模式识别第二次作业

2120190442 周宝航

**目录**

[**一、** **问题描述** 1](#_Toc21966617)

[**二、** **基本思路** 2](#_Toc21966618)

[**三、** **算法描述和算法实现** 3](#_Toc21966619)

[**四、** **结果与分析** 10](#_Toc21966620)

1. **问题描述**

问题1：

在一维模式特征空间的两类问题中，两类模式的概率密度分布函数分别为和。试证明最小平均风险的分类阈值为，其中假设。

问题2：

1. 生成两个各包含个二维随机矢量的数据集合和。数据集合中随机矢量来自于三个分布模型，它们分别满足均值矢量和和协方差矩阵，其中是的单位矩阵。在生成数据集合时，假设来自三个分布模型的先验概率相同；而在生成数据集合时，先验概率分别为0.6、0.3和0.1。
2. 分别画出所生成的两个数据集合中随机矢量的散布图。
3. 在两个数据集合上分别使用“似然率测试规则”、“贝叶斯风险规则”（其中）、“最大后验概率规则”和“最短欧式距离规则”进行模式分类实验，给出实验过程设计和实验结果。
4. 对每个数据集合给出上述每种分类规则的分类错误率，分析结果并给出你的结论。
5. **基本思路**

**问题1：**根据最小贝叶斯风险决策规则以及给定的两类模式的概率密度分布函数，求得最小平均风险的分类阈值。

**问题2：**

**2.1** 已知样本服从多元正太分布，我们可以直接利用numpy包的multivariate\_normal函数直接生成相应均值、协方差矩阵的数据。为了测试模型的性能，我们直接将生成的数据作为测试集。为了使得测试结果具有统计意义，我们设置100个不同的随机数种子来生成同分布的不同数据，然后得到100组测试结果并求其均值与标准差。

**2.2** “似然率测试规则”对于两类问题的决策规则为：

我们只需根据样本、先验概率与似然函数按照上式进行不同类别的比较，即可得到该样本的类别。

**2.3** “贝叶斯风险规则”首先计算样本分别属于三个类别时的后验概率，然后将它们与对应的系数相乘并求和即可得到该样本属于对应类别的风险值。我们选取最小风险值时对应的类别作为该样本的预测结果。

**2.4** “最大后验概率规则”将根据贝叶斯公式计算样本属于每个类别的概率。我们选择后验概率最大值对应的类别作为该样本的预测结果。

**2.5** “最短欧式距离规则”将计算测试样本与不同类别的样本均值之间的欧式距离，并选择距离最小的样本均值对应的类别作为预测结果。

1. **算法描述和算法实现**

**问题1：**

**证明：**

已知决策模式特征为模式特征类别的条件风险为：。那么给定的两类条件风险分别为：

已知最小化条件风险等价于满足最小化的模式类别，那么最优决策规则为：。已给定和贝叶斯公式，将公式（1）（2）带入上述规则可得：

则最小平均风险的分类阈值即为：的解。

已知正态分布的概率密度分布为：，那么依照题意则有：

对上式两边取对数，解得：

**问题2：**

1. **生成数据**

****

实验要求生成两组数据，其中两组数据的对应维度服从相同的多元正太分布，而不同的是两组数据的先验分布。我们使用multivariate\_normal函数直接生成相应均值、协方差矩阵的数据，而通过传入不同的size变量生成相应先验概率的数据。

1. **数据散布图**



我们采用matplotlib包中的scatter函数画出二维数据的散点图。生成结果如下：

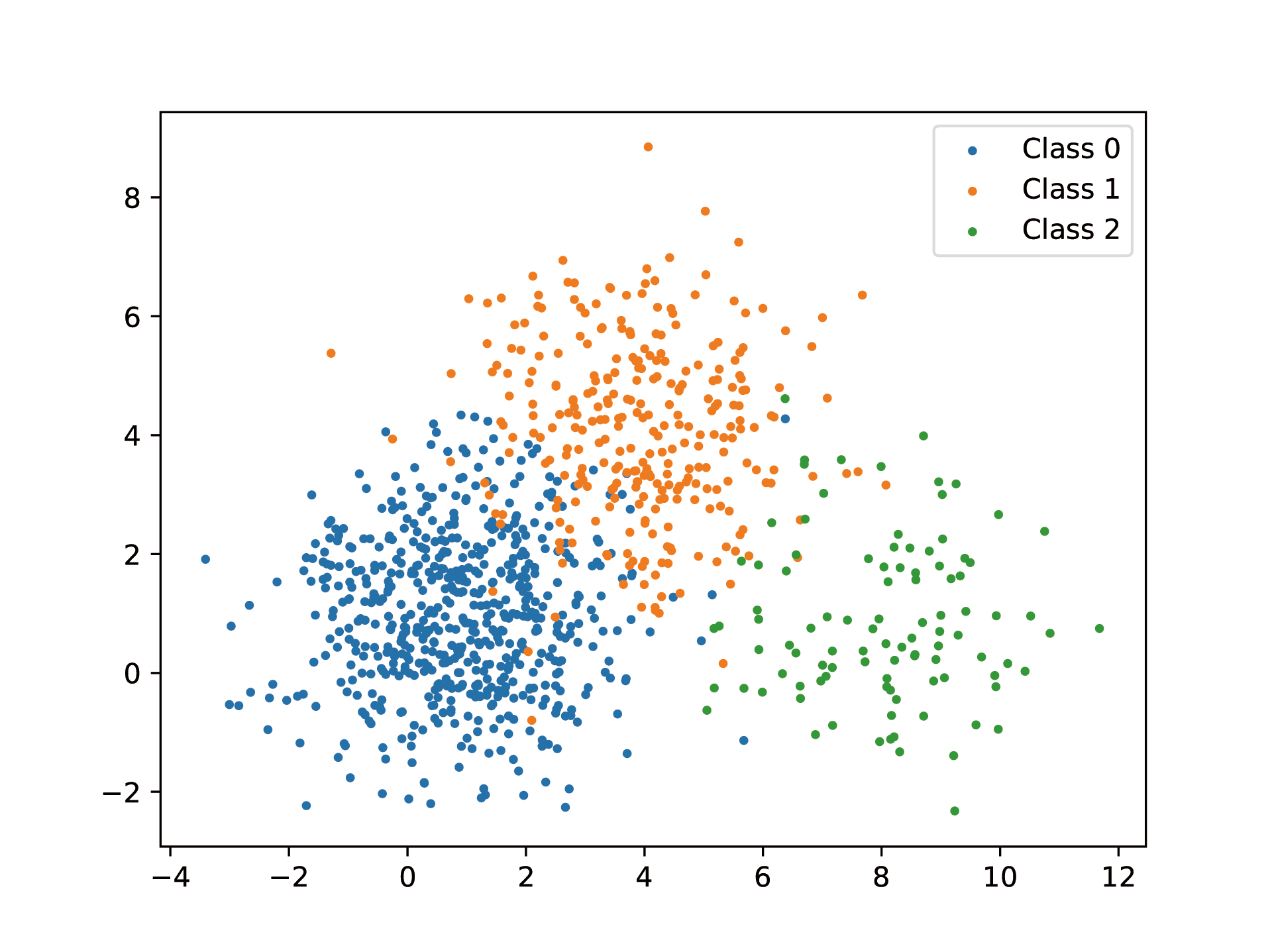
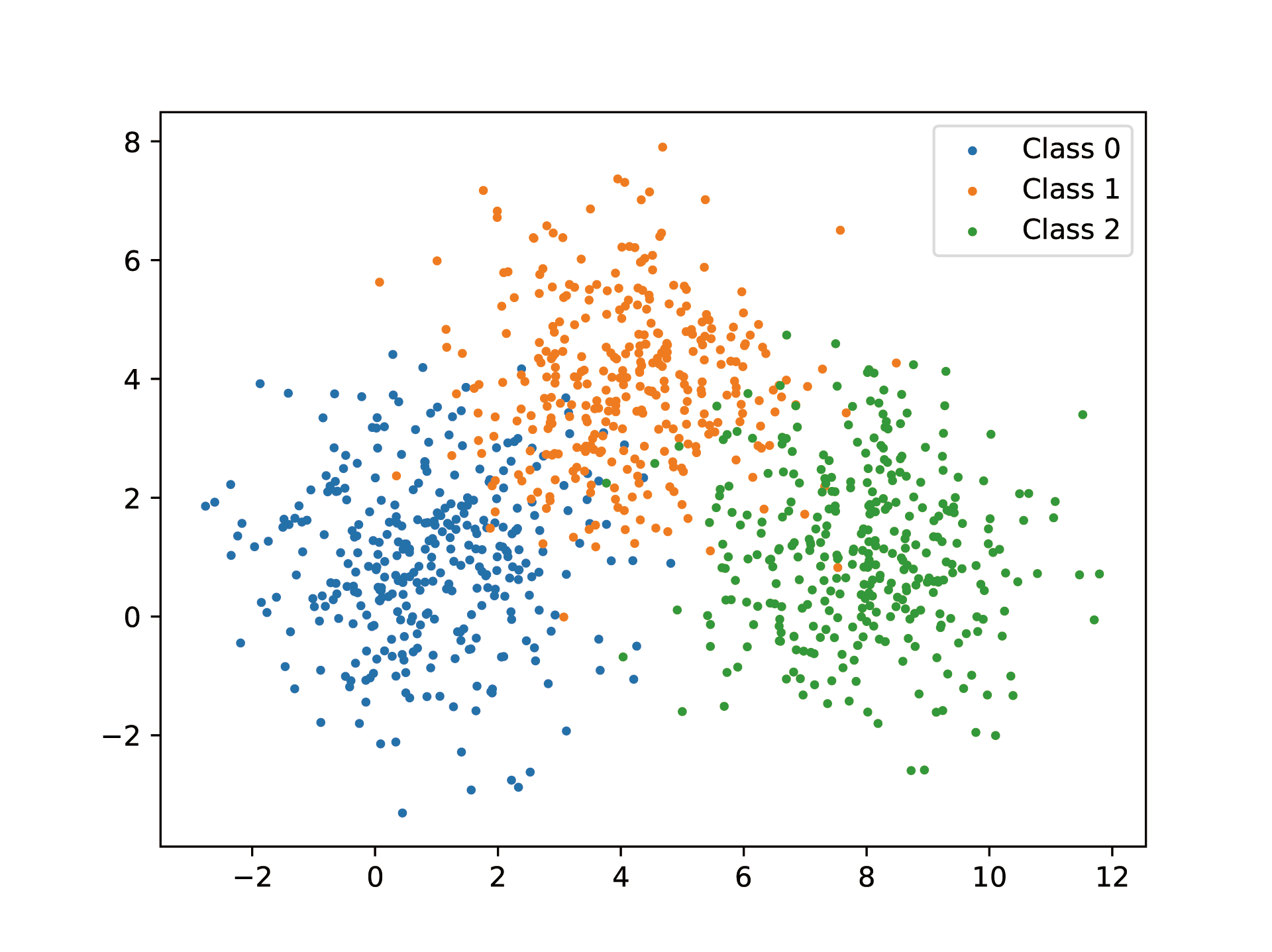


图2 数据集合2

图1 数据集合1

1. **实验要求对上面生成的两个数据集分别采用“似然率测试规则”、“贝叶斯风险规则”、“最大后验概率规则”和“最短欧式距离规则”进行分类实验。**
   1. **实验设计过程：**

**实验数据集：**我们通过多元高斯分布生成了两个数据集，它们每个维度的均值、协方差矩阵以及先验概率均已给定。每个数据集包含三个类别，分别标记为0、1、2，如图1、图2所示。

**实验分类器：**

1. “似然率测试规则”分类器
2. “贝叶斯风险规则”分类器
3. “最大后验概率规则”分类器
4. “最短欧式距离规则”分类器

**实验评价指标：**

* 1. **算法分析及核心代码：**

1. **“似然率测试规则”分类器**

我们已知数据服从多元正太分布，那么其似然函数可知为：

接着，我们在程序中定义了如下函数，通过传入函数的样本x、均值mean以及协方差矩阵cov得到计算结果。

****

我们已知二分类类问题的决策规则为：

那么对于多分类问题，我们只需两两进行似然率测试，找到最大似然率对应的类别作为预测结果。程序函数如下所示：

****

1. **“贝叶斯风险规则”分类器**

我们已知决策的代价函数为。那么多分类问题的风险可定义为：

其中c为样本可能归属的总类别数目。

在实验要求中，我们已经知道。我们将样本的似然函数值与对应的代价函数值相乘并求和，即可得到该样本分类为该类别的风险。我们取最小风险值对应的类别作为该样本的预测结果。

****

1. **“最大后验概率规则”分类器**

在“似然率测试规则”中，我们已经定义了似然函数。已知贝叶斯公式：

我们需要计算样本属于每个类别的后验概率，并选择最大后验概率对应的类别作为该样本的预测结果。其中，作为比例因子对分类结果没有影响。我们只需要根据似然函数、先验概率计算乘积即可。

在实验要求中，我们已经知道了样本的均值、协方差矩阵以及先验概率。程序函数如下所示：

****

1. **“最短欧式距离规则”分类器**

最短欧式距离的决策规则为：

其中，为待测试样本，为类的样本均值。我们计算样本到每个类别中心的欧式距离，并取最近类别作为预测结果。

****

1. **结果与分析**

表 1实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 平均值  （标准差） | 似然率测试规则 | 贝叶斯风险规则 | 最大后验概率规则 | 最短欧式距离规则 |
| 数据集合1 | 0.07059（0.00781） | 0.07745  （0.00804） | 0.07059（0.00781） | 0.07059（0.00781） |
| 数据集合2 | 0.06812  （0.00813） | 0.07443  （0.00865） | 0.06812  （0.00813） | 0.07566  （0.00859） |

从上述实验结果可以看出：

1. 在两个数据集上，“似然率测试规则”和“最大后验概率规则”表现相同。这一点印证了：似然率测试规则观察到模式后的最优决策就是最大后验概率决策。
2. 在数据集合1上，“最短欧式距离规则”与“最大后验概率规则”的表现相同。而在数据集合2上，前者表现远不如后者。已知多元正太分布下的判别函数为：。因为两个数据集中每个类别的协方差矩阵相同，所以判别函数只与样本到类中心的距离和先验概率有关。而数据集合1中每个类别的先验概率相同，故该情况下最大后验概率计算与最短欧式距离规则没有区别。而数据集合2的每个类别具有不同的先验概率，“最短欧式距离规则”没有将其考虑进去，所以表现不如“最大后验概率规则”。
3. “贝叶斯风险规则”在两个数据集上的表现都不好。因为该规则引入了风险函数来标识将某一类分错的风险更高。如此一来，在“最大后验概率规则”中被正确分类的样本会在“贝叶斯风险规则”下被错误分类。