性别分类实验报告

2013011570 自 31 唐静娴

2016/10/11

一、实验方法与数据说明

1. 实验方法和原理

由于人的身高、体重数据满足正态分布，因此不妨采用多元正态分布的概率密度函数来作为类条件概率密度函数，其表达式为：

$$p(x \mid \omega_I) = \frac{1}{(2\pi)^{d/2} \mid \sum_i \mid^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)\right]$$

本次实验只涉及到 2 类对象，即“男性”和“女性”，先验概率根据经验取为 $P(w_1) = P(w_2) = 0.5$。一般情况下，这两类对象身高、体重协方差矩阵是不等的，因此判别函数式可以化简为

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i) - \frac{1}{2}\ln \mid \sum_i \mid + \ln P(\omega_i) \quad , i = 1,2$$

其中，

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$
$$w_i = \Sigma_i^{-1}\mu_i$$
$$w_{i0} = -\frac{1}{2}\mu_i^T \Sigma_i^{-1} \mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

对于某个待分类的 x，分别计算$g_i(x)$，$g_j(x)$，若 $g_i(x)/g_j(x) > thre$，则认为 x 属于第 i 类，否则认为 x 属于第 j 类。其中 $i, j \in \{1,2\}$，thre 代表阈值。

本次实验中定义了 lamda 矩阵作为损失函数矩阵，在一个函数中实现最小错误率和最小风险贝叶斯决策。该做法是基于“最小错误率贝叶斯决策看作是最小风险贝叶斯决策的特例”这一理论基础。进行最小错误率贝叶斯决策时，将 lamda 矩阵赋值为反单位阵，即$\lambda_{11} = \lambda_{22} = 0$，$\lambda_{12} = \lambda_{21} = 1$；进行最小风险贝叶斯决策时，将 lamda 矩阵赋值为$\begin{bmatrix} 0 & 5 \\ 4 & 0 \end{bmatrix}$。

绘制 ROC 曲线时，为了遍历 Sp 的取值可能性，将阈值 thre 从 0.1 开始，以 0.1

为步长逐步取值到 10，得到 100 对(Sp,Sn)对，并以此绘制曲线。

2. 数据来源

采集小组名单：张骁骏、唐静娴、孙超逸、陈熙

采集对象：家人、同学、朋友

采集过程：网上问卷发放与收集

3. 程序来源

自行编写

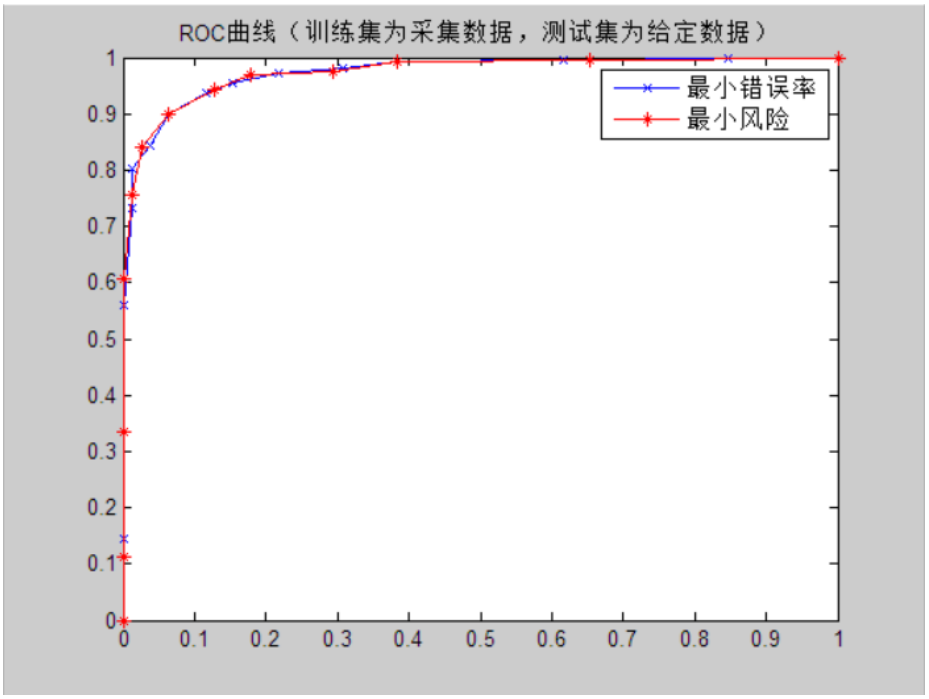
二、实验内容与结果

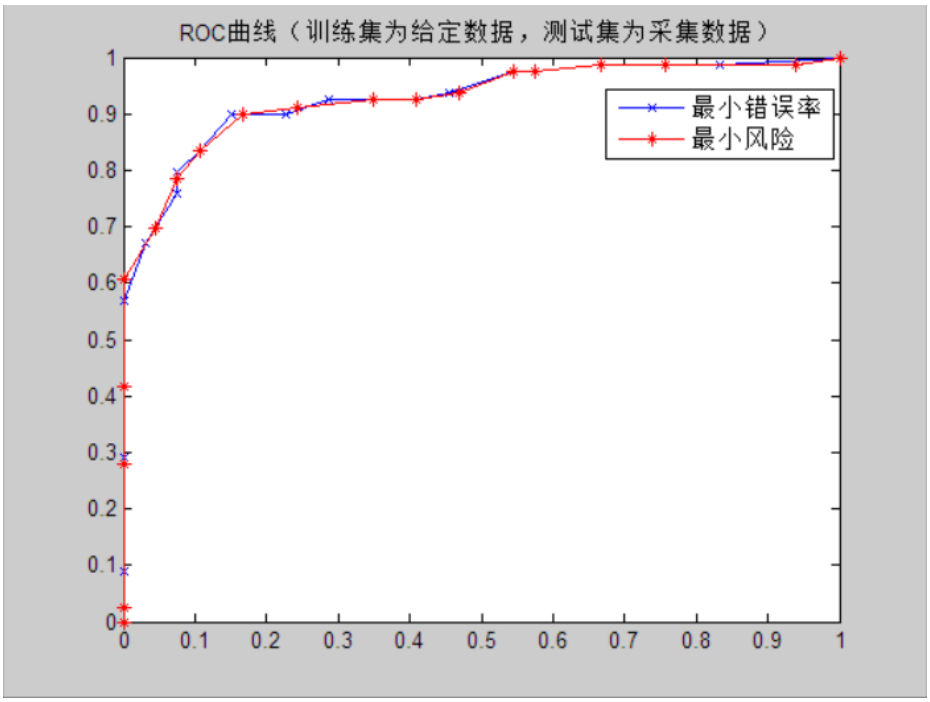
两种贝叶斯决策得到的结果如下：

```
*********最小错误率贝叶斯决策结果*********
A.训练集为采集数据，测试集为给定数据
男性错误率：0.1   女性错误率：0.0641
B.训练集为采集数据，测试集为采集数据
男性错误率：0.177   女性错误率：0.106
C.训练集为给定数据，测试集为采集数据
男性错误率：0.165   女性错误率：0.106
D.训练集为给定数据，测试集为给定数据
男性错误率：0.165   女性错误率：0.106

*********最小风险贝叶斯决策结果*********
A.训练集为采集数据，测试集为给定数据
男性错误率：0.032   女性错误率：0.179
B.训练集为采集数据，测试集为采集数据
男性错误率：0.0886   女性错误率：0.273
C.训练集为给定数据，测试集为采集数据
男性错误率：0.0886   女性错误率：0.242
D.训练集为给定数据，测试集为给定数据
男性错误率：0.0886   女性错误率：0.242
```

以分类结果男性作为阳性，横坐标为 1-特异度，纵坐标为灵敏度，改变阈值，得到 ROC 曲线如下，可以看出曲线都在对角线左上角，且曲线越靠近左上角，说明方法的性能越好。





三、分析与体会

从本次实验过程中，我体会到决策和分类过程非常依赖先验知识，改变先验概率或是损失函数矩阵，都会导致分类结果发生极大的变化，甚至是出现极端结果。此外，通过调整决策阈值，可以兼顾假阳性率和假阴性率，达到错误率的不同情况，满足某些特定的实际需求。