目录

最小距离分类

1 问题描述

w1 = {(2,3); (2,2); (2,4); (3,3); (3,4); (2.5,3); (1.5,2); (3.5,2.5); (4,4); (0.5,0.5)}

w2 = {(0,2.5); (-2,2); (-1,-1); (1,-2); (3,0); (-2,-2); (-3,-4); (-5,-2); (4,-1)}

求：

用最小距离分类判别方法时的识别函数。

最小距离分类判别方法时的识别界面。

画出该识别界面将训练样本的区分结果图示。

2 算法说明

2.1 方法 1

记两类样本集分别是 S_1 和 S_2，对应的样本数目分别是 N_1 和 N_2。类的样本中心（平均）由公式(1)算出。

$$M_i = \frac{1}{N_i}\sum_{x\in S_i} x \quad (i=1,2) \tag{1}$$

根据题目中给出的数据集，样本均值是

$$\begin{aligned} M_1 &= \left(\frac{12}{5},\frac{14}{5}\right)^T \\ M_2 &= \left(-\frac{5}{9},-\frac{5}{6}\right)^T \end{aligned} \tag{2}$$

相应的线性的识别函数是

$$d_i(x) = x^T M_i - \frac{1}{2}M_i^T M_i \quad (i=1,2) \tag{3}$$

公式(3)中，代入 $x=(x_1,x_2)^T$，可得两类的线性识别函数表达式分别是

$$\begin{aligned} d_1(x) &= \frac{12}{5}x_1 + \frac{14}{5}x_2 - \frac{34}{5} \\ d_2(x) &= -\frac{5}{9}x_1 - \frac{5}{6}x_2 - \frac{325}{648} \end{aligned} \tag{4}$$

识别界面即为 $d_1(x)=d_2(x)$，代入公式(3)，可得分界面方程是

$$x^T(M_1-M_2) - \frac{1}{2}\left(M_1^T M_1 - M_2^T M_2\right) = 0 \tag{5}$$

代入题中数据，可得分界面方程是

$$\frac{133}{45}x_1 + \frac{109}{30}x_2 - \frac{20407}{3240} = 0 \tag{6}$$

此时，对未知模式 x 的判定依据是

$$\begin{cases} d_1(x) > d_2(x) & \Rightarrow \quad x \in C_1 \\ d_1(x) < d_2(x) & \Rightarrow \quad x \in C_2 \end{cases} \tag{7}$$

程序处理时，由于计算得到的直接是未知模式与类中心的距离，因此应选取使得该距离最小的类。

2.2 方法 2

本实验是个二分类问题。进一步的，我们假设 $p = (p_1, p_2)^T$ 是分界面上的任一点，即 p 满足

$$p^T (M_1 - M_2) - \frac{1}{2}\left(M_1^T M_1 - M_2^T M_2\right) = 0 \tag{8}$$

由公式(8)即可得到，

$$\left(p - \frac{M_1 + M_2}{2}\right)^T (M_1 - M_2) = 0 \tag{9}$$

公式(9)意味着，两类的分界面正是线段 $M_1 M_2$ 的垂直平分线。因此，对未知模式分类时，我们也可以先得到该分界面，然后判断位置模式 x 与哪个类中心在分界面同一侧。

若记函数 $f(x)$ 为

$$f(x) = x^T (M_1 - M_2) - \frac{1}{2}\left(M_1^T M_1 - M_2^T M_2\right) \tag{10}$$

则对未知模式 x 的判定依据是

$$\begin{cases} f(x) f(M_1) > 0 & \Rightarrow \quad x \in C_1 \\ f(x) f(M_2) > 0 & \Rightarrow \quad x \in C_2 \end{cases} \tag{11}$$

实际上，从理论上来看，上述两个方法是等价的，只是在算法流程上有所差别。它们都是基于最小欧氏距离来分类的。

3 结果分析

程序 `main4.m` 和 `main4s.m` 分别实现了前面介绍的方法 1 和方法 2。

图 1 是方法 1 得到的分类结果，图 2 是方法 2 得到的分类结果。图中各元素所代表的含义在图 2 的图例中给予了说明。

两个程序的输出结果分别是

```
sep-hyper:
2.95555555555556*x1+3.63333333333333*x2-6.29845679012346=0

bias1: -6.8000
bias2: -0.5015
>>
```

```
paras for sep-hyper:
k = -0.8135  b = 1.7335
>>
```

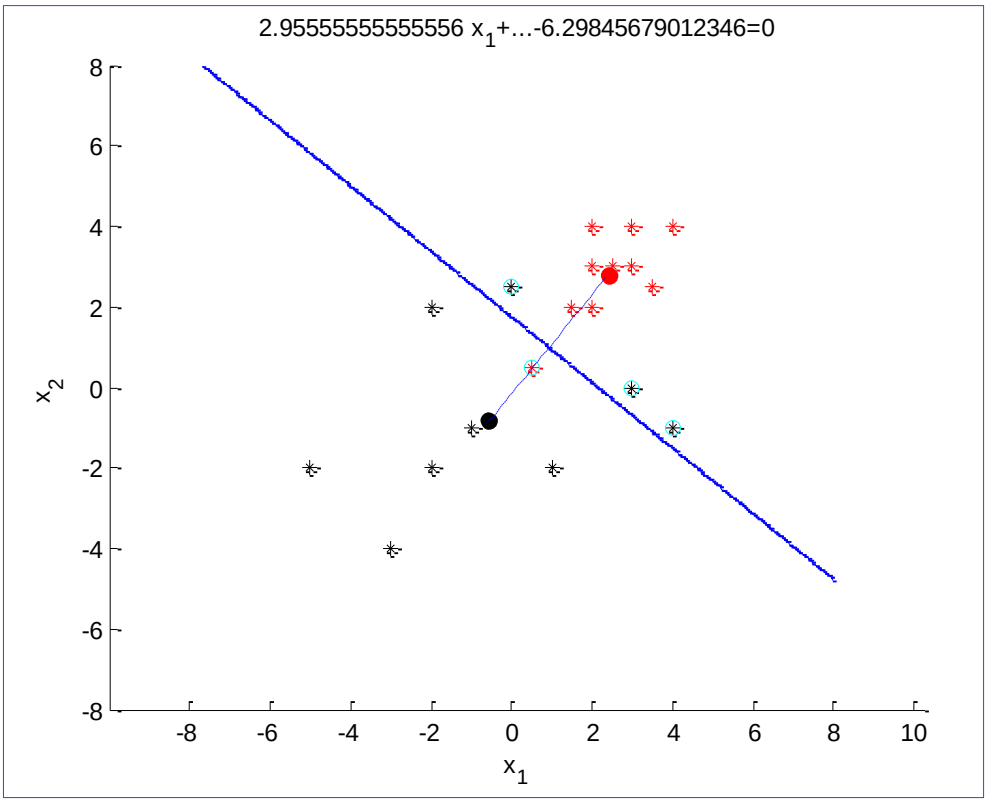

图 1 方法 1 运行结果

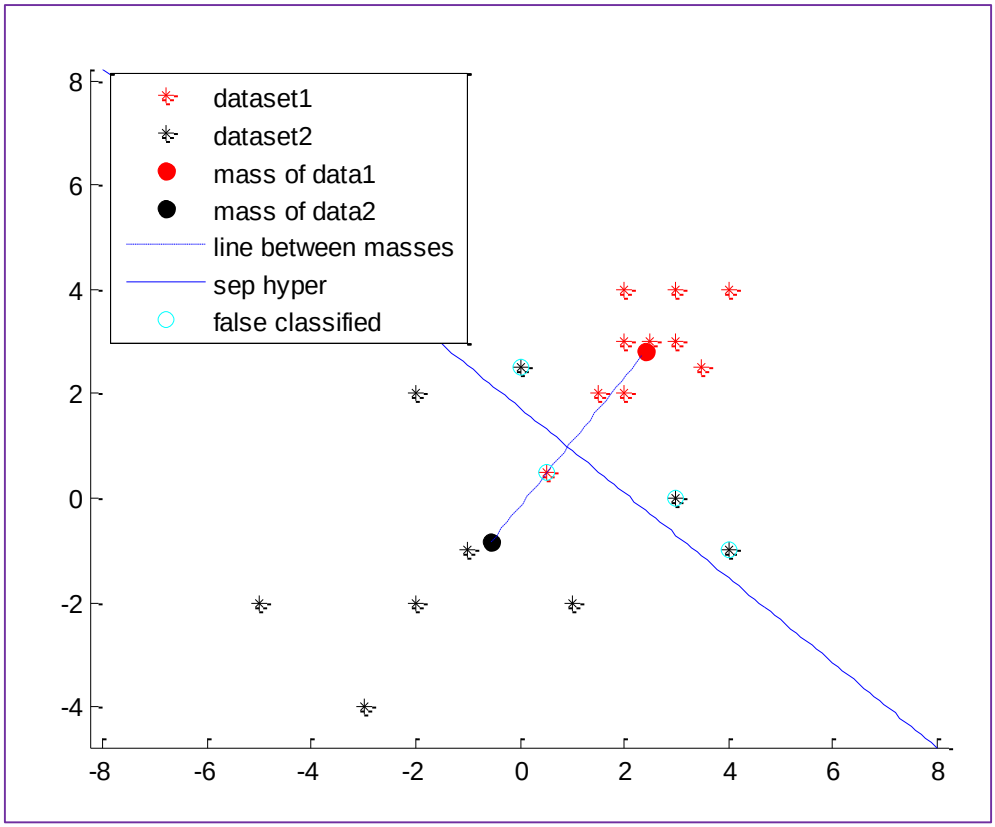

图 2 方法 2 运行结果

对结果的分析如下：

(1) 经验证，上述数据与算法描述中得到的线性识别函数表达式（公式(4)）和分界面方

程（公式(6)）是一致的。

(2) 方法 1 与方法 2 得到的分界面是一样的，这也说明了两个类中心的垂直平分线就是根据最小欧氏距离分类得到的分界面。这与我们的算法分析也是一致的。

(3) 该算法对题中给定的数据集的分类结果如下表所示。

	分类为 1	分类为 2
实际为 1	9	1
实际为 2	3	6

故，对类别 1 与类别 2 分类结果的评估如下。

类别	Precision	Recall	F-score
类别 1	0.750	0.900	0.818
类别 2	0.857	0.667	0.750

(4) 从整体评价来看，数据集 1 的分类结果更好一些。这从图中也可以看出来，因为数据集 1 的内聚性更高一些，而数据集 2 的分布则较为分散。最小欧氏距离分类更适合于类内距离尽可能小，类间距离尽可能大的数据集。