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，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）。

$$P(w_j|x)=\frac{P(x|w_j)P（w_j）}{P(x)} \qquad (1)$$

特别的，正态分布的判别函数可以化为：

$$g_i(x)=-\frac{1}{2}(x-\mu_i)^t\sum_i^{-1}(x-\mu_i)-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\sum_i|+\ln P(w_i) \qquad (2)$$

在连续性分布的问题中，我们可以对（2）进行改进，变成了

$$g_i(x) = \rho(x_{w_i})P(w_i) \tag{3}$$

其中，$\rho(x_{w_i})$ 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 x_1 为特征值对 w_1 和 w_2 进行分类做出图像如图 1 所示

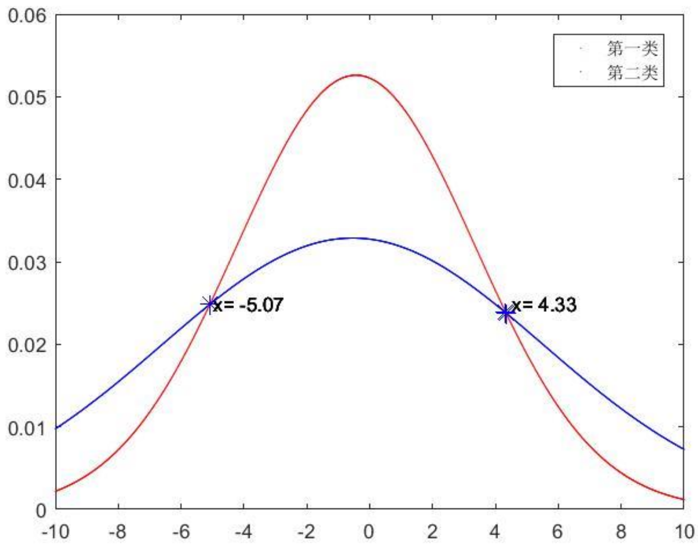


图 1 一类特征值时的判别图

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 x_1 进行分类，误分比为 $\frac{6}{20} = 0.300$，百分比为 30%，按照式（4），求得 Bhattacharyya 造成的误差为 0.4740。

$$\begin{cases} P(error) \le \sqrt{p(w_1)p(w_2)}\int \sqrt{p(x|w_1)p(x|w_2)}dx = \sqrt{p(w_1)p(w_2)}e^{-k(1/2)} \\ k(1/2) = 1/8(\mu_2 - \mu_1)^t[\frac{\sum_1 + \sum_2}{2}](\mu_2 - \mu_1) + \frac{1}{2}\ln\frac{|\frac{\sum_1 + \sum_2}{2}|}{\sqrt{|\sum_1 \| \sum_2|}} \end{cases} \tag{4}$$

至此，按照 x_1 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 x_1, x_2 两类特征值下的空间函数判别式，高程图如图 2 表示。

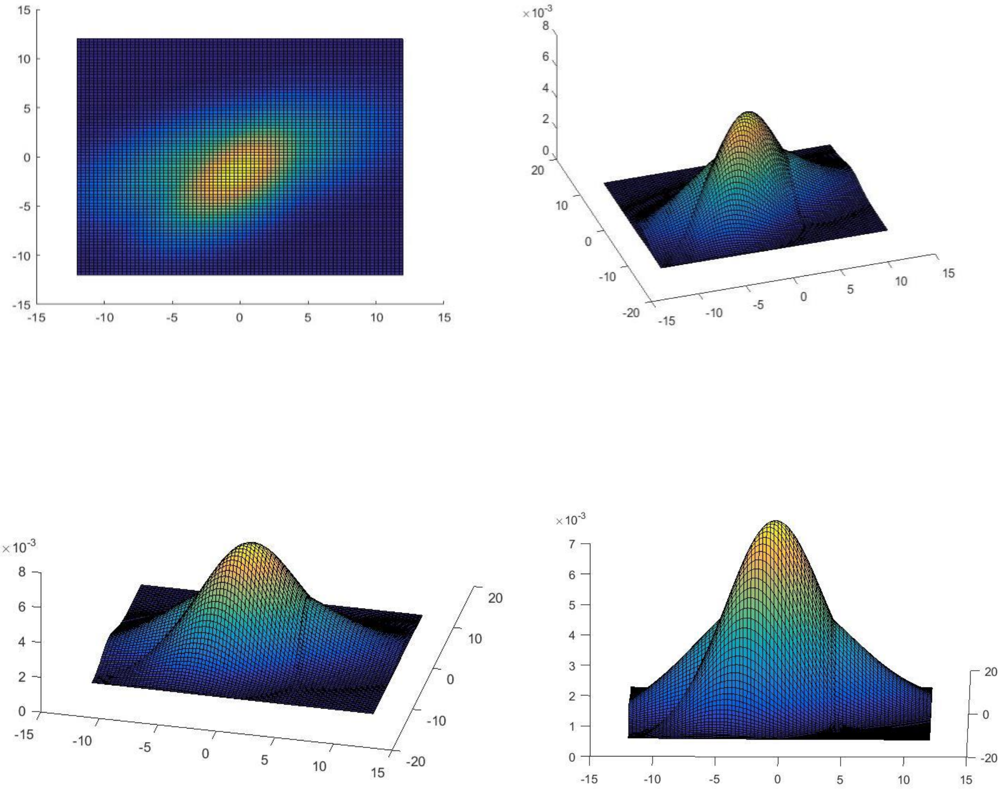
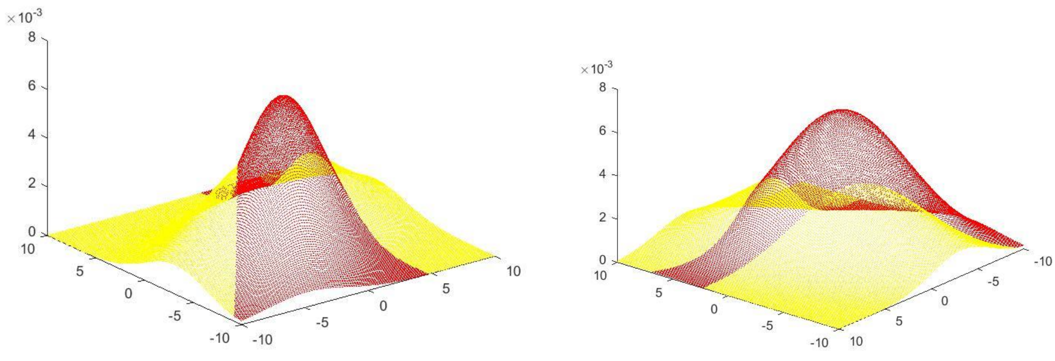


图 2　两个特征值 x_1, x_2 下的贝叶斯函数高程图

函数值的空间散点图如图 3 所示。



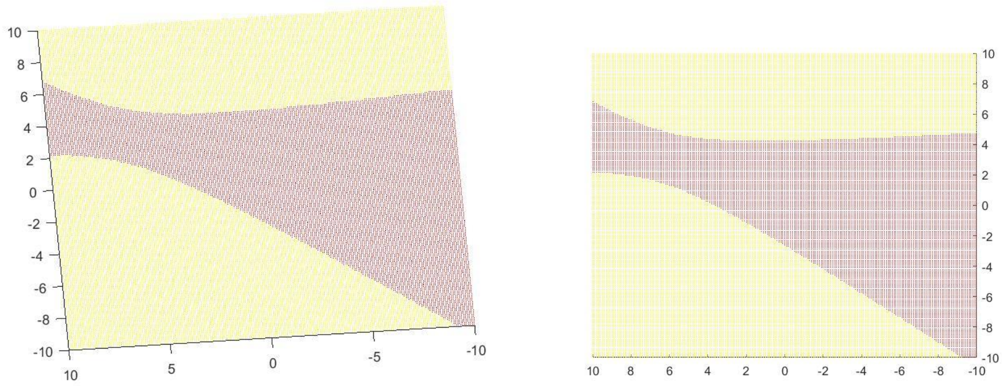


图 3 两个特征值 x_1, x_2 下的贝叶斯函数散点图

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 x_1, x_2 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。

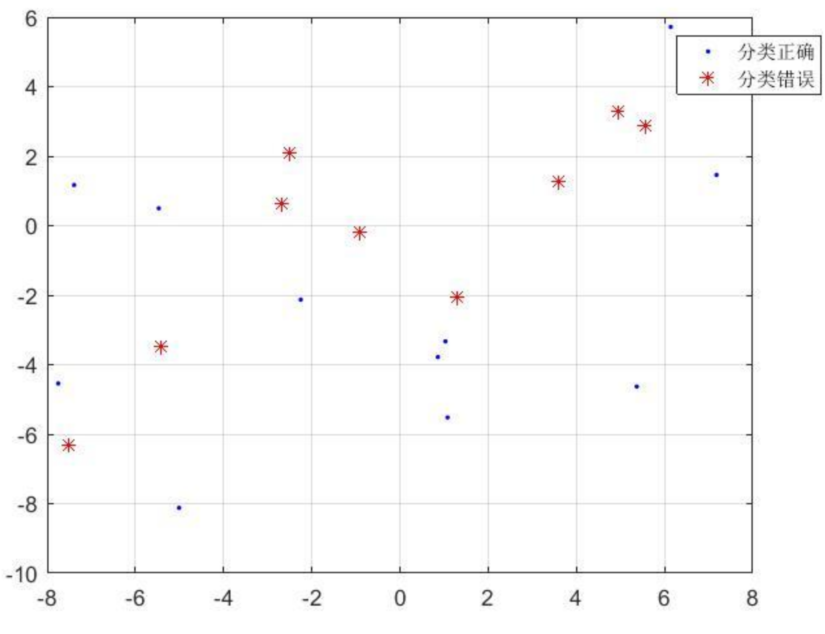


图 4 两个特征值 x_1, x_2 下的误分点

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。

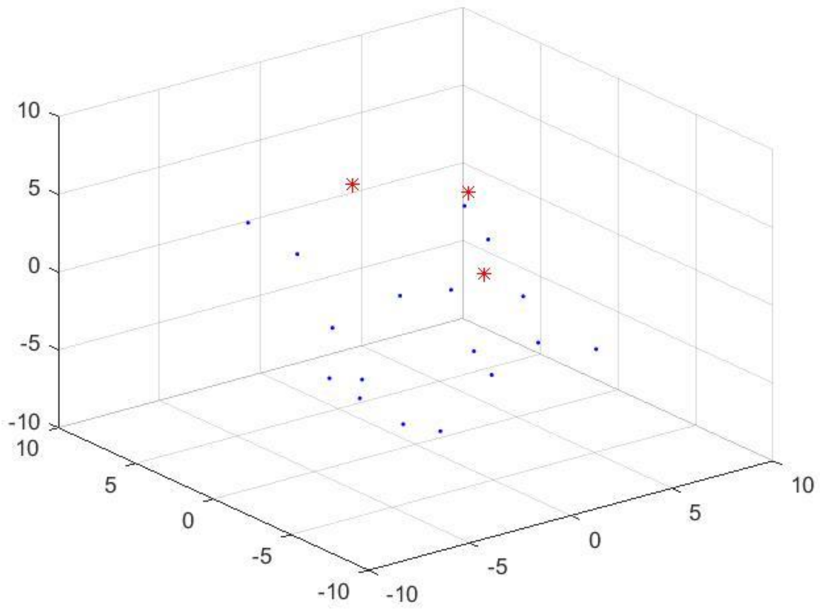


图 5 三个特征值 x_1, x_2, x_3 下的误分点

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。