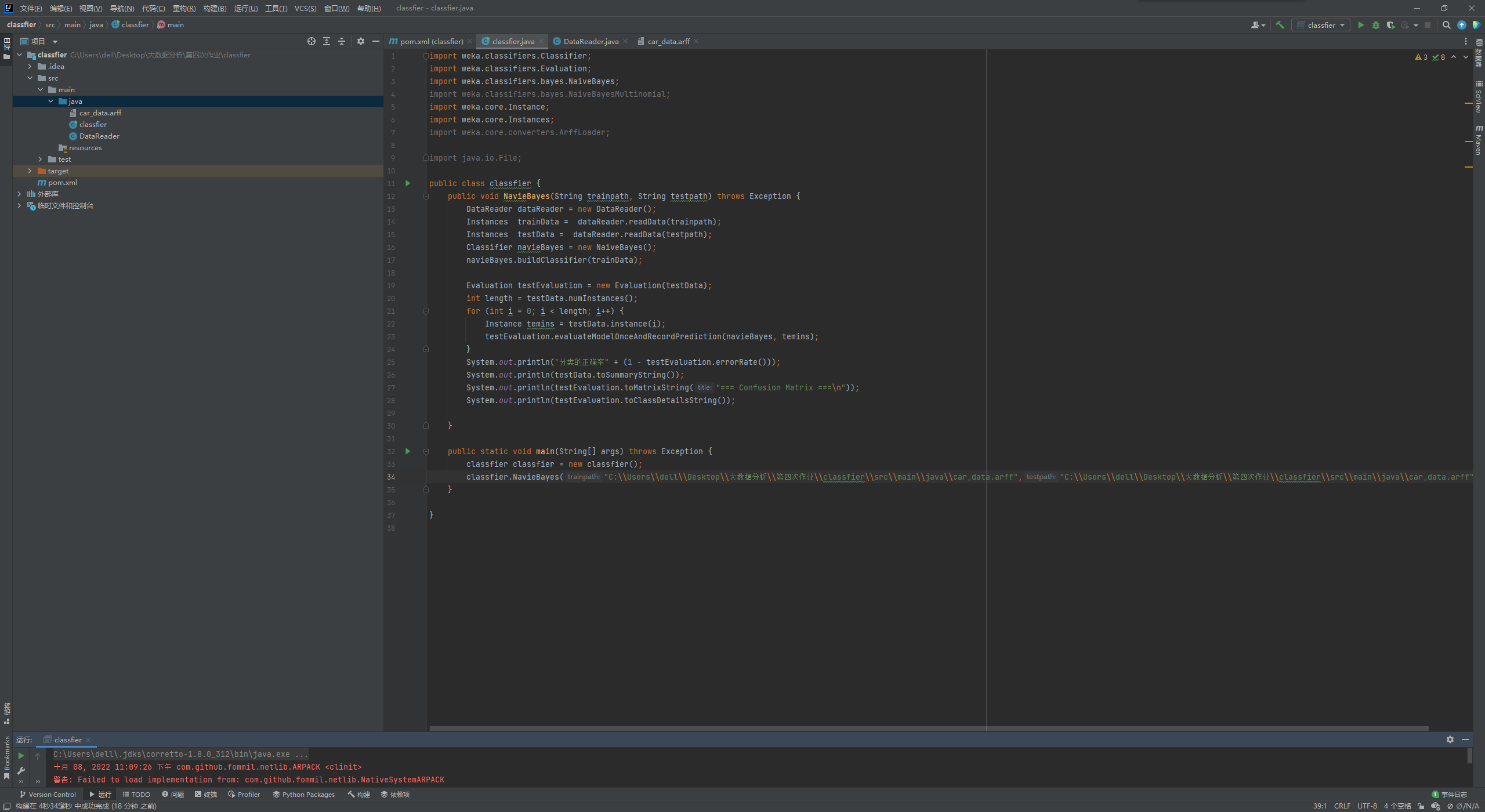
编写的实验代码如下：

这个类封装了数据集文件的读写

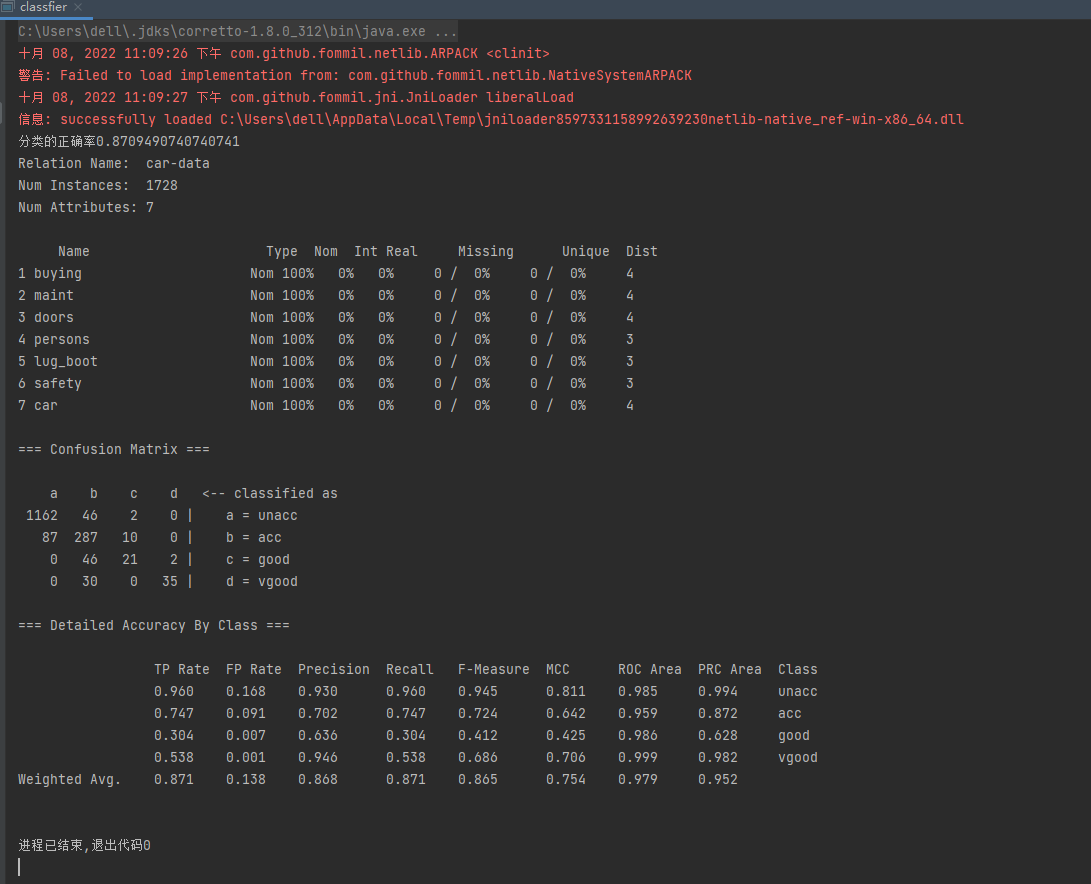
原数据集的最后一个属性是用于分类的，因此读入数据时需要特殊标记，否则会对结果产生影响。



这个类是朴素贝叶斯的实现。



运行结果如下：



第一行是分类正确率，第二行开始是weka的朴素贝叶斯方法自带的总结