《大数据机器学习》第 3 次作业

姓名：刘培源　　学号：2023214278

题目 1： Minsky 与 Papert 指出：感知机因为是线性模型，所以不能表示复杂的函数，如异或 (XOR)。验证感知机为什么不能表示异或。

答： XOR 函数的定义如表1所示：

x_1	x_2	$x_1 \oplus x_2$
0	0	0
0	1	1
1	0	1
1	1	0

Table 1: XOR 函数，红色代表负类，蓝色代表正类

基于以上表格，我们假设点 $(0,0)$ 和 $(1,1)$ 代表负类 0，点 $(0,1)$ 和 $(1,0)$ 代表正类 1。感知机的模型定义如下：

$$f(x) = sign(\omega \cdot x + b), \quad sign(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \tag{1}$$

其中 $\cdot$ 代表向量点乘。我们采用反证法进行证明，假设感知机可以模拟异或运算，且 $x = [x_1, x_2]$，$\omega = [\omega_1, \omega_2]$，为了满足以上四个点的分类，我们需要有：

- 要想 $x_1 = 0, x_2 = 0, f(x) = 0$，则需要 $b < 0$。
- 要想 $x_1 = 0, x_2 = 1, f(x) = 1$，则需要 $w_2 + b > 0$。
- 要想 $x_1 = 1, x_2 = 0, f(x) = 1$，则需要 $w_1 + b > 0$。
- 要想 $x_1 = 1, x_2 = 1, f(x) = 0$，则需要 $w_1 + w_2 + b < 0$。

显然，由前三点可以推出 $w_1 > 0, w_2 > 0, w_1 + b > 0$，所以 $w_1 + w_2 + b > 0$，这与第四点 $w_1 + w_2 + b < 0$ 矛盾，假设不成立，所以感知机不可以模拟异或运算。

进一步，我们可以把以上四个点可视化如图1，从图中可以看出，在二维平面内，没有一个超平面可以将这两类分开。

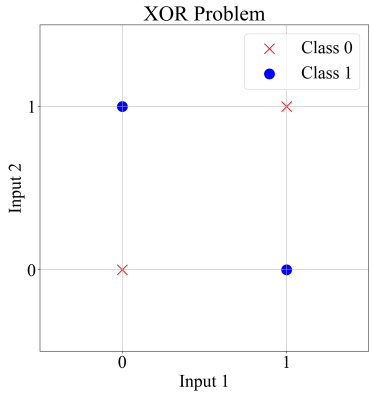

Figure 1: XOR 分类问题可视化

题目 2： 利用课本例题 3.2 构造的 kd 树求点 $x = (3, 4.5)^T$ 的最近邻点。

答： 课本例题 3.2 构造的 kd 树如图2所示。

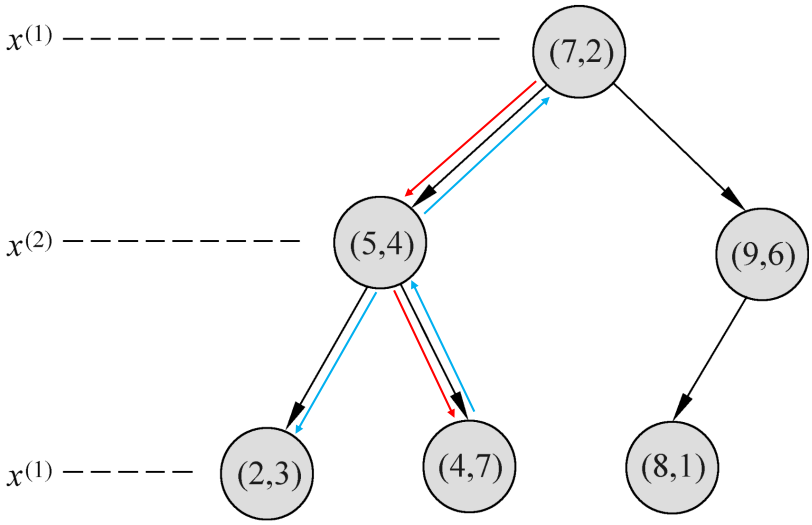

Figure 2: 课本例题 3.2 构造的 KD 树。红线代表对于点 $(3, 4.5)^T$ 进行正向搜索最近邻的过程，蓝线代表对于点 $(3, 4.5)^T$ 进行回溯搜索最近邻的过程。

求点 $x = (3, 4.5)^T$ 的最近邻点的搜索过程如下：

1. **从根节点开始：**
   - 当前节点：$(7, 2)^T$。
   - 由于 $3 < 7$，移动到左子节点。
2. **第二层：**
   - 当前节点：$(5, 4)^T$。
   - 由于 $4.5 > 4$，移动到右子节点。

3. **第三层：**

   - 当前节点：$(4,7)^T$。
   - 作为叶节点，将其设为当前最近的邻居，并计算其到查询点的距离：$\sqrt{(4-3)^2+(7-4.5)^2} \approx 2.69$。

4. **开始回溯：**

   - 回溯到节点 $(5,4)^T$。
   - 以查询点 $(3,4.5)^T$ 为中心，画一个半径为 2.69 的圆。
   - 检查节点 $(5,4)^T$ 是否在该圆内。实际上在圆内，因此需要检查左子树。

5. **检查左子树：**

   - 左子节点：$(2,3)^T$ 在圆内。
   - 计算此节点到查询点的距离，并发现它比当前最近的邻居 $(4,7)^T$ 更近。
   - 更新最近邻为 $(2,3)^T$。

6. **继续回溯：**

   - 回溯到根节点 $(7,2)^T$。
   - 检查节点 $(7,2)^T$ 是否在该圆内。实际上不在圆内，因此无需检查右子树。

7. **搜索结束：**

   - 已检查所有可能的路径，确定点 $(3,4.5)^T$ 的最近邻是点 $(2,3)^T$。