

山东大学机器学习课程

实验报告

——实验一：贝叶斯分类器

姓名：陈竞帆

学院：软件学院

班级：软件八班

学号：201500150180

**一、实验目的：**

（1）熟悉python实验软件及相关函数，

（2）学习以贝叶斯分类器为例的常见聚类方法的思想与算法

（3）根据已给数据集，用贝叶斯分类器进行分类

**二、实验环境：**

（1）硬件环境：

英特尔® 酷睿™ i5-7500U 处理器

256 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

Python3.6

**三、实验内容**

**3.1贝叶斯分类器概述**

贝叶斯分类器是在具有一般风险的情况下，让g­i(x)=-R(ai|x)，此时判别函数越大，对应的风险越小。由于这道题并没有提供条件风险函数，因此我们默认为题目是在最小误差概率的情况下进行分类，即：

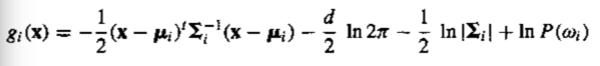


在如上的损失函数下进行分类。为了便于计算机的计算，我们的分类器选为：

../Library/Containers/com.tencent.qq/Data/Library/Application%20Support/QQ/Users/951718744/QQ/Temp.db/D3CF8111-5B0E-474F-AF3F-A27A8BE8FD8C.png

因为题目已经说明，概率密度函数为正态分布，且先验概率已知为p(w1) = p(w2)=0.5, p(w3) =0, 因此只需要对已有的样本数据进行正态分布的参数确定即可获得一个贝叶斯分类器。

因此我们得到如下的正态判别函数：



其中ui为某一类下的正态分布的期望，sigma为协方差矩阵，d为输入数据的维度。

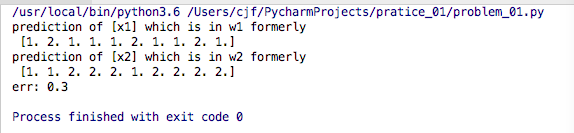
**3.2 贝叶斯分类器算法**

设计程序的伪代码，实现图1所示：

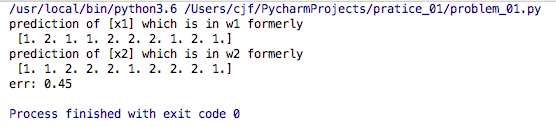
1. 读取json文件中的数据集
2. 计算每一类下的样本均值和样本协方差，并作为正态概率密度函数的参数
3. 代入先验概率和第二步计算得出的参数，作为贝叶斯分类器的参数。
4. 将数据集带入，计算出后验概率，若P(wi|x)>P(wj|x),则当前样本点x属于wi，否则属于wj。

**图1. 贝叶斯分类器 算法伪代码**

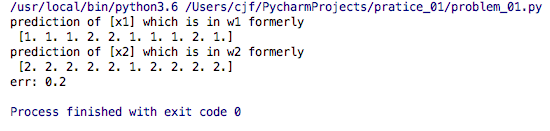
**四、实验结果**



**图2. 只针对x1数据的实验结果**



**图3. 针对x1和x2的实验结果图**

****

**图4. 针对x1、x2和x3的实验结果图**

**误差公式为：错判数目/样本总数**

**从图2中我们可以看出，仅利用x1为特征值的时候，误差为0.3**

**从图3中我们可以看出，仅利用x1、x2为特征值的时候，误差为0.45**

**从图4中我们可以看出，仅利用x1、x2、x3为特征值的时候，误差为0.2**

**五、总结与归纳**

**本次实验数据集较少，而我对正态分布的参数估计是基于训练数据的mean和cov，因此存在偏差。同时我们也可以看到，当维度增大时，准确率会有所提升，我认为是因为有更多的信息去刻画样本点的分类，所以会更加准确。但是，并不是说明，维度更高，对应的经验误差就会越低，我认为也有可能出现经验误差升高的情况。从我们的第二个例子就可以看出这一结论。**