清华大学《模式识别》

作业 5

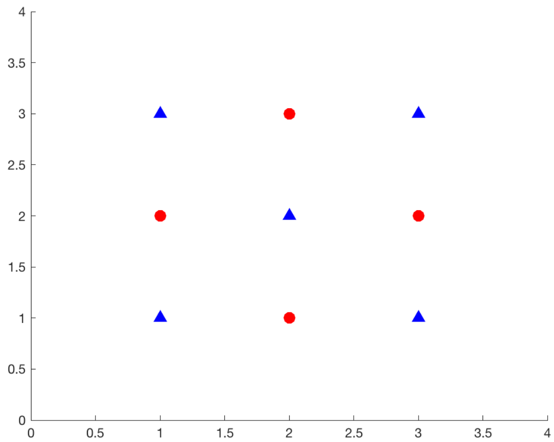
郭一隆

April 26, 2018

1 非线性分类器

1.1 问题描述

求分离 ● 和 ▲ 的函数。



1.2 求解思路

记 ▲ 为第 0 类， ● 为第 1 类。
从数据本身的分布特点考虑：

- 周期性:

  两类数据点在二维平面的分布具有很强的周期性。由于数据恰好分布在格点上，可考虑用奇偶性来分类。即对于数据 $\mathbf{x} = (x_1, x_2)$，其分类

$$i = (x_1 + x_2) \mod 2$$

  上述分类函数只能应对坐标为整数的数据，扩展至全平面:

$$i = (\text{round}(x_1) + \text{round}(x_2)) \mod 2$$

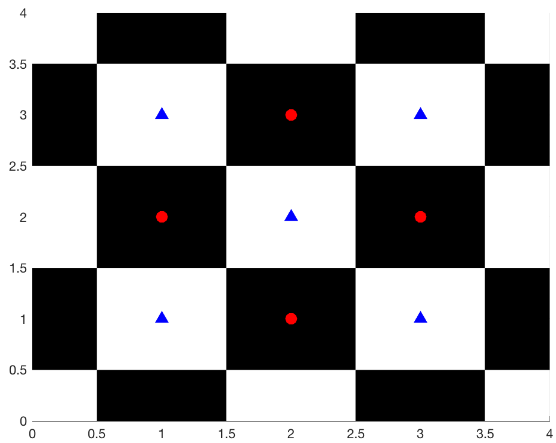


Figure 1: 取整奇偶性分类函数

- 对称性:

  由于数据点全体分别关于直线 $x_1 = 2$ 和直线 $x_2 = 2$ 对称，考虑映射

$$(x_1, x_2) \to (x_1', x_2') = (|x_1 - 2|, |x_2 - 2|)$$

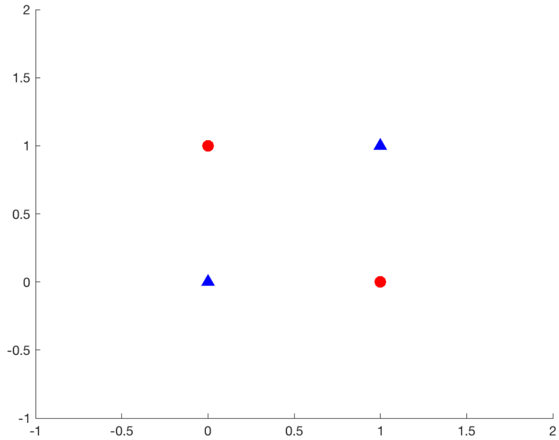


Figure 2: 经过对称折叠后的数据分布

得到的新数据分布（如图2），与课件中提到的广义线性识别函数的例子相同。可通过增加第三维度

$$x_3' = x_1' x_2'$$

进一步构造出高维平面

$$d(\mathbf{X}') = x_1' + x_2' - 2x_3' - \frac{1}{3} = x_1' + x_2' - 2x_1' x_2' - \frac{1}{3} = 0$$

映射回原坐标系即

$$d(\mathbf{X}) = |x_1 - 2| + |x_2 - 2| - 2|x_1 - 2||x_2 - 2| - \frac{1}{3} = 0$$

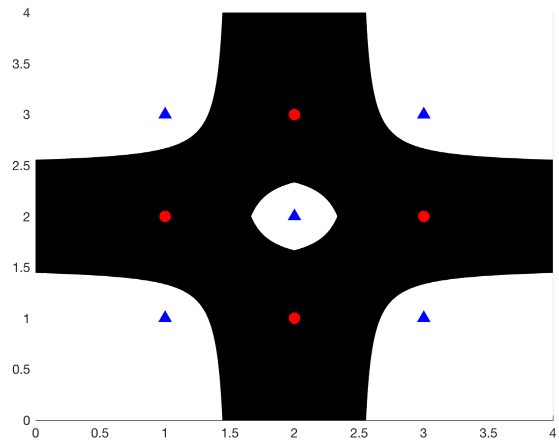


Figure 3: 对称双曲线分类函数

如图3所示，分类界面由对称的 4 支双曲线交叠而成。

- 两条直线分割:
  仍考虑映射
  $$(x_1, x_2) \to (x_1', x_2') = (|x_1 - 2|, |x_2 - 2|)$$
  得到如图2所示的数据分布，可考虑用两条直线所夹区域来识别:
  $$d(\mathbf{X}') = (x_1' - x_2')^2 - \frac{1}{4} = 0$$
  即
  $$d(\mathbf{X}) = (|x_1 - 2| - |x_2 - 2|)^2 - \frac{1}{4} = 0$$

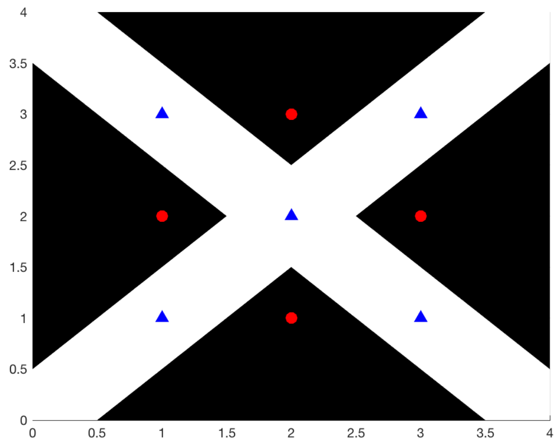


Figure 4: 多直线分类函数

2 迭代修正求权向量

2.1 问题描述

$$\omega_1 = \{(1,1), (2,0), (2,1), (0,2), (1,3)\}$$
$$\omega_2 = \{(-1,2), (0,0), (-1,0), (-1,-1), (0,-2)\}$$

- 迭代修正求权向量的方法，求上述两类的线性识别函数
- 迭代修正求权向量的方法，求上述两类的线性识别界面
- 画出该识别界面将训练样本的区分结果图示

2.2 求解思路

两类数据分布如图5所示。

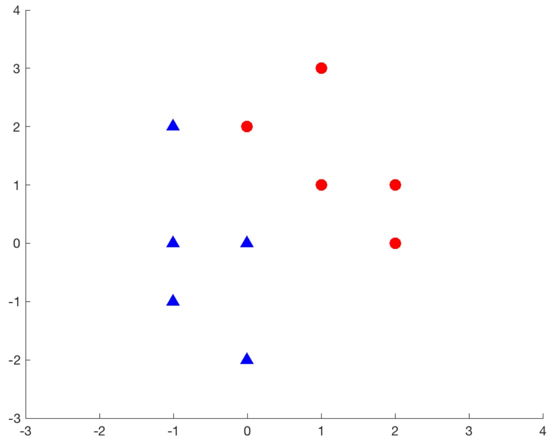


Figure 5: 数据分布

直观可判断两类线性可分，利用梯度下降法训练权向量可得：

$$\mathbf{W} = (3.30, 0.386, -0.439)$$

即线性分界面为

$$d(\mathbf{X}) = 3.30x_1 + 0.386x_2 - 0.439 = 0$$

如图6所示。

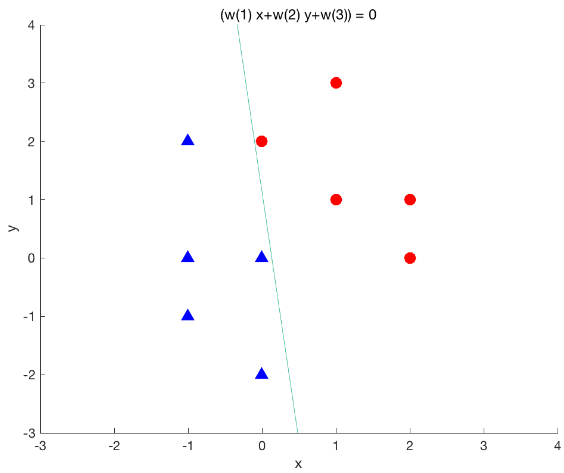


Figure 6: 线性分界面示意图

3 总结

- 第一题中采用两条直线区分课件样例，实际上两条直线即二次双曲线的极限退化形式。如果结合梯度下降法，并采用一组升维基底（升维至四次是可行的），也可以求得广义线性识别函数。
- 实际应用中应该尽量避免过拟合。