

山东大学机器学习课程

实验报告

——实验二：贝叶斯分类器的设计与实现

姓名：刘梦源

学院：计算机科学与技术学院

班级：计算机14.4

学号：201400301007

**一、实验目的：**

（1）设计贝叶斯分类器

（2）体会监督学习的思想，理解训练与训练误差等概念

（3）根据已给数据集，用贝叶斯分类器实现分类，并绘制图像，找出误分点，讨论训练误差的影响因素。

**二、实验环境：**

（1）硬件环境：

英特尔® 酷睿™ i7-7500U 处理器

512 GB PCIe® NVMe™ M.2 SSD

8 GB LPDDR3-1866 SDRAM

（2）软件环境：

Windows10家庭版64位操作系统

Matlab R2016a

**三、实验内容**

**（1）贝叶斯分类器**

贝叶斯是一种基于概率的学习[算法](http://lib.csdn.net/base/datastructure" \o "算法与数据结构知识库" \t "_blank)，能够用来计算显式的假设概率，它基于假设的先验概率，给定假设下观察到不同数据的概率以及观察到的数据本身。

  我们用P(h)表示没有训练样本数据前假设h拥有的初始概率，也就称为h的先验概率，它反映了我们所拥有的关于h是一个正确假设的机会的背景知识。当然如果没有这个先验知识的话，在实际处理中，我们可以简单地将每一种假设都赋给一个相同的概率。类似，P(D)代表将要观察的训练样本数据D的先验概率（也就是说，在没有确定某一个假设成立时D的概率）。然后是P(D/h)，它表示假设h成立时观察到数据D的概率。在机器学习中，我们感兴趣的是P(h/D),也就是给定了一个训练样本数据D,判断假设h成立的概率，这也称之为后验概率，它反映了在看到训练样本数据D后假设h成立的置信度。（注：后验概率p(h/D)反映了训练数据D的影响，而先验概率p(h)是独立于D的）。

 （1）

特别的，正态分布的判别函数可以化为：

 （2）

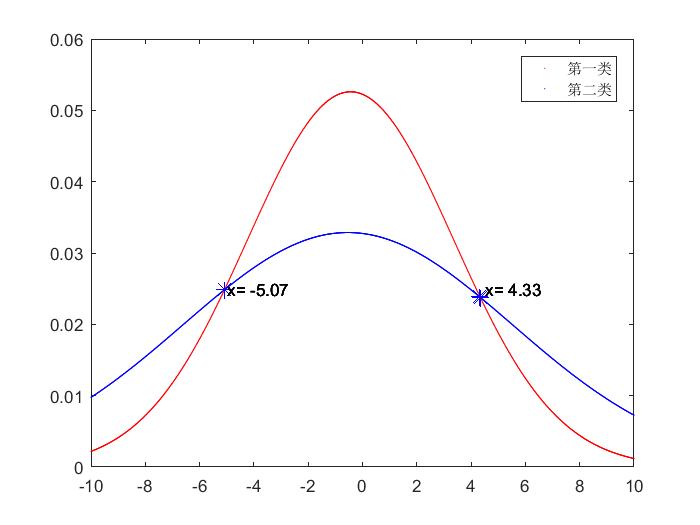
在连续性分布的问题中，我们可以对（2）进行改进，变成了

 （3）

其中，为***x***在第***i***类下的概率密度。

**（2）一类特征值下的两类分类问题**

用为特征值对和进行分类做出图像如图1所示



**图1** 一类特征值时的判别图

由图像可知，当x>4.33或者当x<-5.07时，应该判给第二类，当-5.07<x<4.33时，应该判给低一类。按照表格查找数据，可发现20个点中有6个判错了，分别是：第一类中的-5.43，4.94，-2.55；第二类中的-0.91，1.30，3.60。

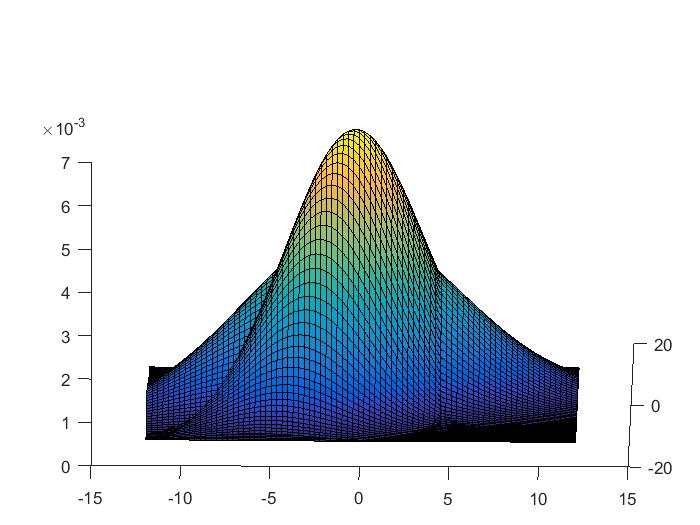
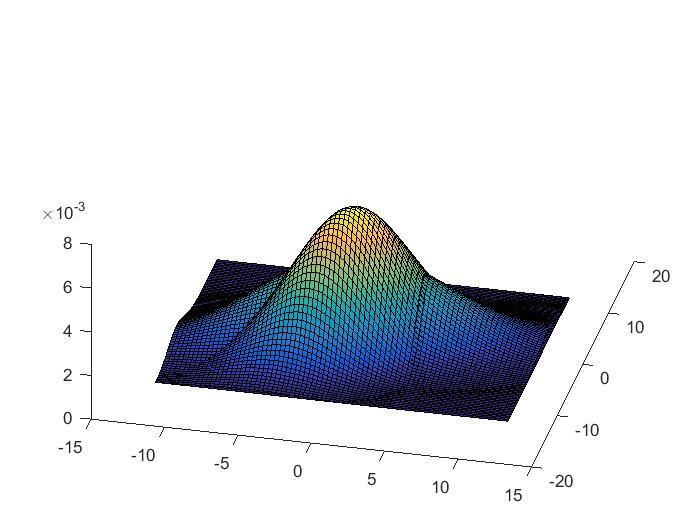
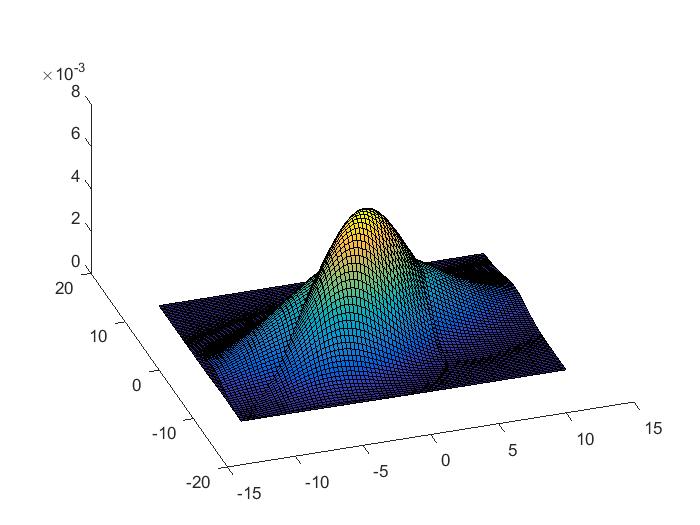
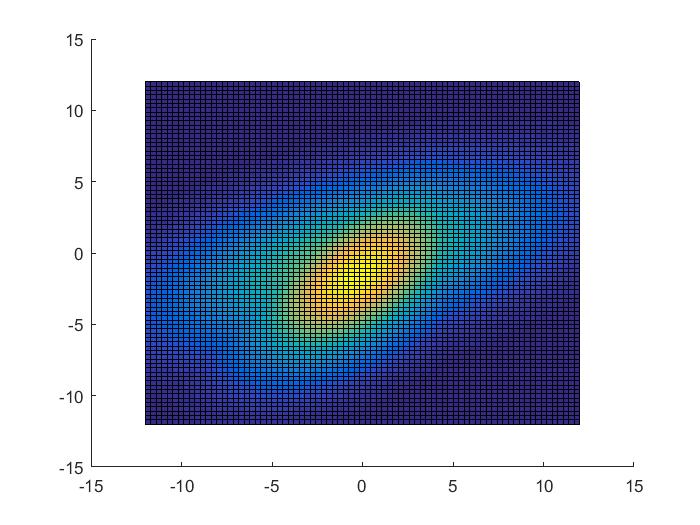
所以经验按照特征值进行分类，误分比为，百分比为30%，按照式（4），求得Bhattacharyya造成的误差为0.4740。

 （4）

至此，按照一类特征值的分类工作完成。

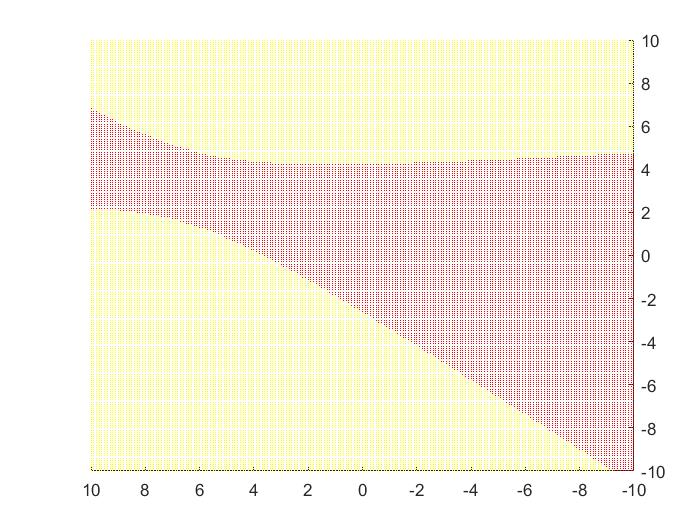
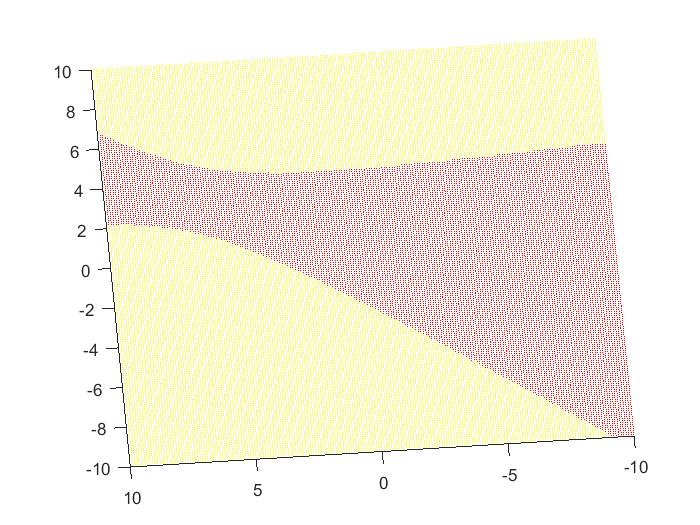
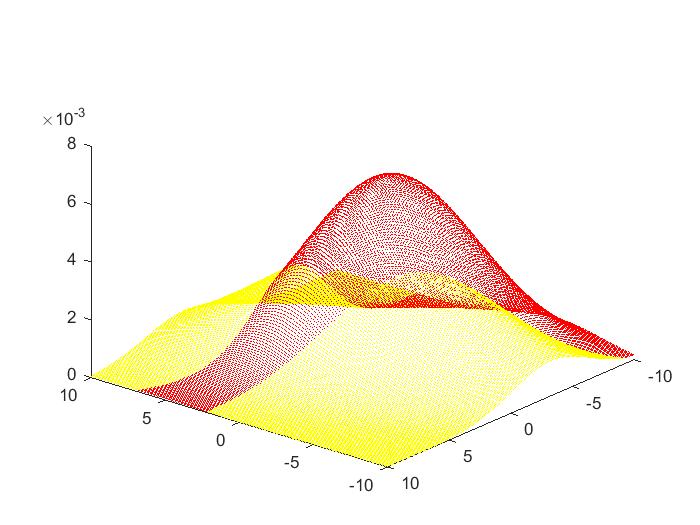
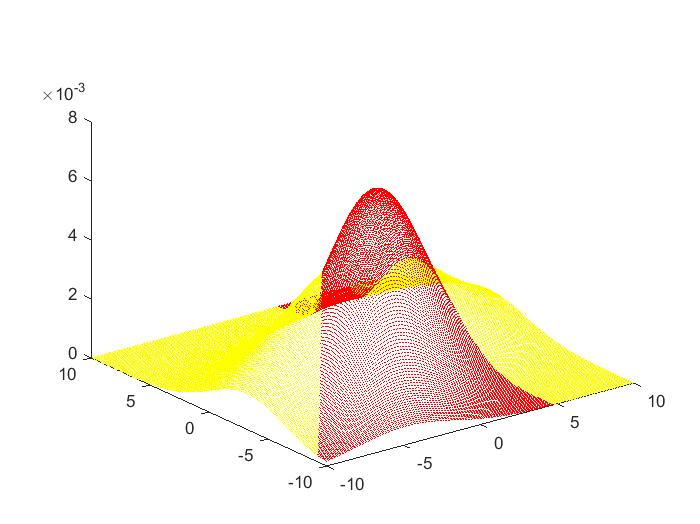
**（2）两类特征值下的两类分类问题**

可做出按照两类特征值下的空间函数判别式，高程图如图2表示。



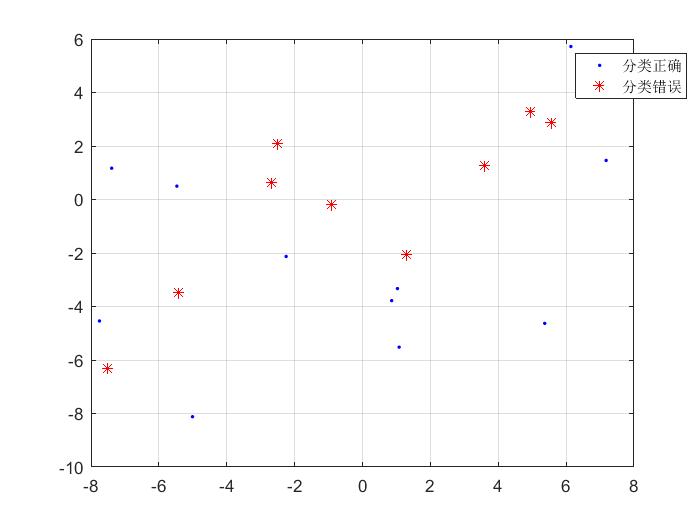
**图2** 两个特征值下的贝叶斯函数高程图

函数值的空间散点图如图3所示。



**图3** 两个特征值下的贝叶斯函数散点图

对高程图来讲，可清楚看见两个正态空间形状叠加的图像，对于散点图就更加直观了，红色表示第一类，黄色表示第二类。散点图在水平面的投影直接反映了两个特征值的区间（图3后两幅图）。

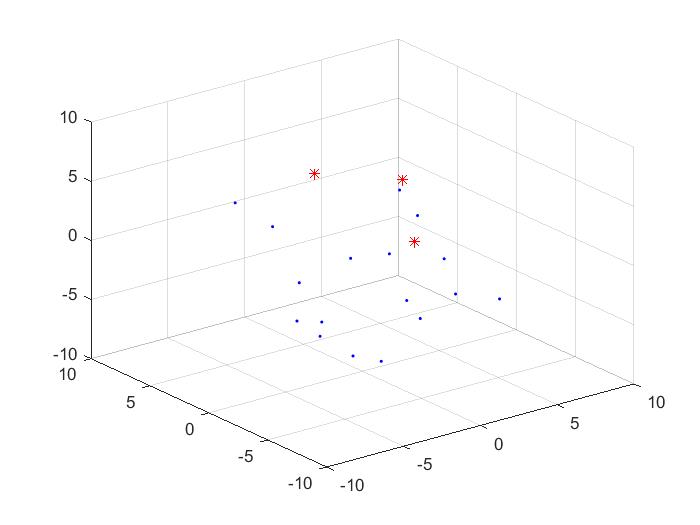
用设计的分类器判定表中的误分点，做出图像。得到图4。

**图4** 两个特征值下的误分点

误分点有九个，所以误分点百分比为0.45（45%），按照式（4），求得Bhattacharyya造成的误差为0.4604。

**（3）三类特征值下的两类分类问题**

将全部三类特征值进行分类，通用可以得到误分点，如图5所示。



**图5** 三个特征值,下的误分点

误分点有三个，所以误分点百分比为0.15（15%），按照式（4），求得Bhattacharyya造成的误差为0.4119。

**（5）讨论：影响误差的影响因素**

首先，在训练集较少的情况下，可能会造成较大的不稳定性，往较小的训练集下得到的结论并不是可以反应整体属性的情况：例如，训练集与整体分布存在明显差异，并不严格符合正态分布。

其次，特征值的选取也是重要的问题，例如只考虑x1和考虑x1,x2的情况，x1似乎比x1,x2还要稳定，这说明x2并不是一个十分理想的特征，两种类别的x2分布十分接近，在这种情况下，我们可以认为非显著特征x2“拖了x1的后腿”。

综上所述，我们并不是一定可以说，对于一个有限的数据集，更高的数据维度一定可以保证误差的减小。