北京航空航天大學

BEIHANG UNIVERSITY

模式识别实验报告

实验二、贝叶斯分类器

院（系）名称	自动化科学与电气工程学院
专业名称	自动化
学生学号	15031117
学生姓名	柳天宇
指导教师	

贝叶斯分类是一种依据概率统计理论进行决策的模式分类方法，它基于先验知识和条件概率的分布通过贝叶斯公式求得后验概率，将后验概率作为判别样本归属类别的依据，在理论上可以使分类错误最小化。贝叶斯分类的方法可以简单有效地解决二分类和多分类问题，作为模式识别中的一种基本方法而得到广泛应用。

1.知识准备

1. 先验概率：是指根据以往经验和分析得到的概率
2. 后验概率：使用了有关自然状态更加全面的资料，既有先验概率资料，也有补充资料
3. 贝叶斯公式：$P(wi \mid X)=\dfrac{P(wi)P(X \mid wi)}{P(X)}$，它提供了求取后验概率的手段
4. 最小错误规则：在一般的模式识别问题中，人们的目标往往是尽量减少分类的错误，追求最小的错误率。即求解一种决策规则，使得：$\min P(e)=\int P(e \mid x)P(x)dx$
5. 最小风险规则：考虑各种错误造成损失不同时的一种最优决策，通过决策表对后验概率进行加权操作从而求得损失函数，根据损失函数做出风险最小的分类决策。

2.贝叶斯分类设计的一般性步骤：

6. 依据先验知识获得每个类别的先验概率
7. 依据样本数据求得类条件概率
8. 通过贝叶斯公式求得后验概率
9. 依据决策规则做出分类决策

3.实验内容及结果分析

3.1 设计贝叶斯分类器完成对于给定一维特征样本的二分类

3.1.1 设计流程

（1）由先验知识获取先验概率

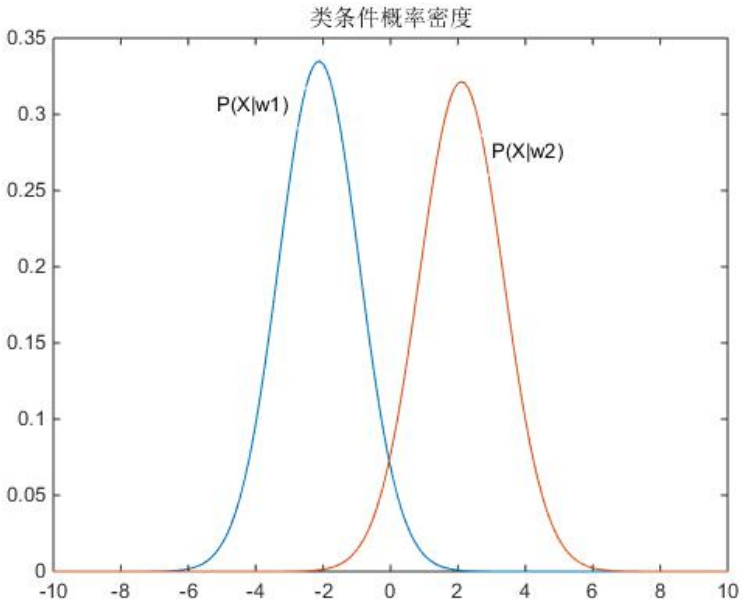
（2）类条件概率的分布形式已知，根据给定样的数据样本求取每个类条件概率分布中未知参数：均值、方差的极大似然估计，从而获得条件概率的似然函数

（3）由先验概率与类条件概率求得后验概率：$P(wi \mid X)=\dfrac{P(wi)P(X \mid wi)}{P(X)}$

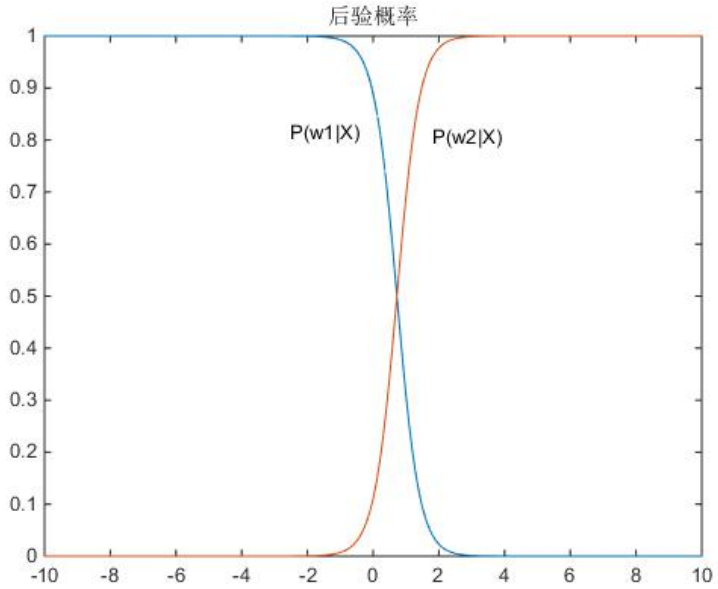
（4）依据后验概率的大小在给定的决策规则下做出分类决策

3.1.2 按最小错误率的规则做出决策

（1）类条件概率密度曲线：



（2）后验概率曲线：



（3）根据最小错误率决策规则：

$$\begin{cases} 若P(\omega_1/x) > P(\omega_2/x), 则x \in \omega_1 \\ 若P(\omega_1/x) < P(\omega_2/x), 则x \in \omega_2 \end{cases}$$

故决策边界方程： $P(\omega_1/x) = P(\omega_2/x)$

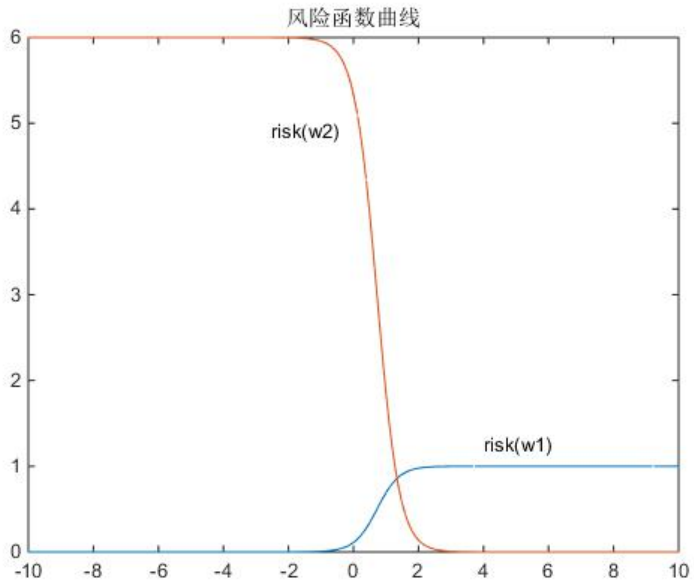
对于该实验，若样本点落在决策边界左侧，则判为ω_1类；若落在决策边界右侧，则判为ω_2类。

3.1.3 考虑错误分类带来的损失，在最小风险规则下做出决策

（1）由决策表中误判产生的损失，可以求出判别为每个类别风险：

$$risk(1)=\lambda_{11}P(w1\mid x)+\lambda_{12}P(w2\mid x)=P(w2\mid x)$$
$$risk(2)=\lambda_{21}P(w1\mid x)+\lambda_{22}P(w2\mid x)=6*P(w1\mid x)$$

（2）风险函数曲线：



（3）根据最小风险决策规则：

$$若risk(1)<risk(2),则x\in w1;$$
$$若risk(1)>risk(2),则x\in w2;$$

故决策边界方程：$risk(1)=risk(2)$。对于该实验，若样本点落在决策边界左侧，则判为第一类；若样本点落在决策边界右侧，则判为第二类。

（4）结果分析与讨论

通过两组实验的对比可以发现，贝叶斯分类器在不同的决策规则下获得的决策边界是不同的，就是说对于同一个样本点按照不同规则去进行贝叶斯分类可能会得到不同的分类结果。在该实验中，相比于最小错误原则，最小风险原则会使决策边界右移，这样，更多的样本点会被归入 w1 类，这是由于将 w1 误判成 w2 时造成的损失远小于 w2 误判成 w1 时的损失。尽管决策边界的右移会增大分类错误的概率，但却在理论上保证了每次决策承担的平均风险是最小的。实际上，最小错误可以视为最小风险的一种特殊情形，也就是所有的误判所

造成的损失是相同的，而最小风险是基于已有的经验和实际的情况对于后验概率进行加权修正的操作。最小风险将更多的实际因素纳入了考虑，它会放弃对于分类正确率的一位追求而偏向于做出更为稳妥的决策，从实验结果看，当误判所造成的损失差别不大时，最小错误和最小风险的分类结果差别不是很大，这时对于分类结果起主导作用的仍然是后验概率，而当不同种类误判产生风险差异性较大时，误判的损失对于分类会起到很强的影响作用时原先的决策边界发生显著的移动。最小风险原则做出决策的效果好坏非常依赖于对于风险的量化，如果量化非常准确，最小风险往往能够做出非常好的分类决策。

3.2.设计贝叶斯分类器对多维特征的数据进行多分类

此次实验针对于三类二维特征的样本点进行分类。

3.2.1 设计流程

（1）根据样本点的数目估计每个类别的先验概率 $P(w1), P(w2), P(w3)$

（2）假设类条件概率服从二元正态分布，依据样本数据对条件概率中未知参数进行极大似然估计，获得条件概率密度的极大似然函数

（3）由于样本点是二维特征，因此需要计算每个特征均值以及每个类别的协方差矩阵

均值的极大似然估计：$\mu_{ij} = \frac{1}{N_i}\sum_{x\in w_i} X$；$i = 1,2,3; j = 1,2$；

方差和协方差的极大似然估计：$s_{jk}^{i} = \frac{1}{N_i - 1}\sum_{l=1}^{N_i}(x_{lj} - \mu_j^{w_i})(x_{lk} - \mu_k^{w_i}) \qquad j,k = 1,2$

协方差矩阵：$\Sigma^i = \begin{pmatrix} \Sigma_{11}^i & \Sigma_{12}^i \\ \Sigma_{21}^i & \Sigma_{22}^i \end{pmatrix}$（i=1，2，3）

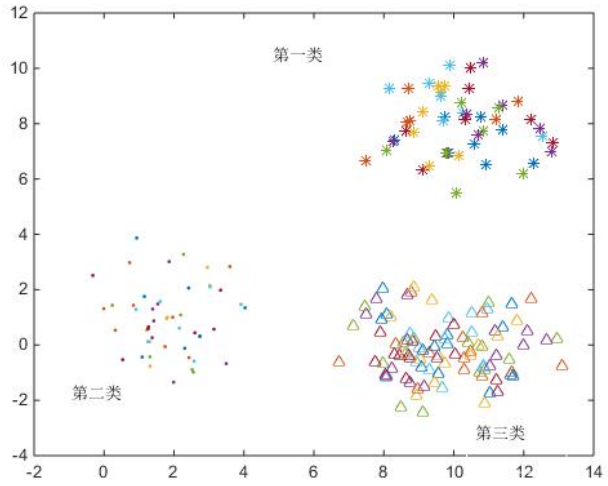
类条件概率的极大似然函数：$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right\}$

（Σ^{-1} 是Σ的逆矩阵，$|\Sigma|$是Σ的行列式）

由类条件概率与先验概率求得后验概率，在不同决策规则下做出决策。

3.2.2 按最小错误率的规则做出决策

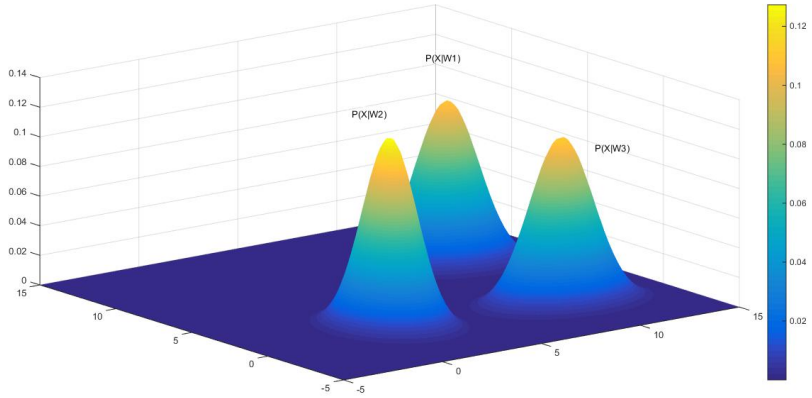
（1）样本数据的分布：



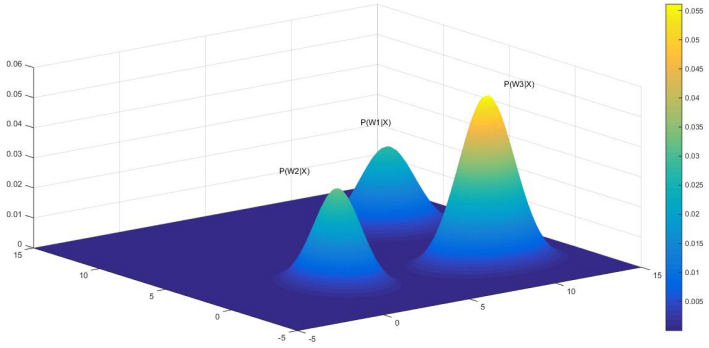
三类样本由二维高斯分布产生：

```
x1=mvnrnd([10 8],[2 0;0 1],50)';
x2=mvnrnd([2 1],[1 0;0 2],50)';
x3=mvnrnd([10 0],[2 0;0 1],100)';
```

（2）w1,w2,w3 类条件概率的似然函数：



（3）后验概率分布：



（4）判别函数的计算：

$g_i(X) = P(w_i)P(X \mid w_i); i = 1,2,3$

依据最小错误决策规则：

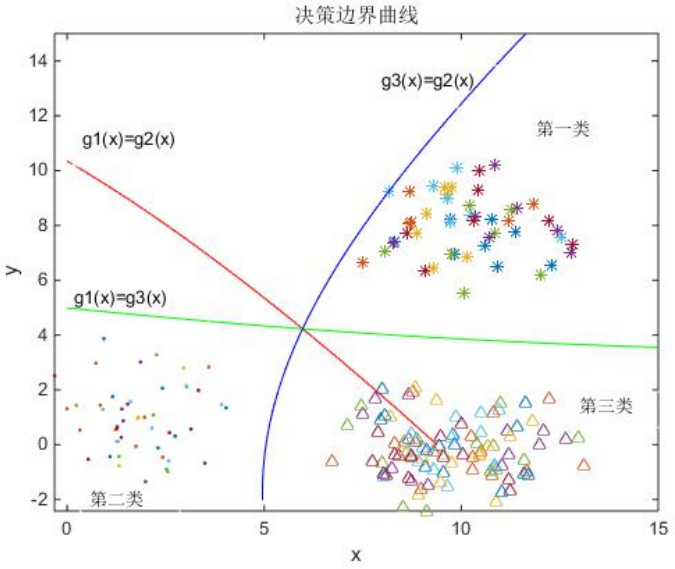
若$g_1(X) > g_2(X)$且$g_1(X) > g_3(X) \Rightarrow X \in w1;$

若$g_2(X) > g_1(X)$且$g_2(X) > g_3(X) \Rightarrow X \in w2;$

若$g_3(X) > g_1(X)$且$g_3(X) > g_2(X) \Rightarrow X \in w3;$

故决策边界为：$g_1(X) = g_2(X), g_2(X) = g_3(X), g_3(X) = g_1(X)$

（5）依据后验概率对样本进行分类，求取决策边界：



每两条决策边界会划分出一个类别所在的区域，由上图可见，在最小错误规则下，通过贝叶斯方法做出的决策边界对于该数据集中的每个样本点均可做出正确的分类。

3.2.3 根据最小风险规则做出决策：

（1）假设误判时的损失：

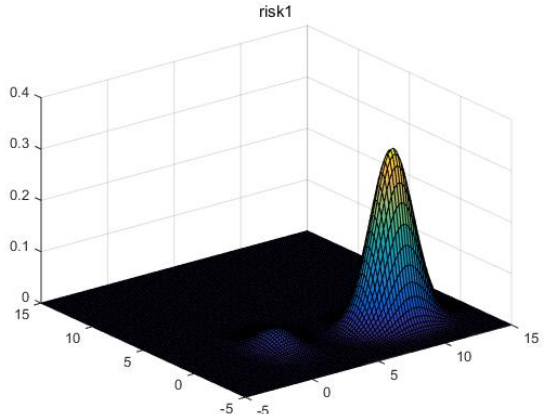
实际 / 判断	W1	W2	W3
a1	0	2	6
a2	1	0	1
a3	6	1	0

风险函数的计算：

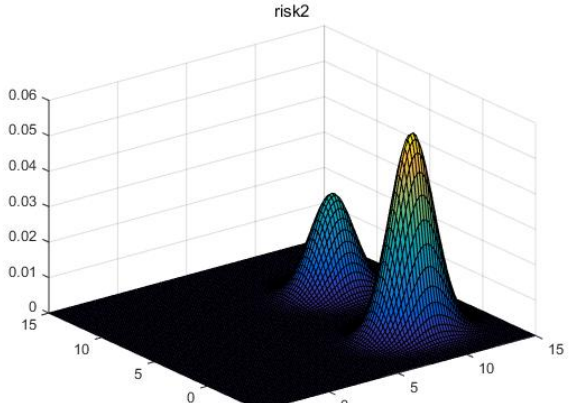
$$risk(1) = \lambda_{11}P(w1 \mid x) + \lambda_{12}P(w2 \mid x) + \lambda_{13}P(w3 \mid x) = 2 * P(w2 \mid x) + 6 * P(w3 \mid x)$$
$$risk(2) = \lambda_{21}P(w1 \mid x) + \lambda_{22}P(w2 \mid x) + \lambda_{23}P(w3 \mid x) = P(w1 \mid x) + P(w3 \mid x)$$
$$risk(3) = \lambda_{31}P(w1 \mid x) + \lambda_{32}P(w2 \mid x) + \lambda_{33}P(w3 \mid x) = 6 * P(w1 \mid x) + P(w2 \mid x)$$

（2）风险函数曲线：

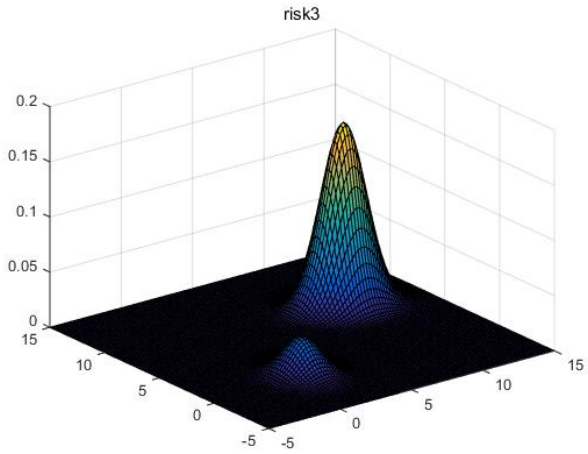
Risk1:



Risk2:



Risk3:



（3）根据风险函数的大小，在最小风险规则下做出决策：

若$risk1 < risk2$且$risk1 < risk3 \Rightarrow X \in w1$;

若$risk2 < risk1$且$risk2 < risk3 \Rightarrow X \in w2$;

若$risk3 < risk1$且$risk3 < risk2 \Rightarrow X \in w3$;

4.问题讨论

二分类问题是分类问题中最简单的一种情形，只需通过一个判别函数就可以求得决策边界从而对样本进行分类，在该实验中，通过贝叶斯方法解决二分类问题使用的判别函数是两类后验概率或者两类损失函数求差，根据结果为正还是为负来判断所属类别，决策边界为使判别函数等于 0 的点的集合。多分类问题的解决有多种思路，但都离不开二分类的思想。

思路一：我们可以将多分类问题拆解成若干个二分类，然后采用二分类方法进行分类：将 N 个类别两两配对，产生 $N(N-1)/2$ 个二分类任务，根据 $N(N-1)/2$ 个分类的结果选择被预测最多的那个类别作为多分类的结果。

思路二：仿照二分类的方法，直接计算样本点属于每个类别的后验概率，$P(w1 \mid X), P(w2 \mid X) \cdots P(w_n \mid X)$，选择后验概率最大的那个类别作为分类结果。

显然第一种思路下需要产生 $N(N-1)/2$ 个分类器，每次分类用到了两个类别的样本；第二种思路需要产生 N 个分类器，每个分类器都需要使用所有的训练样本。在类别很多，样本容量很大时，第一种思路更具优势，每个分类器使用的样本数目少，训练时间更短，但对于该实验，类别总共只有三类，且样本数目较少，故考虑使用第二种更为简单的思路。实际工程实践中，我们应根据样本类别和数目的多少，妥善地选择分类方法，提高分类器的性能。

5.总结

本次实验通过贝叶斯决策的方法实现了对给定样本的二分类，对比了在最小错误规则和最小风险规则下分类结果的差异性，同时尝试使用贝叶斯方法解决多分类问题，完成了对三类样本的正确分类，验证了贝叶斯分类方法对于解决模式分类问题的效果。同时研究了先验概率、类条件概率、风险等参数对于分类起到的作用，加深了对贝叶斯决策思想的理解。

Github 网址：https://github.com/ECHOliuty/PR-report