用身高体重数据进行性别分类实验报告

2013012245
基科 31 白可
2015 年 10 月 8 日

实验内容：

- 利用最小错误率贝叶斯决策解决根据身高体重进行性别分类的问题
- 利用最小风险贝叶斯决策重复上述实验。自行确定决策表
- 粗略画出 ROC 曲线

实验过程：

- 利用收集到的数据，估计正态分布的参数μ, σ^2。并求出各类别的条件概率概率密度$p(x|w_i)\ (i = 1,2)$。
- 估算各类别的先验概率$P(\omega_i)$，统计训练集和测试集的正确率
- 制定决策表，再次计算训练集和测试集的正确率
- 改变阈值，统计 TP,FP,FN,TN, 计算假阳性率和真阳性率（这里假设男性为“阳性”）

实验数据：

一、最小错误率贝叶斯分类器

由于关于最小错误率贝叶斯书上已经有详细的推导过程，我这里就不再赘述。

样本：共计 62 份样本，其中男性 25 人，女性 37 人。样本年龄在 18——22 岁之间。

方程:

$$\mu_{mh} = 176.96$$
$$\sigma_{mh}^2 = 33.78$$
$$\mu_{mw} = 68.6$$
$$\sigma_{mw}^2 = 9.09$$
$$r_m = \begin{bmatrix} 1 & 0.5437 \\ 0.5437 & 1 \end{bmatrix}$$

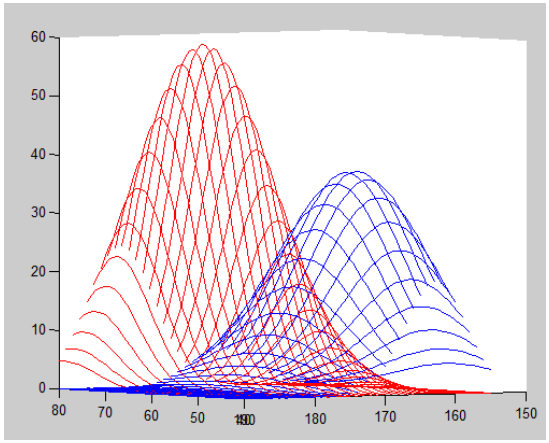
最终，男性的 Guassian 分布为：

$$P_{male} = \frac{1}{0.0169}\exp\left(-\frac{1}{2\left(1 - r_m^{1,2^2}\right)}\left(0.0296(x - 176{,}96)^2 - 0.022\right.\right.$$
$$\left.\left.(x - 176.96)(y - 68.6) + 0.0138(y - 68.6)^2\right)\ \right)$$

同理算出女性的 Guassian 分布为。

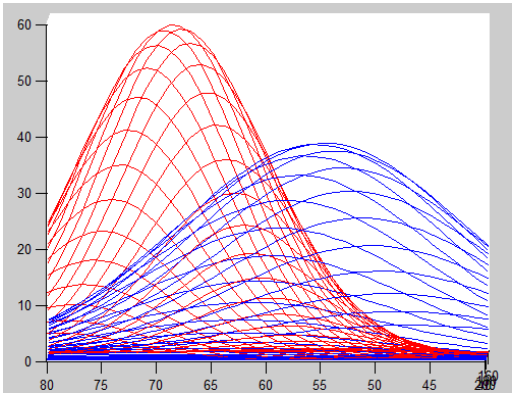
$$P_{female} = \frac{1}{0.0159}\exp\left(-\frac{1}{2\left(1-r_w^{1,2^2}\right)}\left(0.0474(x-164.35)^2 - 0.096\right.\right.$$
$$\left.\left.(x-164.35)(y-55.16) + 0.0058(y-55.16)^2\right)\right)$$

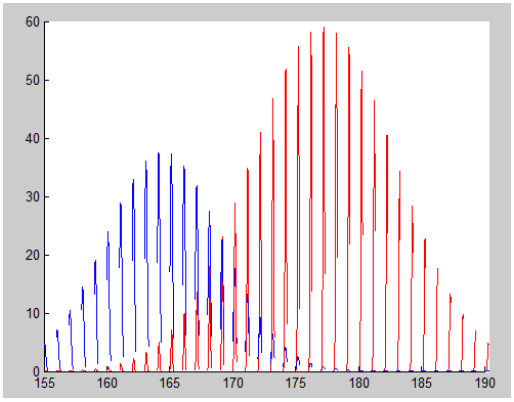
使用 Plot3 函数画出的图像是这样的：



由于是立体图像，因此不便于展示...

不过如果我们分体重和身高两个维度分别观察：





左侧为体重，右侧为身高，可以看到男女的身高差异和体重差异。

先验概率（男\女）	训练集正确率	测试集正确率
25\37	90.32%	86.28%
1\1	90.32%	90.24%
2\1	90.32%	91.77%
6\1	74.19%	93.60%

且男女的身高体重本身就存在着较大的差异，这是这个分类器效果比较好的一个重要原因。

当先验概率取成训练集的男女比时，测试集正确率偏低，猜测可能是因为测试集的男女比偏低，那么男女比较小的训练集样本可能会不太适合。

但从另一方面分析，样本点更多就统计意义上就更加准确，意味着可能方差会更小，乘上训练集的较大的先验概率可能会更加准确。

然后 ，我尝试着改变了先验概率。但这种改变不是以“得到最佳准确率”为目的的，因为就算是在这个测试集上达到了比较好的效果，在其他的也不一定。我仅仅是通过猜测不同的具有具体含义的先验概率来验证其正确率。

a. 如果在不知道“测试集”的性质为“清华学生”的情况下，我们这时候较为保守的选择时选择先验概率为 0.5/0.5。

b. 但是如果考虑到“清华学生”这一属性，我将先验概率调整至 0.66/0.33(清华男女比)，测试集正确率迅速上升。

c. 继续探索，考虑到自动化等理工科院系的男女比，正确率进一步上升…但是训练集准确率有所下降。

这也启示我：在采集样本的时候，尽量均匀准确。

二、最小风险贝叶斯决策

清华的理工院系中，女生人数较少，男女比例大于 4:1，我们不妨假定将一个男生判断为女生的代价为将一个女生判断为男生的 0.25。

决策	男(pre)	女(pre)
男(fact)	0	1
女(fact)	4	0

根据实验一，我们选取先验概率为 33%/66%一组。得到结果如下：

先验概率（男\女）	训练集正确率	测试集正确率
2\1	90.32%	86.28%

此时正确率并不能说明什么了

关键在下表中：

决策	男(fac)	女(fac)
男(pre)	208（TP）	3 (FP)
女(pre)	75(FN)	42(TN)

决策	男(fac)	女(fac)
男(pre)	223（TP）	5 (FP)
女(pre)	73(FN)	27(TN)

可以看到，女生只有 3 个人被误判为男生。不加 cost 时，为 5 人，误判率下降。

三、ROC 曲线

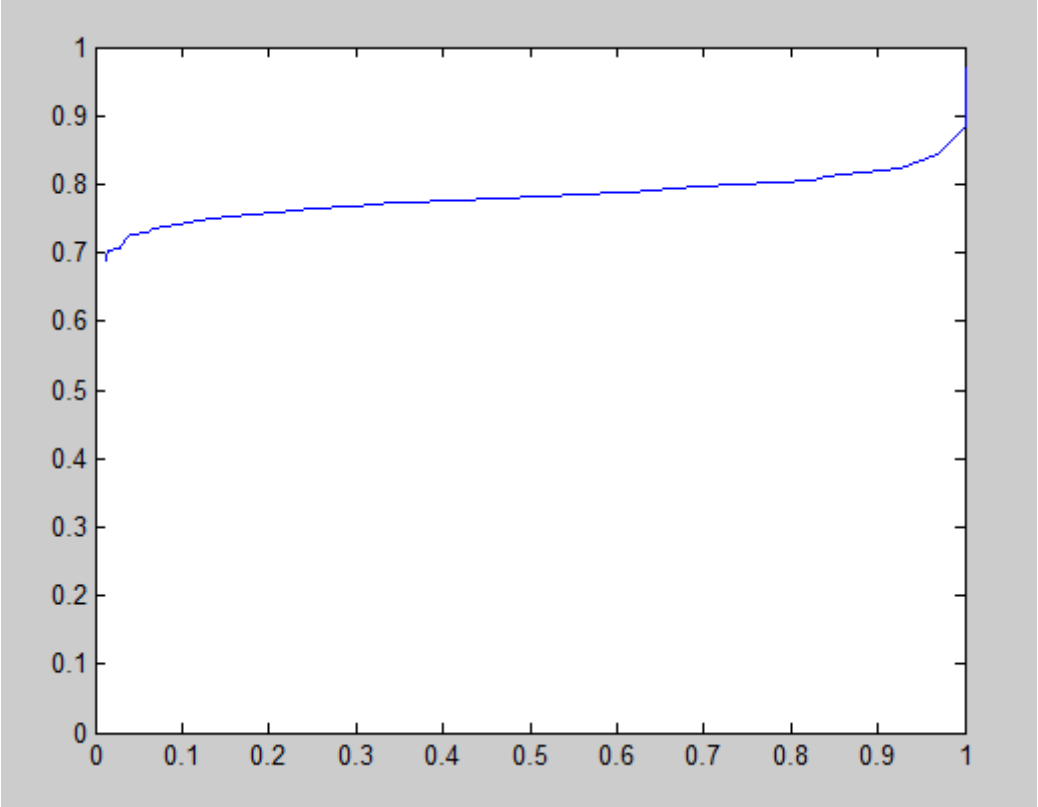
我将 t 的值从 0.01 以 0.01 为步长一直取到了 5 得到的 ROC 图如下所示。红色段是我补上去的，基本与实验一实验二相吻合。

实验二所在的点为：（0.067，0.735）处，此时假阳性率较低，但是真阳性率也较低。这就意味着，如果我们要求尽可能少的女生被误判成为男生，那么男生被判断为男生的准确率必然下降。

实验一所在的点为：（0.37,0.77）处，此时真阳性上升，假阳性也上升，这就意味着如果我们需要有更高的精度（更多的男生被判断为男生），就意味着更高的假阳性（更多的女生被误判成为男生）。

考虑到整个图像处在 y=x 的上部，因此，这个决策器整体的性能还不差。

两侧出现明显截断，是因为数据量不够的原因造成。此图只是一个粗略估计。



四、实验方法和数据说明

小组成员：白可（基科 31） 张一铭（基科 32） 张格菲（基科 31）

采集对象：张一铭的同学，白可的同学，共计 62 份样本，其中男性 25 人，女性 37 人。

采样方式：网上问卷调查。

程序来源：自己写的。

（注：采集数据检查后发现有两个明显的错误，于是更正，已经通知其他同学，但不确定他们是否记得更正）