Pattern Recognition: Homework #7

Professor Wang Guijin

Qingyun Fang

2017311003

Use LaTeX in 5.6 2018

Problem One

正态分布时分类

设有符合正态分布的两类训练样本，并且设 $P(\omega_1) = P(\omega_2) = 0.5$，且 ω_1 和 ω_2 满足下列关系：

$$\omega_1 = \{(3,4)(3,8)(2,6)(4,6)\}$$
$$\omega_2 = \{(3,0)(3,-4)(1,-2)(5,-2)\}$$

求识别函数、识别界面方程、作图。

Solution

识别函数　　在多元正态函数概率型 $(p(\boldsymbol{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, ..., c)$，其判别函数为：

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}{\boldsymbol{\Sigma}_i}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_i|) + \ln P(\omega_i)$$

将题目中的数据代入可得：

参数	数值	参数	数值
$\boldsymbol{\mu}_1$	$(3,6)^{\mathrm{T}}$	$\lvert\boldsymbol{\Sigma}_1\rvert$	1
$\boldsymbol{\mu}_2$	$(3,-2)^{\mathrm{T}}$	$\lvert\boldsymbol{\Sigma}_2\rvert$	4
$\boldsymbol{\Sigma}_1$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}$	${\boldsymbol{\Sigma}_1}^{-1}$	$\begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}$
$\boldsymbol{\Sigma}_2$	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	${\boldsymbol{\Sigma}_1}^{-1}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$

表 1: 参数表

因此 $g_1(\boldsymbol{x})$ 和 $g_2(\boldsymbol{x})$ 如下：

$$g_1(\boldsymbol{x}) = -(x_1 - 3)^2 - 0.25(x_2 - 6)^2 - \ln 2\pi + \ln 0.5$$
$$g_2(\boldsymbol{x}) = -0.25(x_1 - 3)^2 - 0.25(x_2 + 2)^2 - \ln 2\pi - \frac{1}{2}\ln 4 + \ln 0.5$$

识别界面

显然识别界面方程为 $g_1(\boldsymbol{x}) = g_2(\boldsymbol{x})$, 将其化简可得：

$$y = \frac{3}{16}(x - 3)^2 - \frac{\ln 4}{8} + 2$$

界面显示

最终的识别结果如下：

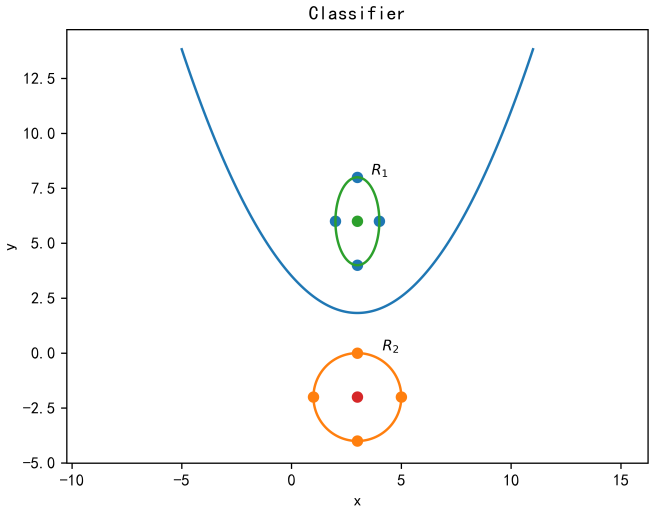

图 1: 识别结果

根据上图中 R_1 和 R_2 分别用来表征两类数据的等概率密度线，两类密度在 y 方向上具有相同的方差，仅在 x 上 $P(\omega_1)$ 比 $P(\omega_2)$ 小，因此 x 大的样本可能来自于 ω_1，并且其决策面为抛物线。

贝叶斯分类下错误率计算

在几类协方差不相同的情况下，其错误率很难用公式得出，因此在这里可以考虑用蒙特卡洛的方法来计算错误率。随机生成两维的正态分布数据，其方差和均值都满足表 1。生成的数据分布如下图 2：

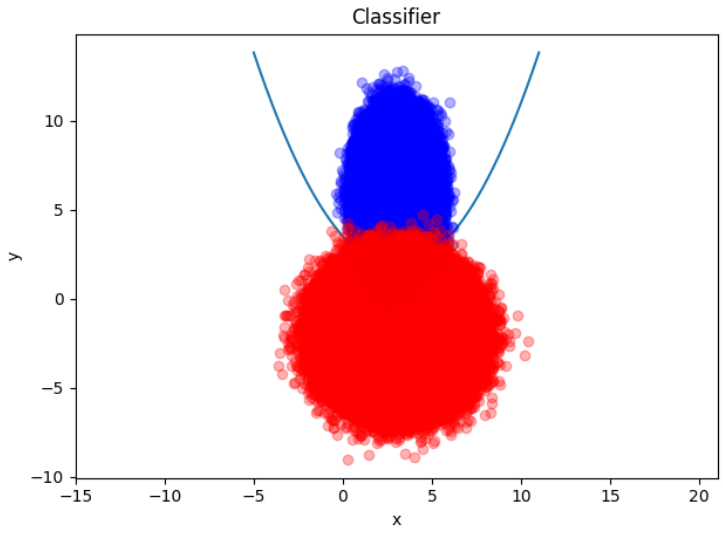

图 2: 随机数据分布

随机生成 2000000 个数据，每类都 1000000 个，统计错误率如下：

类别	分类错误个数	精确率	召回率
ω_1	2001	99.7944%	99.7999%
ω_2	2056	99.7999%	99.7944%

表 2: 参数表

运用蒙特卡洛法，得出的精确率和召回率都极其优秀，显示在数据分布为正态分布的前提下，运用贝叶斯决策的方法是十分有效的。

与 Fisher 准则对比

Fisher 准则虽然不对数据做任何假设，但是在很多情况下，投影到一维空间后其样本分布呈现一种正态分布的情况，显然二维空间中正态分布的数据在任意投影方向下都应该是正态分布的, 因此 Fisher 准则处理这种数据应该有一种天然的优势。

使用 fisher 准则分类，其投影方向为 y 轴，分界面为 $y = 2$，这一点从直观也可以判断出来，往 y 轴投影是类间距离最大的。fisher 分类结果如下图 3，其中黄色的直线为 fisher 分界面：

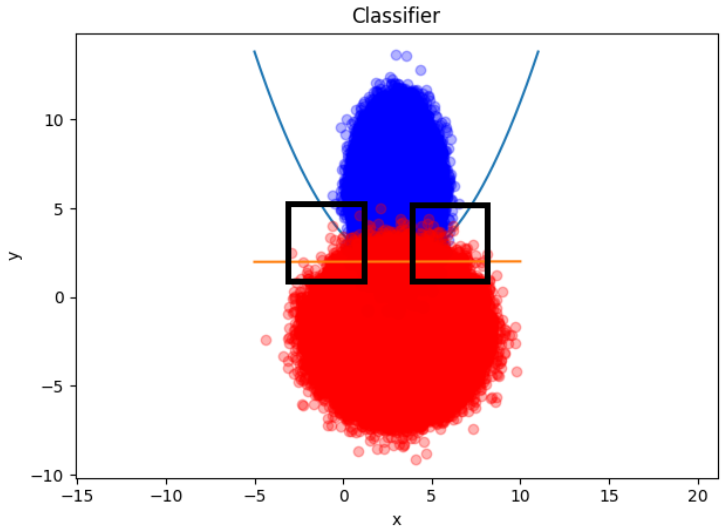

图 3: Fisher 准则判断

接下来我们在统计一下 Fisher 的精准度和召回率，实验参数和之前的贝叶斯法相同：

类别	分类错误个数	精确率	召回率
ω_1	2343	99.7773%	99.7657%
ω_2	2262	99.7657%	99.7738%

表 3: 参数表

对比两种方法，贝叶斯的方法比 Fisher 准则的精度提升了一点，由于这些方法都是逼近与 100% 的准确率，因此很难有大的提升了。在图 3 中用黑框标记出来了，这部分数据就是贝叶斯比 Fisher 准则优势的地方，性能的提升也就是靠这部分数据。