

**模式识别大作业**

题 目 Glass数据集分类

学 院 信息科学与工程

专 业 控制科学与工程

组 员 朱雪婷

指导教师 赵海涛

**完成日期： 2018 年 10 月22日**

**模式识别作业报告——Glass数据集分类**

组员：朱雪婷

经过模式识别课程的学习，在赵海涛老师的辛勤指导下，我对模式识别的贝叶斯算法有了一定的了解，并通过本次针对Glass数据集的评分预测实验来巩固所学内容。下面将详细说明我的解决过程。

**一、Glass数据集简介**

glass数据集是matlab自带的分类实验数据集，是一类多重变量分析的数据集。数据集包含214个数据，分为2类，第一类有163个数据，第二类有51个数据，每个数据包含9个属性，不同维度的数据相互独立。

**二、贝叶斯决策**

**2.1贝叶斯公式**

贝叶斯决策就是在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策。[贝叶斯决策理论](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%86%B3%E7%AD%96%E7%90%86%E8%AE%BA)方法是[统计模型](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%A8%A1%E5%9E%8B/7492984)决策中的一个基本方法，其基本思想是：首先已知类条件[概率密度](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6)参数表达式和[先验概率](https://baike.baidu.com/item/%E5%85%88%E9%AA%8C%E6%A6%82%E7%8E%87)，然后利用[贝叶斯公式](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%85%AC%E5%BC%8F)转换成[后验概率](https://baike.baidu.com/item/%E5%90%8E%E9%AA%8C%E6%A6%82%E7%8E%87)，最后根据后验概率大小进行决策分类。

在没有对样本进行任何观测情况下的概率，我们称为先验概率。通常在进行分类决策时，我们都可以对样本的某个特性进行观测，从而获得样本特性的数值，以此作为判断的依据，这样就可以提高决策的正确率，这种概率称作后验概率，记作𝑃(𝜔|𝑥)。

对于后验概率的求解，就是运用了概率论中的贝叶斯公式，即：

其中是关于x的似然函数（likelihood），p(x)称为证据因子（evidence），是一个标量因子，保证各类的后验概率和为1。

p(x)=

由于证据因子是标量，且不同类别的证据因子是一致的，所以可以得出：

若pP>pP，则x∈;反之，x∈

**2.2朴素贝叶斯**

朴素贝叶斯方法是用一个非常简单的方式来处理然p的估计问题。其思路是简洁的，即假设x∈的不同维度相互独立，也就是

p

这样p的估计就转化为𝐷个相互独立的一维随机变量的密度估计问题，“维数灾难”问题得到了缓解。

整个朴素贝叶斯分类分为三个阶段：

1.准备工作阶段，任务是为朴素贝叶斯分类做必要的准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出是特征属性和训练样本。这一阶段是整个朴素贝叶斯分类中唯一需要人工完成的阶段，其质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。

2.分类器训练阶段。这个阶段的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据前面讨论的公式可以由程序自动计算完成。

3.应用阶段。这个阶段的任务是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。这一阶段也是机械性阶段，由程序完成。

**三、整体解决方案**

**3.1方案分析**

glass数据集包含214个数据集，分为2类，第一类有163个数据，第二类有51个数据，每个数据包含9个属性。将glass\_dataset拆分为两个矩阵：traindata和testdata;因为两类数据数目差别很大，所以第一类测试数据多一些，将第一类数据的后50列用于测试，第二类数据的后20列用于测试，归到testdata,其余用于训练，归到traindata。即共有70个数据用于测试，144个用于训练，并假设9维特征中的每一维都为正态分布，分别求取均值及方差。最后通过散点图展示分类效果。

**3.2程序框图**

读取glass数据

traindata

testdata

求取每个变量的均值和方差

求取每个数据在每类的后验概率

将后验概率大的归入这一类中

判断是否归入正确的类中，计算正确率

画图

**3.2主要程序代码**

假设每一维数据都为正态分布，计算训练数据的均值和方差。代码如下：

for i=1:9

[mu(i,1),sigma(i,1)]=normfit(traindata1(i,:));

[mu(i,2),sigma(i,2)]=normfit(traindata2(i,:));

end

对测试数据进行朴素贝叶斯估计，假设每一维数据独立，计算每个数据的后验概率，并判断数据在哪一类中的概率最大，并归入这一类中。代码如下：

posterior=zeros(2,70);

for i=1:70

for j=1:2

if j==1

priori(j)=163/214;

elseif j==2

priori(j)=51/214;

end

likelihood(j)=1;

for d=1:9

likelihood(j)=likelihood(j)\*normpdf(testdata(d,i),mu(d,j),sigma(d,j));

end

posterior(j,i)=likelihood(j)\*priori(j);

end

[c,s]=max(posterior(:,i)');

category(i)=s;

end

判断是否在正确的类别中，并计算正确率和错误率。代码如下：

category\_true1=0;

category\_false1=0;

category\_true2=0;

category\_false2=0;

for k=1:50

if category(k)==1

category\_true1=category\_true1+1;

else

category\_false1=category\_false1+1;

end

end

for k=51:70

if category(k)==2

category\_true2=category\_true2+1;

else

category\_false2=category\_false2+1;

end

end

category\_false=category\_false1+category\_false2;

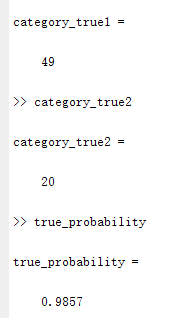
category\_true=category\_true1+category\_true2;

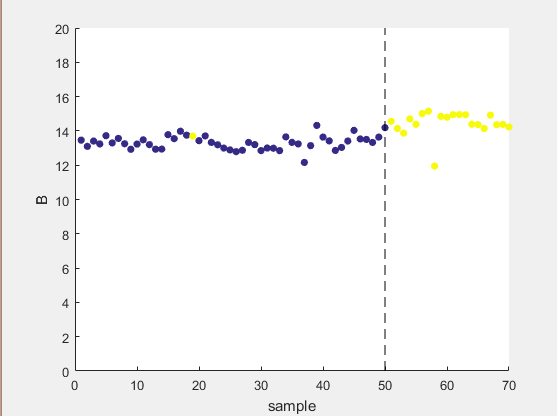
false\_probability=category\_false/70;

true\_probability=category\_true/70;

**3.3运行结果**

朴素贝叶斯法具有较好的分类精度。在我们假设各变量之间不存在相关性的前提下，本算法可以准确地将绝大多数样本进行分类，第一类数据有一个数据被分错，第二类数据全部被分对。最终的正确率为0.9857。



 任取一维图片，第一类的点为蓝色，第二类的点为黄色，可以清楚的看出第一类中有一个点被分错了类别，第二类中的点都是分对的。

**四、作业总结**

通过模式识别课程的学习，我学到了很多解决实际问题的方法，但是因为编程能力比较弱，我选择了比较简单的数据集做大作业，通过编程，我对朴素贝叶斯有了更加清晰的理解，之后我会集中学习每一个课上讲到的方法并提高自己的编程能力，将这些算法真正的掌握。