人工智能——寻找数据的规律

孔静—2014K8009929022

October 17, 2016

Contents

1 概 述

问题：

条件属性 X，continuous value, 值域为 [a,b], 决策属性 Y，值域为 $\{0,1\}$

已知一组数据：(x_i, y_i), 设计一个方法：计算出一个值 c，使得：在区间 [a,c] 和 [c,b] 上，X 与 Y 的变化规律一致

要求：

1. 描述事先的方法，可以使用描述性文字，把方法描述清楚
2. 设计一些数据的例子，画出对应的数据直方图，并显示计算出的区间的划分

2 分 析

idea：

背景：$similarity = cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$

简单起见，在区间 [a,c] 和 [c,b] 上，X 与 Y 的变化规律一致，那么从图像上来看，两段函数要尽可能地想象，由于 X 是连续递增的，两边相同；所以忽略 X，只考虑 Y，如果左右 Y 向量相似度 similarity 越大，说明变化规律越一致。

3 方 法

step. 1

选择一个合适的区间大小，对 X 进行划分，计算每一段 X 上 Y 的平均值，并画出相应的折线图。

step. 2

取 [a,c] 上按顺序取均值如 $(Y_1, Y_2, ..., Y_i)$, 视为向量 A, 在 [c,b] 上的折线图里等间距取同个数即 i 个 Y 值, 视为向量 B, 计算 similarity(A,B)。同理在 [c,b] 上取剩下的均值点视为向量 C, 在 [a,c] 等间距取同样个数均值点视为向量 D, 计算 similarity(C,D)。

两者相加，和最大的，即为相似度最高的，即为我们所寻找的 c 点。

step. 3

可利用二分查找法寻找，先取中点，再去左右部分中点进行比较，若左边大，选择左边继续查找。

Algorithm 1 Find The C

```
procedure FIND THE C(X, Y)
    Δx = Choose(X)
    Drawpicture(X, Y, Δx)
    mid = (left + right) / 2
    while left < right do
        leftmid = (left + mid) / 2
        rightmid = (mid + right) / 2
        if similarity(rightmid) > similarity(leftmid) then
            left = mid
        else
            right = mid
        end if
    end while
    return mid
end procedure
```

4 实 例

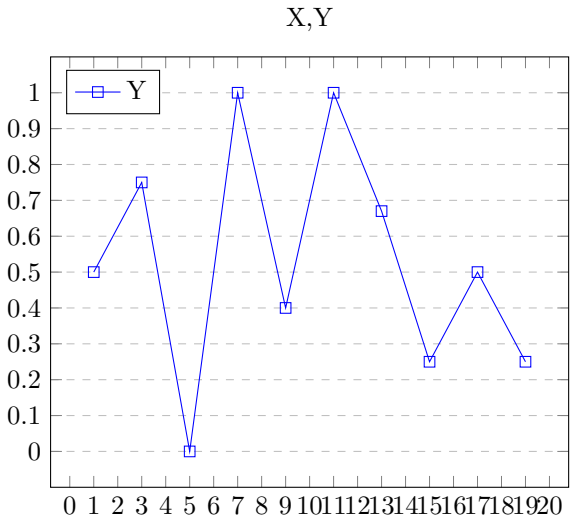
数据

X	1.1	1.6	2.1	3.4	3.5	3.9	5.1	7	8.1
Y	0	1	1	1	0	1	0	1	0
X	9	9.5	9.9	10	11.5	12.7	13.4	13.7	14
Y	1	0	1	0	1	1	1	1	0
X	14.3	14.5	15.1	16.6	17.1	18.4	18.5	19.9	20
Y	0	0	1	0	1	1	0	0	0

划分

X	1	3	5	7	9	11	13	15	17	19
Y	0.5	0.75	0	1	0.4	1	0.67	0.25	0.5	0.25

折线图



结果

在整数精度下，程序运行结果：9