華東理工大學

模式识别大作业

题　　目　小麦种子数据集分类

学　　院　信息科学与工程学院

专　　业　控制科学与工程

组　　员　徐泽铭

指导教师　赵海涛

完成日期：　2018 年 10 月 25 日

模式识别作业报告——用朴素贝叶斯进行数据集分类

姓名：徐泽铭　　学号：Y30180685

1　报告综述

因本人的编程水平有限，对于Python的学习刚起步，对模式识别知识的掌握也不够深入，因此就用比较基础的问题来完成此次作业。通过使用MATLAB编写程序对小麦种子数据集的数据进行分类测试，以体现朴素贝叶斯在分类问题方面的优势。数据来源于网络资源。

在报告中首先介绍所使用的分类方法——朴素贝叶斯的原理，然后展示MATLAB编写的代码，使用注释对程序进行解释，最后对实验结果进行展示，总结完成作业中遇到的问题和解决方法，以及我个人在完成报告过程中的心得体会。

2　算法原理

2.1　模式识别的分类概述

模式识别的应用广泛，而每个使用模式识别的人对于如何识别和分类使用的方法也不尽相同，一般的，按照所采用的模式表达方式，我们可以把模式识别领域中研究者已使用的识别方法分为：统计模式识别和结构模式识别。报告中使用的朴素贝叶斯分类就属于统计模式识别。统计模式识别，顾名思义，也可大概猜出其特点在于“统计”，采用特征向量作为输入，对应的识别工作在特征向量空间中进行，采用的识别方法的理论则是基于统计学。由于结构模式识别不是老师授课的重点，也不是本次报告的主要内容，因此这里就不多做介绍。除了上述分类之外，还可以根据用于分类的观测样本的类别属性是否已知将模式识别分类方法分为：有监督的模式分类方法和无监督的模式识别分类方法。报告中用到的朴素贝叶斯分类算法则属于有监督的模式识别方法。

2.2　朴素贝叶斯（Naive Bayesian）

在介绍朴素贝叶斯之前，我们首先了解一下贝叶斯分类的概念。贝叶斯分类是一系列分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类。朴素贝叶斯算法是其中应用最为广泛的分类算法之一。朴素贝叶斯的“朴素”在于其基于一个简单的假定：给定目标值时属性之间相互条件独立。1997 年，微软研究院的 Domingos 和 Pazzani 通过实验证

明，即使在其前提假设不成立的情况下，该分类器依然表现出良好的性能。对这一现象的一个解释是，该分类器需要训练的参数比较少，所以能够很好的避免发生过拟合（overfitting）。朴素贝叶斯分类的正式定义如下：

1、设$\mathrm{x}=\{a_1,a_2,\ldots,a_m\}$为一个待分类项，而每个 a 为 x 的一个特征属性；

2、有类别集合$\mathrm{C}=\{y_1,y_2,\ldots,y_n\}$；

3、计算$\mathrm{P}(y_1|x),P(y_2|x),\ldots,P(y_n|x)$；

4、如果$\mathrm{P}(y_k|x)=max\{P(y_1|x),P(y_2|x),\ldots,P(y_n|x)\}$，则$\mathrm{x}\in y_k$ 。

那么现在的关键就是如何计算得到第 3 步中的各个条件概率。我们可以这么做：

1、找到一个已知分类的待分类项集合，这个集合叫做训练样本集；

2、统计得到在各类别下各个特征属性的条件概率估计，即$\mathrm{P}(x|y_i)$，$i\in\mathrm{n}$。

3、在朴素贝叶斯的假设下各个特征属性是条件独立的，则根据贝叶斯定理有如下推导：

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)}$$

因为分母对于所有类别为常数，因为我们只要将分子最大化即可。又因为各特征属性是条件独立的，所以有如下推导：

$$\mathrm{P}(x|y_i)P(y_i)=P(a_1|y_i)P(a_2|y_i)\ldots P(a_m|y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j|y_i)\,P(y_i)$$

到这一步，我们需要的值都求出来了，代入原公式即可求出在某个特征下，某一类发生的概率，用更直接的表示即为下面的表达式，这个表达方法是我在查阅资料的时候摘录下来的。

$$\mathrm{p}\left(类别\middle|特征\right)=\frac{p\left(特征\middle|类别\right)p\left(类别\right)}{p\left(特征\right)}$$

在特征向量中的每个特点的概率我们都可以通过极大似然估计(maximum-likelihood estimate)来求得，也就是简单地求某个特征在某个类别中的频率，公式如下：

$$\mathrm{P}(a_j|y_i)=\frac{N_{(a_j|y_i)}}{N_{y_i}},(j=1,2,\ldots,m)$$

其中，分子代表所有属于类别y_i的样本中，特征a_j出现的次数；分母代表属于类别y_i的

样本中，所有特征出现的次数。

基于以上推导，结合查阅的论坛里有关朴素贝叶斯的笔记，总结如下：

朴素贝叶斯的优点在于算法逻辑简单，易于实现；分类过程中占用的存储空间小（假设特征相互独立，只会涉及到二维存储）。但它也存在缺点：理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率，但是实际上并非总是如此，这是因为朴素贝叶斯模型假设属性之间相互独立，在属性相关性较小时，朴素贝叶斯性能最为良好。但考虑到这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

3 数据结构

报告所用的测试数据集的部分数据如下表所示，小麦种子数据集共有 210*7 维，最后需划分为三类，其 1-7 列分别代表小麦种子的 7 个评估标准，分别为：区域、周长、压实度、籽粒长度、籽粒宽度、不对称系数、籽粒腹沟长度。根据数据集里提供的分类标准，前 70 维为第一类，中间 70 维为第二类，最后 70 维为第三类。

表 1　小麦种子数据集部分数据

	1	2	3	4	5	6	7
1	15.2600	14.8400	0.8710	5.7630	3.3120	2.2210	5.2200
2	14.8800	14.5700	0.8811	5.5540	3.3330	1.0180	4.9560
3	14.2900	14.0900	0.9050	5.2910	3.3370	2.6990	4.8250
4	13.8400	13.9400	0.8955	5.3240	3.3790	2.2590	4.8050
5	16.1400	14.99000	0.9034	5.6580	3.5620	1.3550	5.1750
6	14.3800	14.2100	0.8951	5.3860	3.3120	2.4620	4.9560
7	14.6900	14.4900	0.8799	5.5630	3.2590	3.5860	5.2190

4 算法实现

使用朴素贝叶斯实现小麦种子分类的算法如下所示：

```
clear
clc
load wheatseeds.mat   %共 210*7 维数据
%wheatseeds.mat 可以分为 3 类(1,2,3);
%第一类为第 1-70 行数据，第二类为第 71-140 行数据，第三类为第 141-210 行数据;
%将 wheatseeds.mat 的数据分成两个矩阵：训练数据 trdata 和测试数据 tedata;
%trdata 包含每类数据的前 40 行;tedata 包含每类数据的后 30 行；
%数据分类如下：

trdata1=wheatseeds(1:40,1:7);
tedata1=wheatseeds(41:70,1:7);
trdata2=wheatseeds(71:110,1:7);
tedata2=wheatseeds(111:140,1:7);
trdata3=wheatseeds(141:180,1:7);
tedata3=wheatseeds(181:210,1:7);
trdata=[trdata1;trdata2;trdata3];
tedata=[tedata1;tedata2;tedata3];
%假设 wheatseeds.mat 中每一类、每一维都服从正态分布;
%分别计算 3 类种子的 7 种评估标准的均值和方差;
mean=zeros(3,7);       %均值
sigma=zeros(3,7);      %方差
for i=1:7
    [mean(1,i),sigma(1,i)]=normfit(trdata1(:,i));
    [mean(2,i),sigma(2,i)]=normfit(trdata2(:,i));
    [mean(3,i),sigma(3,i)]=normfit(trdata3(:,i));
end
```

```
%用朴素贝叶斯计算似然函数;
posterior=zeros(3,90); %后验概率
priori=zeros(3,1);        %先验概率
class=zeros(90,1);
for i=1:90
    for j=1:3
        if j==1
            priori(j,1)=70/210;   %类别 1 的先验概率
        elseif j==2
            priori(j,1)=70/210;   %类别 2 的先验概率
        else
            priori(j,1)=70/210;   %类别 3 的先验概率
        end
        likelihood=ones(3,1);     %计算似然函数
        for d=1:7
            likelihood(j,1)=likelihood(j,1)*normpdf(tedata(i,d),mean(j,d),sigma(j,d));
        end
        posterior(j,i)=likelihood(j,1)*priori(j,1);   %计算后验概率
    end
    C=posterior(:,i);
    [m,n]=max(C);     %选择最大后验概率对应的类别
    class(i,1)=n;         %对 90 个测试数据分类，产生 90*1 的矩阵
end
```

5 结果展示

朴素贝叶斯分类结果如下图所示：

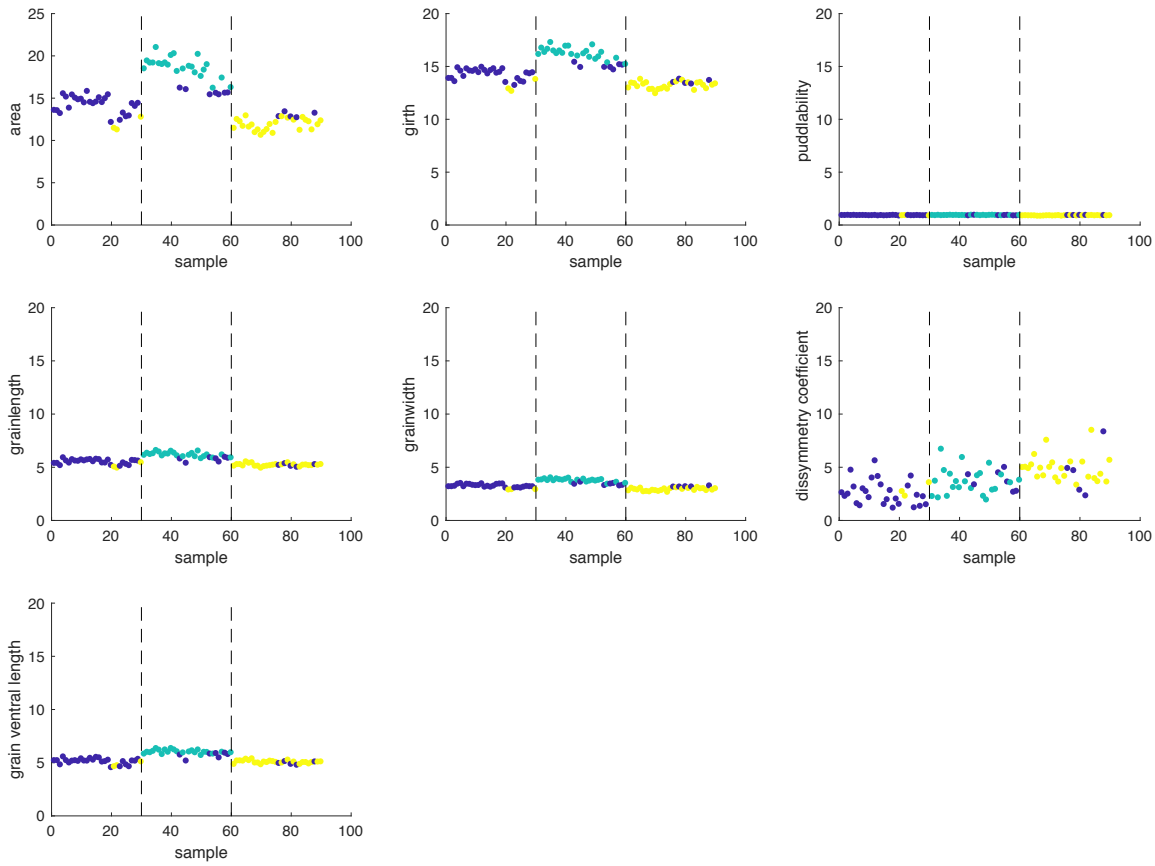

图 1 小麦种子数据的 7 维散点图

从图上可以看出，分类的误差主要出现在第一类数据分类上，本来位于第二和第三两类的数据点却被识别为了第一类数据，因此，如何提高识别率降低识别误差是一个改进方向，可以从程序上改进，也可以改进识别方法。总的来说，使用朴素贝叶斯对数据分类的正确率还是有一定的保障，经典方法毕竟是经典。但是这种分类器的准确率，还是要依赖于训练的样本，借用在查阅资料的过程中看到的一句话：朴素贝叶斯分类器其实是比较依赖于训练语料的，机器学习算法就和纯洁的小孩一样，取决于其成长（训练）条件，“吃的是草挤的是奶”，但，“不是所有的牛奶，都叫特仑苏”。

附上前三维画图程序，如下所示：

```
%作图展示在 7 种标准下数据的分类情况;
%第一类数据为紫色，第二类数据为绿色，第三类数据为黄色;
figure;
subplot(3,3,1);
hold on
scatter(1:90,tedata(:,1),20,class,'filled');
plot([30,30],[0,25],'--k');     %分割线
plot([60,60],[0,25],'--k');
xlabel('sample');
ylabel('area');                  %标准 1：区域
hold off
subplot(3,3,2);
hold on
scatter(1:90,tedata(:,2),20,class,'filled');
plot([30,30],[0,20],'--k');     %分割线
plot([60,60],[0,20],'--k');
xlabel('sample');
ylabel('girth');                 %标准 2：周长
subplot(3,3,3);
hold on
scatter(1:90,tedata(:,3),20,class,'filled');
plot([30,30],[0,20],'--k');     %分割线
plot([60,60],[0,20],'--k');
xlabel('sample');
ylabel('puddlability');       %标准 3：压实度
hold off
```

6 报告总结

程序方面有受到课程笔记的启发，也请教了同学，对于我这种编程入门选手来说，将理论知识变成实际应用的程序，着实有些摸不着头脑，在此要感谢同学对我的关爱。其实完成编程得出较好结果的时候，内心还是比较欣喜的，为自己完成了作业高兴，也为自己又学习了知识高兴。同时也体会到了老师上课时劝告我们的含义，在模式识别上花费的功夫仅仅是每星期两节课是远远不够的，必须多看多思考，博采众长，才能找到适合自己的学习方法，提升自己的能力。

在编写程序过程中，遇到的大大小小的问题不少，在数据的选择上就费了一些功夫，因为避开了 MATLAB 里自带的测试数据集，但是付出都是值得的，通过查阅网上的资料，我也了解了一些模式识别常用于测试的数据集，在课余时间可以拿出来再练练手。由于对数据结构理解的不够透彻，导致分类矩阵在维数上不匹配，没办法作出最后的图像。后来经过反复的研究数据，理解数据与理论知识的联系，请教同学以及在网上搜索 MATLAB 里一些基本函数的含义和作图的教程，最终解决了数据维数的问题，得到了最终结果。另外其实应该有一段计算准确率的程序，我觉得不是要展示的重点，就没有放进报告里。

不断查阅资料的过程中，我对朴素贝叶斯也有了更清晰的认识，更加体会到了“学无止境”对于模式识别的学习，甚至于对任何一种知识的学习都是没有止境的，人生也是因为不断学习，去体验，去感悟，才变得充实。愿我常怀求知之心，不忘初衷，不负时光。