模式识别作业一

姓名：李宁　　学号：2120180459　　专业：计算机科学与技术

问题一

一、问题描述

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$N(0,\sigma^2)$和$N(1,\sigma^2)$。试证明最小平均风险的分类阈值为$x_0=\frac{1}{2}-\sigma^2\ln\frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}$，其中假设$C_{11}=C_{22}=0$。

二、基本思路

对于两类的贝叶斯风险

$$R=\mathbb{E}[C]=\sum_{i=1}^{2}\sum_{j=1}^{2}C_{ij}Pr\left(decide\ \ \omega_i|\mathrm{x}\in\omega_j\right)=\sum_{i=1}^{2}\sum_{j=1}^{2}C_{ij}p\left(x\in R_i|\omega_j\right)Pr\left(\omega_j\right)$$

因为$p\left(x\in R_i|\omega_j\right)=\int_{R_1}p\left(x|\omega_j\right)dx$，故贝叶斯风险可表示为

$$R=\int_{R_1}[C_{11}p(x|\omega_1)Pr(\omega_1)+C_{12}p(x|\omega_2)Pr(\omega_2)]\,dx+$$

$$\int_{R_2}[C_{21}p(x|\omega_1)Pr(\omega_1)+C_{22}p(x|\omega_2)Pr(\omega_2)]\,dx$$

对于每个似然函数，有

$$\int_{R_1}p(x|\omega_i)\,dx+\int_{R_2}p(x|\omega_i)\,dx=\int_{R_1\cup R_2}p(x|\omega_i)\,dx=1$$

将上式代入贝叶斯风险公式，整理后消除在区域R_2上的积分项，可得

$$R=C_{21}Pr(\omega_1)+C_{22}Pr(\omega_2)$$

$$+(C_{12}-C_{22})Pr(\omega_2)\int_{R_1}p(x|\omega_2)\,dx$$

$$-(C_{21}-C_{11})Pr(\omega_1)\int_{R_1}p(x|\omega_1)\,dx$$

最小化贝叶斯风险的决策区域R_1为

$$R_1=arg\min_R\left\{\int_R[(C_{12}-C_{22})Pr(\omega_2)p(x|\omega_2)-(C_{21}-C_{11})Pr(\omega_1)p(x|\omega_1)]\,dx\right\}$$

$$=arg\min_R\left\{\int_R g(x)\,dx\right\}$$

又$g(x)<0$，即分类为ω_1时，等价于

$$(C_{12}-C_{22})Pr(\omega_2)p(x|\omega_2)<(C_{21}-C_{11})Pr(\omega_1)p(x|\omega_1)$$

整理得到

$$\text{Decide } \omega_1 \text{ if } \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(C_{12}-C_{22})}{(C_{21}-C_{11})} \cdot \frac{Pr(\omega_2)}{Pr(\omega_1)}$$

三、问题求解

由题可知，两个分类的似然函数分别为 $p(x|\omega_1) = N(0,\sigma^2)$，$p(x|\omega_2) = N(1,\sigma^2)$。

由最小贝叶斯风险规则

$$\text{Decide } \omega_1 \text{ if } \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(C_{12}-C_{22})}{(C_{21}-C_{11})} \cdot \frac{Pr(\omega_2)}{Pr(\omega_1)}$$

可得

①对于类别一，令

$$\frac{N(0,\sigma^2)}{N(1,\sigma^2)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-1)^2}{2\sigma^2}}} = e^{\frac{(x-1)^2}{2\sigma^2}-\frac{x^2}{2\sigma^2}} = e^{\frac{-2x+1}{2\sigma^2}} > \frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}$$

对上式两边取对数得

$$\ln\left(e^{\frac{-2x+1}{2\sigma^2}}\right) > \ln\left(\frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}\right)$$

则

$$\frac{-2x+1}{2\sigma^2} > \ln\frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}$$

求解得

$$x_1 < \frac{1}{2} - \sigma^2 \ln\frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}$$

②同理，对于类别二，可以得到

$$x_2 > \frac{1}{2} - \sigma^2 \ln\frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}$$

因此，可以得到最小贝叶斯风险分类得阈值为

$$x_0 = \frac{1}{2} - \sigma^2 \ln\frac{C_{12}Pr(\omega_2)}{C_{21}Pr(\omega_1)}$$

问题二

一、问题描述

随机生成各包含 1000 个二维随机矢量的数据集合 X 和 X'。两个数据集合中的随机矢量来自于三个分布模型，它们的均值矢量分别为 $\mathbf{m}_1=[1,1]^T$，$\mathbf{m}_2=[4,4]^T$，$\mathbf{m}_3=[8,1]^T$，它们的协方差矩阵为 $\mathbf{S}_1=\mathbf{S}_2=\mathbf{S}_3=2\mathbf{I}$，其中 **I** 是 2*2 的单位矩阵。生成数据集合 X 时，来自三个分布模型的先验概率相同；而生成数据集合 X'时，来自三个分布的先验概率分别为 0.6、0.3 和 0.1。

本次作业要求对于上述两个数据集合 X 和 X'。

① 首先分别画出两个数据集合的随机矢量的散布图；

② 分别用似然率测试规则、最小贝叶斯风险规则、最大后验概率规则和最短欧氏距离规则对两个数据集合进行分类；

③ 对每种分类规则计算错误率并分析结果

二、基本思路

① 对于随机生成两个数据集合的散布图。

生成散布图的最关键因素是近似计算出两个数据集合分别在三个模型分布中生成的数量。

由题给出的已知条件，数据集合的数据总量、分布函数、分布比例均已知，可以通过随机生成数在（0,1）之间的生成的概率来近似确定每类分布在 1000 个数中所需要生成的数量。

② 对于似然率测试规则、最小贝叶斯风险规则、最大后验概率测试规则。

三种决策规则均是在贝叶斯决策规则的基础上推到而来。

由贝叶斯公式

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

将贝叶斯规则扩展到随机向量可以得到如下公式

$$P(\omega_i|x)=\frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

其中$P(\omega_i)$为先验概率，在本题目中，对于数据集合 X，$P(\omega_1)=P(\omega_2)=P(\omega_3)=1/3$；对于数据集合 X'，$P(\omega_1)=0.6$，$P(\omega_2)=0.3$，$P(\omega_3)=0.1$。

$p(x|\omega_i)$为似然函数，在本题中，三个模型均服从二维正态分布，故似然函数为

$$p(x|\omega_i)=\frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u}_i)^T\sum_i^{-1}(\boldsymbol{x}-\boldsymbol{u}_i)\right)$$

在上式中，i=1，2，3；d=2。Σ_i为协方差矩阵，$\boldsymbol{u}_i$为模型 i 的均值。在本题中，$\Sigma_1=\Sigma_2=\Sigma_3=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}$，$\boldsymbol{u}_1=[1,1]^T$，$\boldsymbol{u}_2=[4,4]^T$，$\boldsymbol{u}_3=[8,1]^T$。

$p(x)$为比例因子，$P(\omega_i|x)$为后验概率，两者未知，但$p(x)$可以视为常量。

贝叶斯决策标准如下：

$$decide\ \ \omega_i\ \ if\ \ p(\omega_i|x)>p(\omega_j|x),\qquad \forall i\neq j$$

因此，可以利用已知的$p(x|\omega_i)$和$P(\omega_i)$来对三个决策规则进行求解。

而对于最小欧氏距离分类器，可以直接计算每个点距离三个分布均值点的距离的平方进行分类。

③ 该实验中，三种分类模型已知

因此可以将数据集合 X 和 X'直接作为测试数据集，本实验不需要训练数据集。之后可通过对分类后重新为集合中各个点分配的类标签（1，2，3）与集合初始化时为每个点分配的类标签做对比，进而求出每种分类器的错误率。

三、原理及算法

① 散布图的生成

算法流程：

- 通过先验概率计算三种分布的数量 count1，count2，count3
- M_1=[] <—— 生成 count1 个满足$N(u_1,\sigma^2)$的散点
- M_2=[] <—— 生成 count2 个满足$N(u_2,\sigma^2)$的散点
- M_3=[] <—— 生成 count3 个满足$N(u_3,\sigma^2)$的散点
- 绘制散点集合 M_1，M_2，M_3

② 分类规则设计原理

• 似然率测试规则

似然率测试规则的决策为

$$decide\ \ \omega_i\ \ if\ \ \frac{p(\boldsymbol{x}|\omega_i)}{p(\boldsymbol{x}|\omega_j)}>\frac{P(\omega_j)}{P(\omega_i)},\qquad \forall i\neq j$$

等价于

$$decide\ \ \omega_i\ \ if\ \ p(\boldsymbol{x}|\omega_i)P(\omega_i)>p(\boldsymbol{x}|\omega_j)P(\omega_j),\qquad \forall i\neq j$$

因此似然率测试规则的决策是使得$p(\boldsymbol{x}|\omega_i)P(\omega_i)$取值最大的类别。

• 贝叶斯风险规则

贝叶斯风险的定义公式如下

$$R=\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}P(decide\ \ \omega_i|\mathbf{x}\in\omega_j)=\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}p(\boldsymbol{x}\in R_i|\omega_j)P(\omega_j)$$

其中，C_{ij}表示为原来是 j 的类别被判别为 i 类别的风险系数，n 表示总的类别数。

根据在多类别问题中最小贝叶斯风险决策规则，决策模式特征$\mathbf{x}$为模式类别ω_i的条件风险为

$$\Re(a(\boldsymbol{x})\rightarrow\omega_i)=\Re(\omega_i|\boldsymbol{x})=\sum_{j=1}^{C}C_{ij}P(\omega_j|\boldsymbol{x})$$

其中

$$P(\omega_i|\boldsymbol{x})=\frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})}=\frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{C}p(\boldsymbol{x}|\omega_j)P(\omega_j)}$$

因为对于所有的类别，$\sum_{j=1}^{C}p(\boldsymbol{x}|\omega_j)P(\omega_j)$是一个常数，不影响最终的决策，因此最小化贝叶斯决策的条件风险$\Re(\omega_i|\boldsymbol{x})$等价于最小化$\sum_{j=1}^{C}C_{ij}p(\boldsymbol{x}|\omega_i)P(\omega_i)$。

因此贝叶斯风险规则的决策为使得$\sum_{j=1}^{C} C_{ij} p(\boldsymbol{x}|\omega_i) P(\omega_i)$取值最小的类别。

• 最大后验概率规则

最大后验概率规则的决策为

$$decide \quad \omega_i \quad if \quad P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}), \qquad \forall i \neq j$$

其中

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{C} p(\boldsymbol{x}|\omega_j)P(\omega_j)}$$

因此，最大后验概率规则的决策是使得$P(\omega_i|\boldsymbol{x})$取值最大的类别。

• 最短欧氏距离规则

特征矢量$\mathbf{x}$与类别ω_i的均值矢量$\boldsymbol{u}_i$之间的欧氏距离为

$$d(\boldsymbol{x}, \boldsymbol{u}_i) = (\boldsymbol{x} - \boldsymbol{u}_i)^T (\boldsymbol{x} - \boldsymbol{u}_i)$$

因此最短欧氏距离规则的决策是使得$d(\boldsymbol{x}, \boldsymbol{u}_i)$取值最小的类别

三、实验结果与分析

1、数据集合 X 与 X’的散布图

实验结果如下图所示

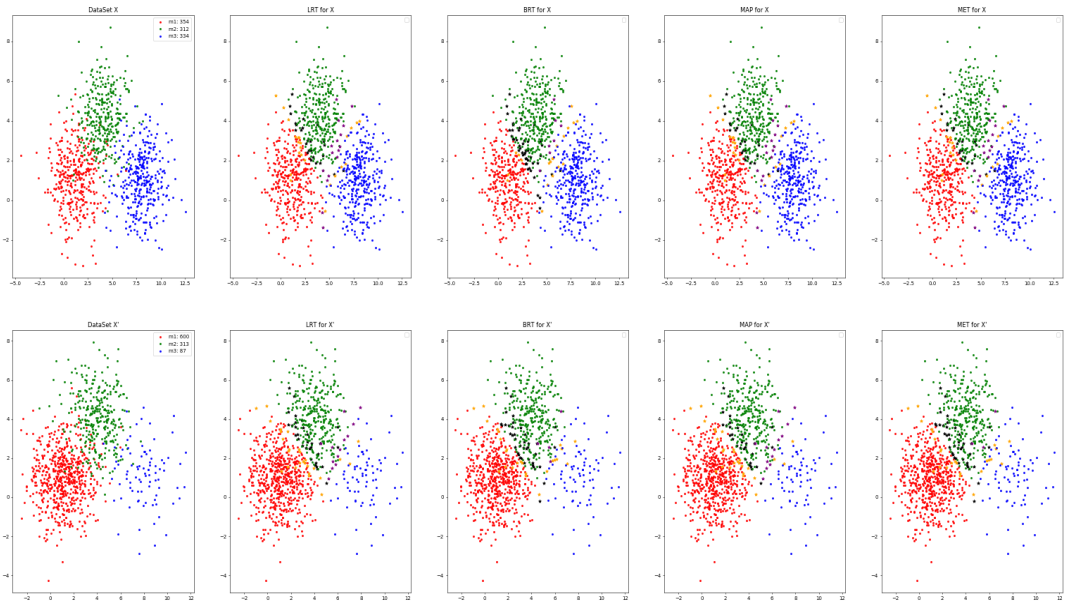

图 1 X 与 X’分布及各分类器分类结果分布

如图 1 所示，M_1，M_2，M_3 三个类别分别如中“红色”、“绿色”、“蓝色”数据所示。

图中第一行为 X 数据集合的数据分布，其中第一张散布图为 X 数据集合的原始数据分布图；之后四张图片分别是 X 数据集合在似然率测试规则、贝叶斯风险测试规则、最大后验测试规则、最小欧氏距离测试规则上进行分类的测试结果，其中“黑色”数据为原本为

M_1 类别被分类错误的数据点，“橙色”数据为原本为 M_2 类别被分类错误的数据点，“紫色”数据为原本为 M_3 类别被分类错误的数据点。

图中第二行为 X’数据集合的数据分布，其他与数据集合 X 相似。

2、错误率

• 在数据集合 X 上

数据集合 X 用似然率测试规则分类的错误率：0.065

数据集合 X 用贝叶斯风险测试规则分类的错误率：0.074

数据集合 X 用最大后验测试规则分类的错误率：0.065

数据集合 X 用最小欧氏距离测试规则分类的错误率：0.065

• 在数据集合 X’上

数据集合 X’用似然率测试规则分类的错误率：0.071

数据集合 X’用贝叶斯风险测试规则分类的错误率：0.070

数据集合 X’用最大后验测试规则分类的错误率：0.071

数据集合 X’用最小欧氏距离测试规则分类的错误率：0.069

3、实验结果分析

① 似然率测试规则和最大后验概率测试规则的分类结果是一致的。这是因为似然率测试规则的决策依据

$$decide \quad \omega_i \quad if \quad \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)} > \frac{P(\omega_j)}{P(\omega_i)}, \qquad \forall i \neq j$$

与最大后验概率测试规则的决策依据

$$decide \quad \omega_i \quad if \quad p(\omega_i|\mathbf{x}) > p(\omega_j|\mathbf{x}), \qquad \forall i \neq j$$

两者是等价的，所以分类结果一致。

② 对于数据集合 X，最短欧氏距离规则与似然率测试规则以及最大后验测试规则的分类结果一致；而对于数据集合 X’，最短欧氏距离规则与后两种分类规则分类结果不一致。

原因在于最短欧氏距离测试规则在一定条件下与贝叶斯测试规则是一致的。

对于最小欧氏距离分类器，贝叶斯最优的条件是：

- 模式类别样本满足正态分布
- 相等的协方差矩阵且与单位矩阵成比例
- 相等的先验概率