

03 k近邻算法

从邻居投票到 kd 树检索

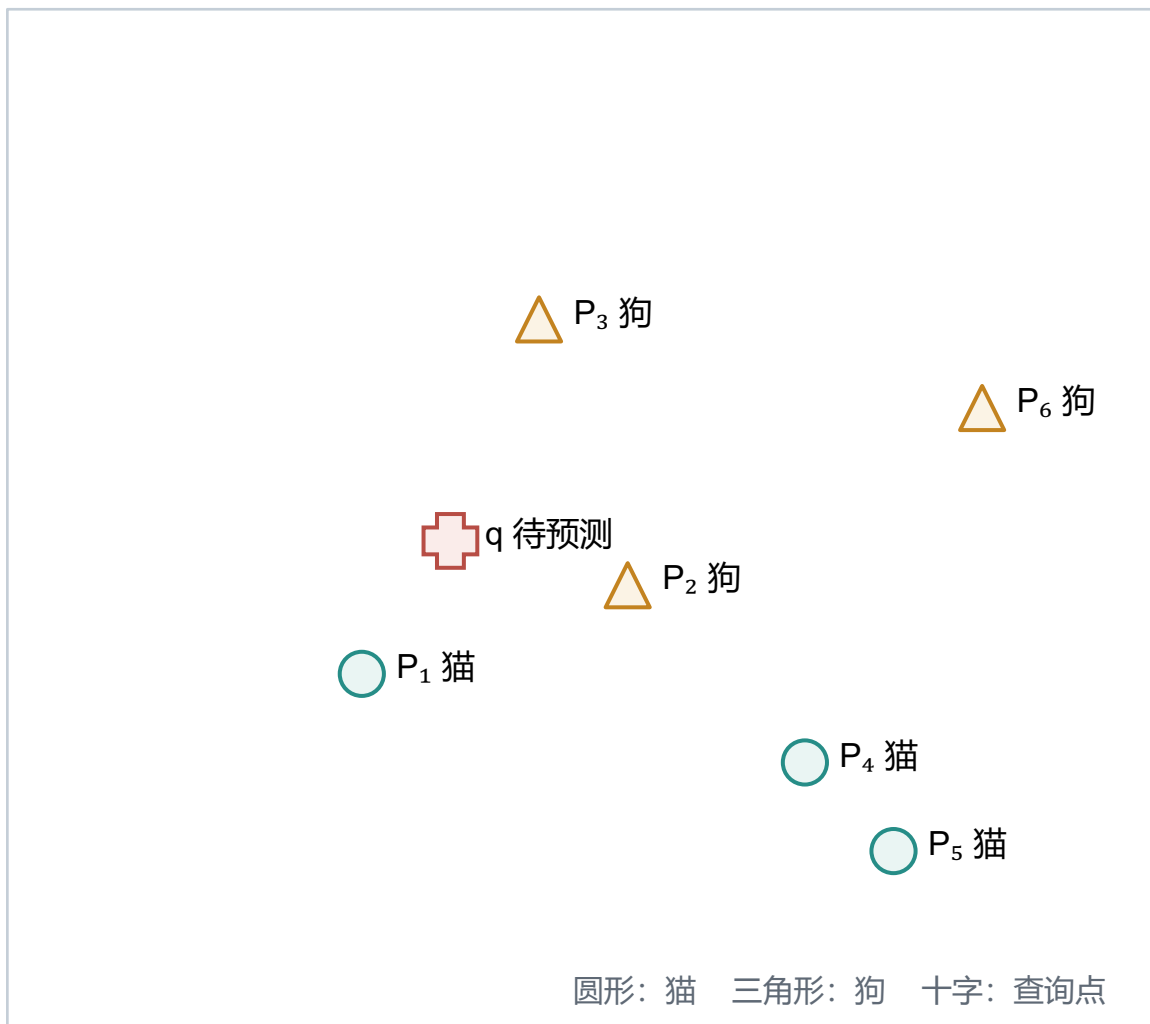
1 k 近邻预测

从附近样本、邻居数量到多数投票

k 近邻分类的基本过程

- **1 衡量距离**: 比较查询样本与带标签训练样本的距离
- **2 选择邻居**: 按距离从小到大排序, 取最近的 k 个训练样本
- **3 多数投票**: 统计这些邻居的标签, 用多数标签作为预测类别
- **影响因素**
 - 特征及其尺度决定 “多近”
 - k 决定 “参考多少个邻居”

已缩放的教学特征 2



已缩放的教学特征 1

■ 训练样本

- 六个样本已有“猫”或“狗”的标签

■ 查询任务

- 查询点 $q = (3, 4.5)$ 的类别未知

■ 特征空间

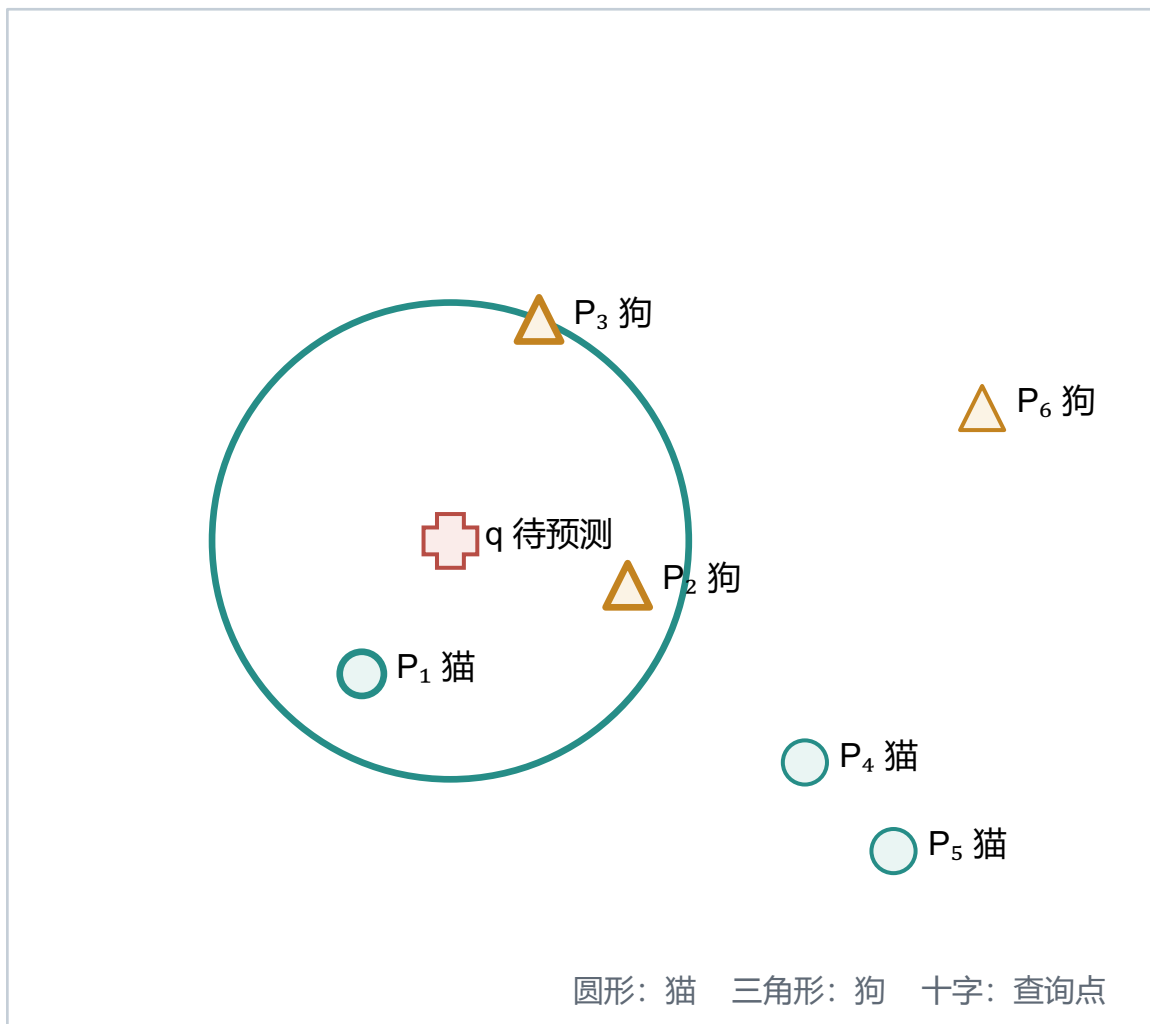
- 两个坐标轴表示已经缩放的教学特征

分类问题

哪些训练样本最能帮助判断 q 的类别?

k 近邻分类规则

已缩放的教学特征 2



已缩放的教学特征 1

检索约定

- 用欧氏距离衡量查询点与训练样本的接近程度
- 选择最近的 $k = 3$ 个样本

邻居与预测

- 最近三个样本为 $P_1(1.80)$ 、 $P_2(2.06)$ 、 $P_3(2.69)$
- 三个标签分别是猫、狗、狗，多数为狗

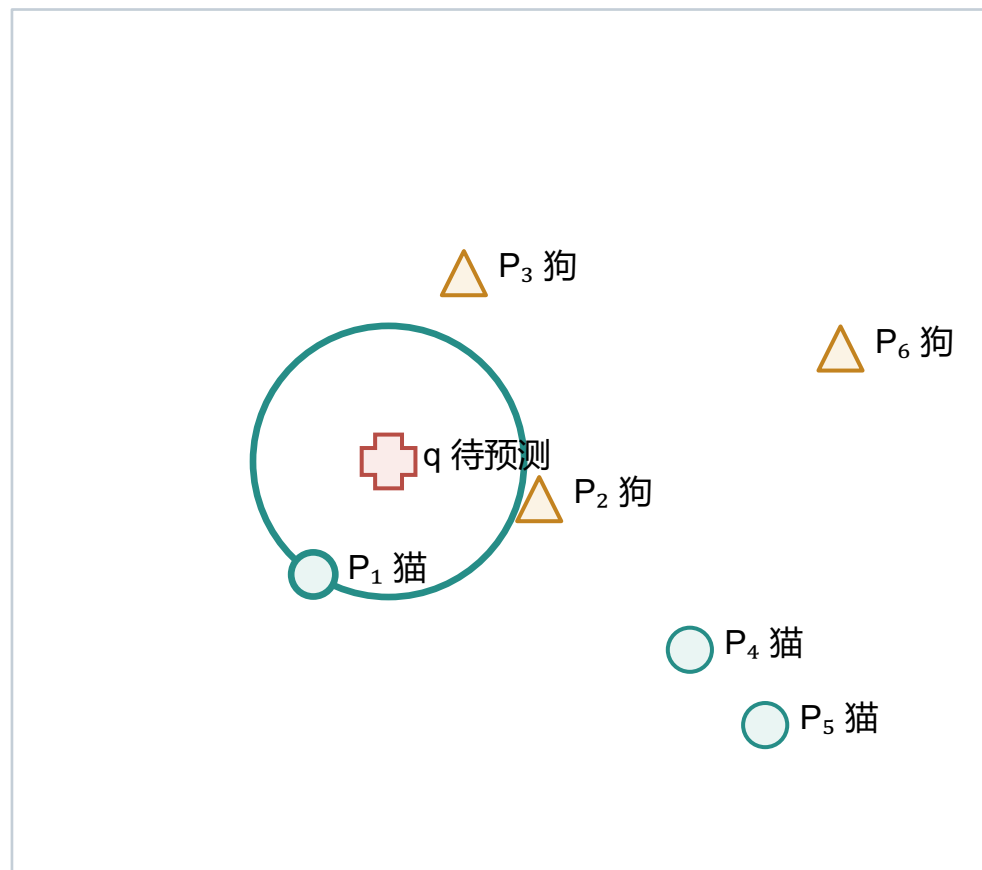
$$P_1: 1.80 < P_2: 2.06 < P_3: 2.69$$

$$P_4: 4.72 < P_5: 6.10 < P_6: 6.18$$

猫、狗、狗 多数为狗

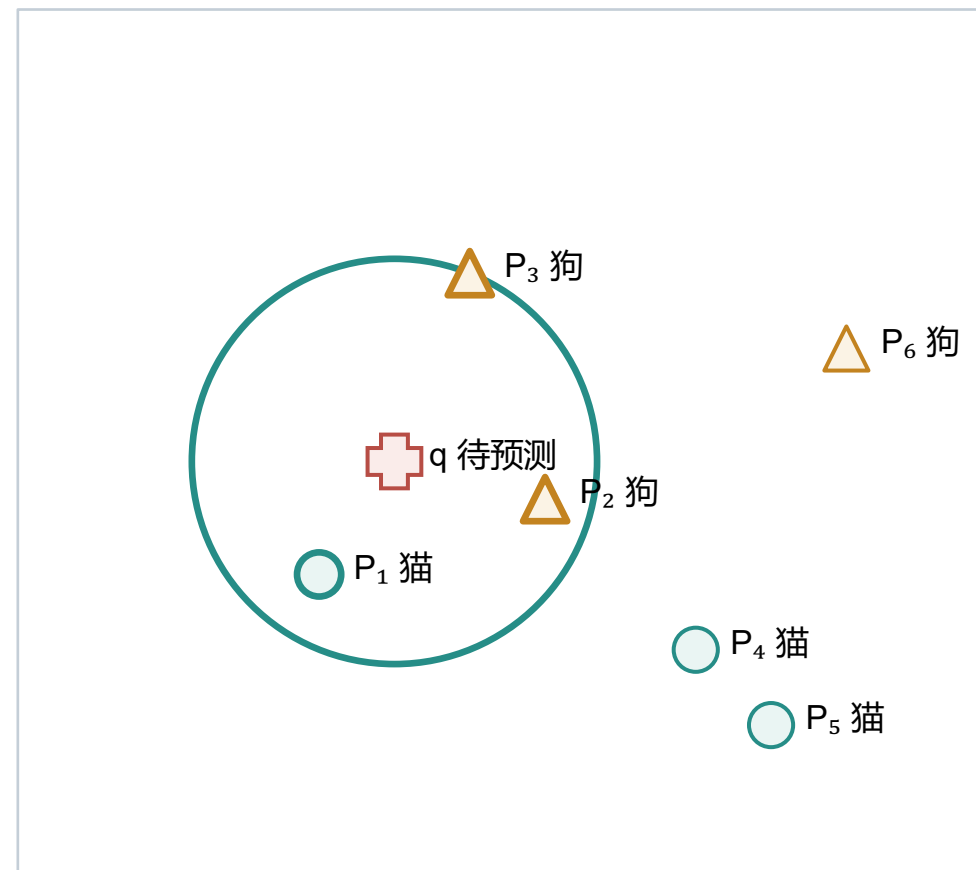
邻居数量对预测结果的影响

$k = 1$



最近邻是 P_1 , 预测为猫

$k = 3$



三个邻居中两只狗, 预测为狗

比较结论: 特征和尺度决定距离的含义; k 决定本次投票采用多大的局部范围

检索职责

- 直接搜索逐一检查全部样本，提供精确近邻基准
- kd 树组织空间，试图安全排除不可能更近的区域

预测职责

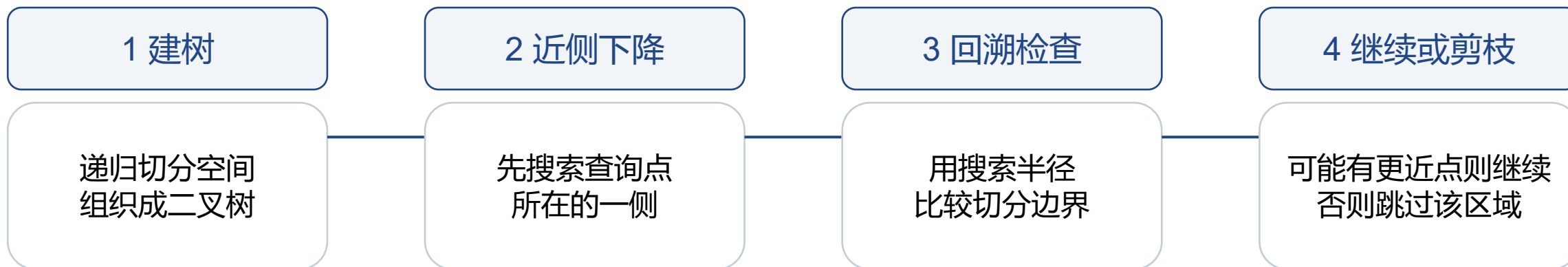
- k 近邻规则决定找到邻居以后怎样形成预测



两条检索路径在同一组精确邻居处汇合；投票规则保持不变

2 kd 树构造与检索

用空间划分组织样本，并通过回溯保证精确查询

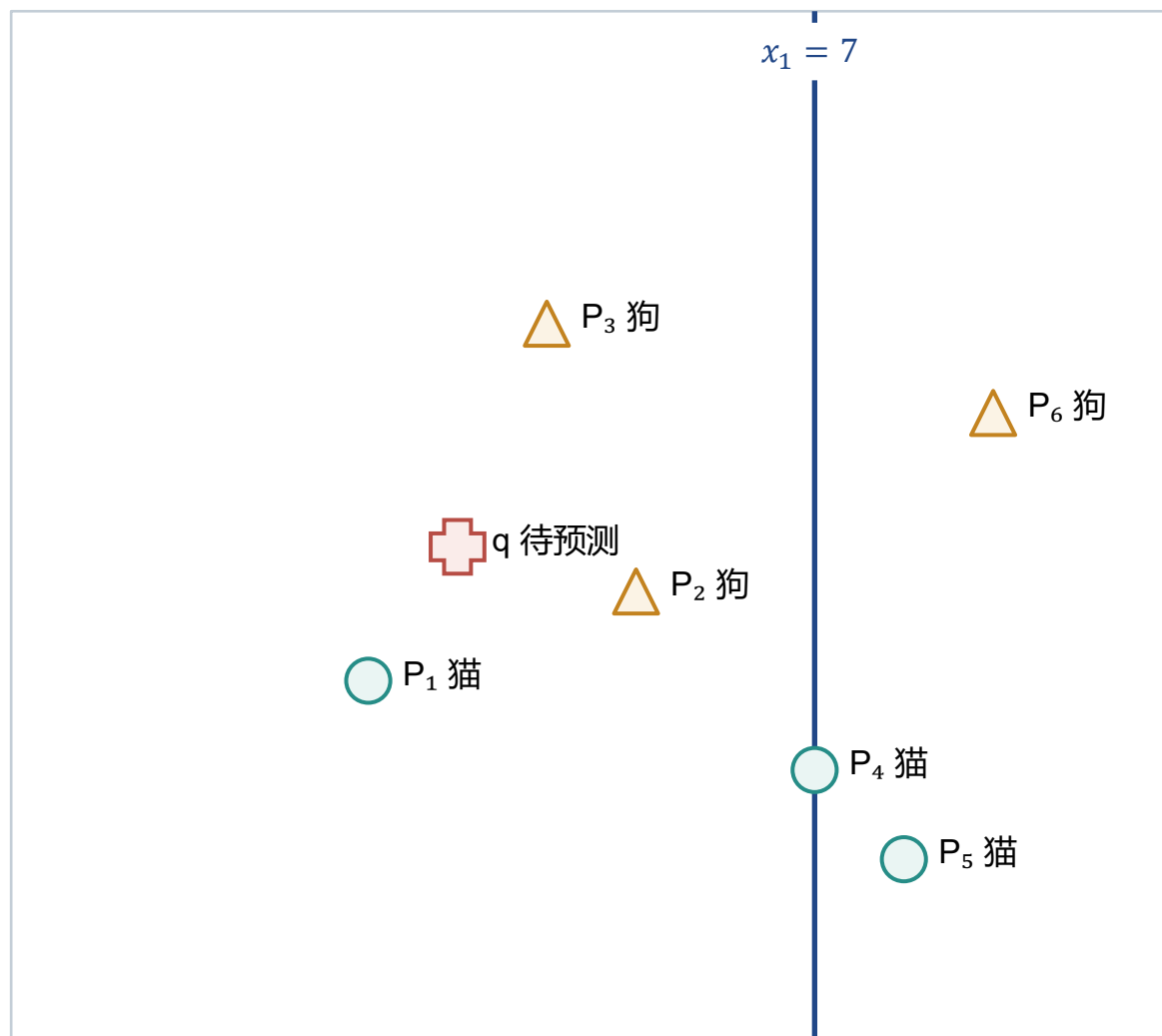


完整回溯后，精确 kd 树检索应返回与直接搜索相同的邻居

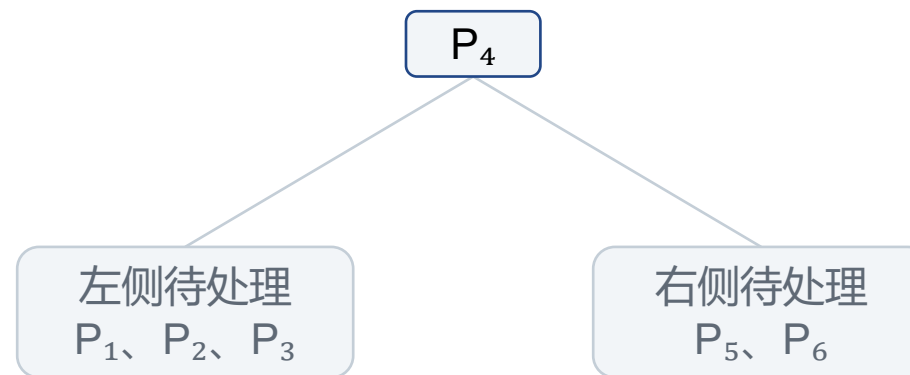
检索结构只改变“怎样找到邻居”，不改变邻居定义和投票规则

kd 树的根节点与第一次切分

已缩放的教学特征 2



已缩放的教学特征 1



■ 构造约定

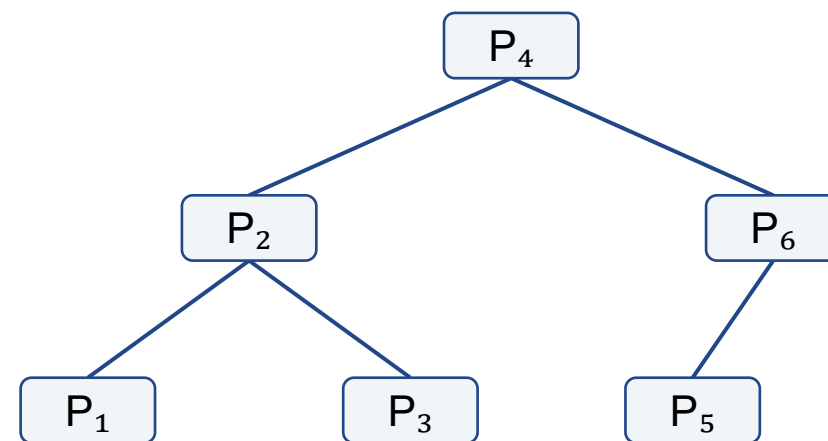
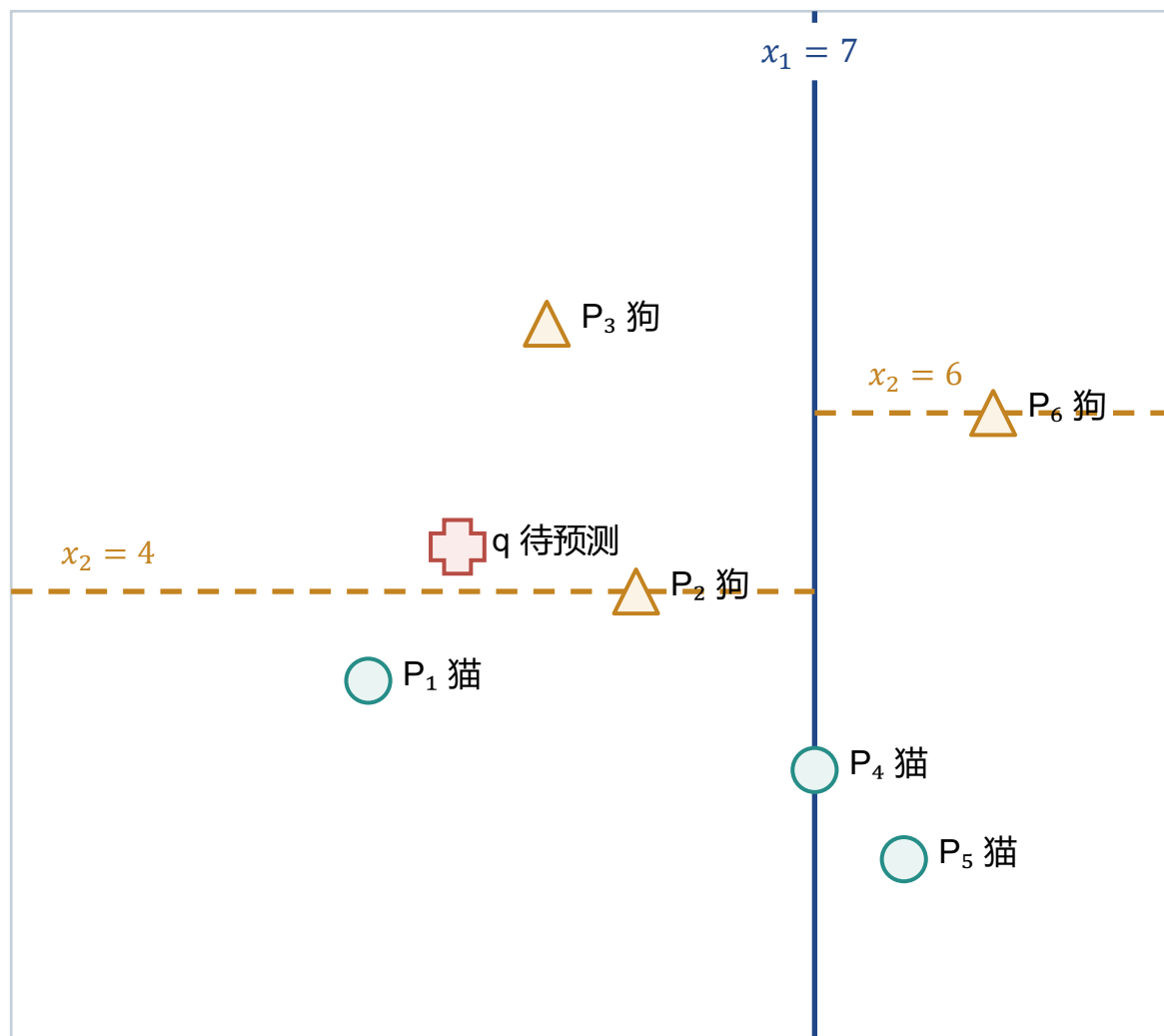
- 按特征 1、特征 2、特征 1 的顺序循环选择切分维度
- 点数为偶数时，选择排序后的上中位数

■ 第一次切分

- 根节点选择 P_4 ，切分边界为 $x_1 = 7$
- 左侧包含 P_1 、 P_2 、 P_3 ，右侧包含 P_5 、 P_6

kd 树的递归切分

已缩放的教学特征 2



递归切分

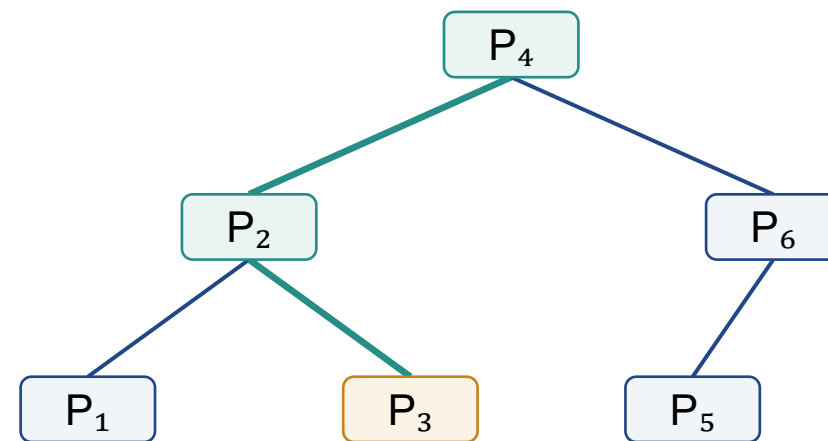
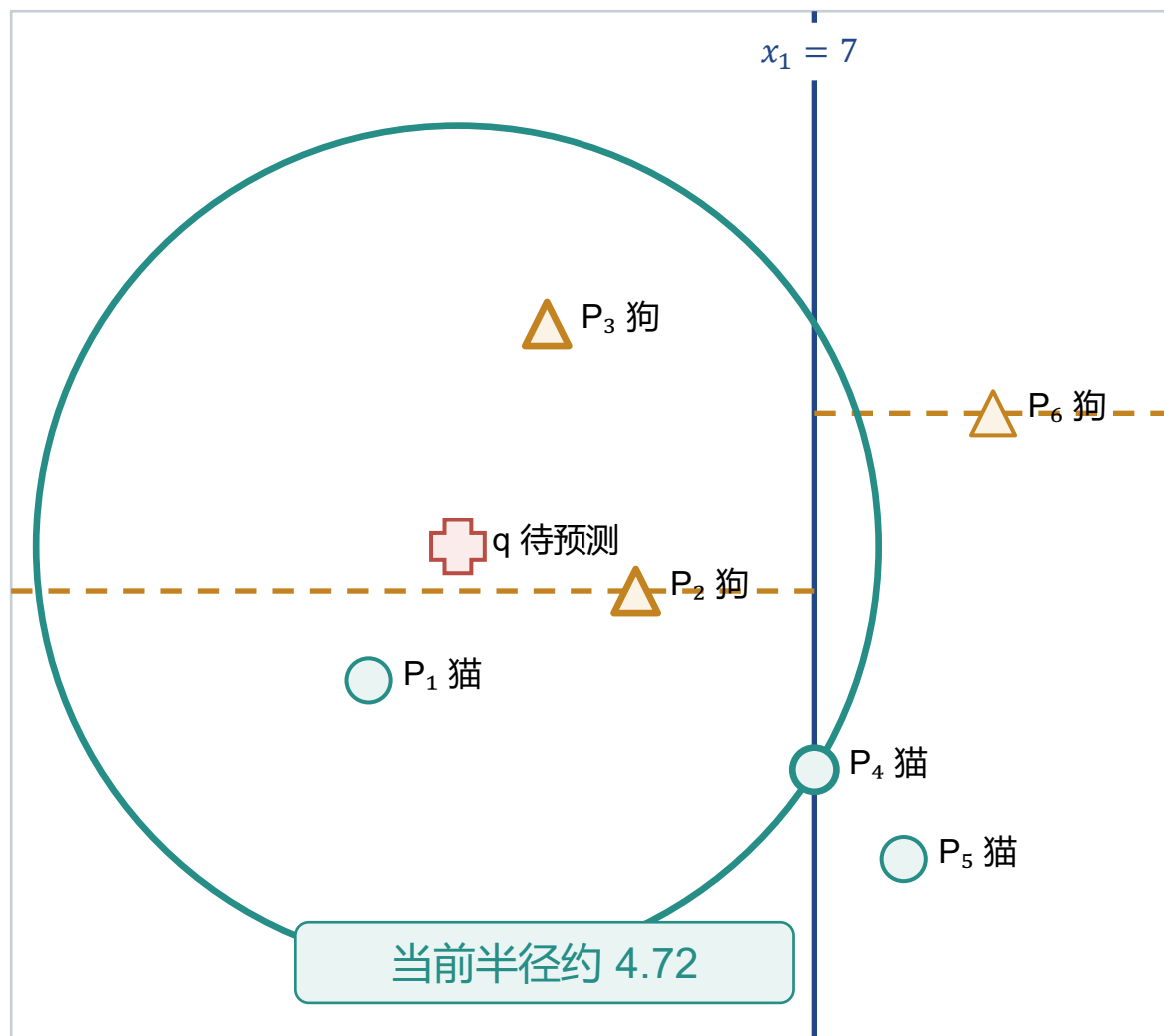
- 左右子集改用特征 2, 分别得到 P_2 与 P_6
- 下一层再改用特征 1, 放置 P_1 、 P_3 、 P_5

空间与树的对应

- 每条切分线只划分当前节点负责的空间区域
- 空间区域与树中的左右子树一一对应

近侧下降与当前候选

已缩放的教学特征 2



近侧下降

- 按当前切分维度选择查询点所在的一侧
- 下降路径为 P_4 、 P_2 、 P_3

当前候选

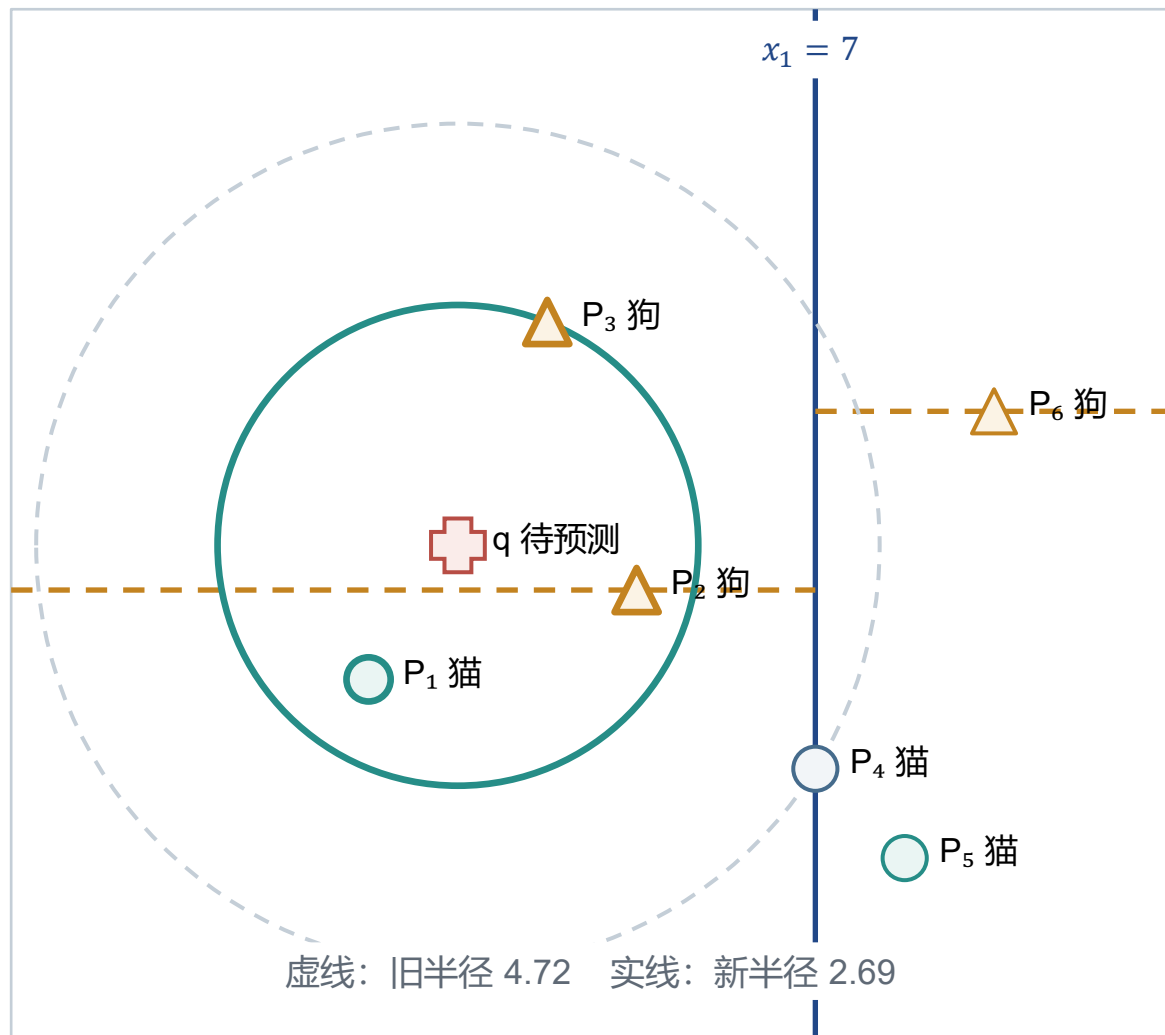
- 三个候选为 P_4 、 P_2 、 P_3
- 最远候选距离约为 4.72, 形成第一版搜索圆

完整性边界

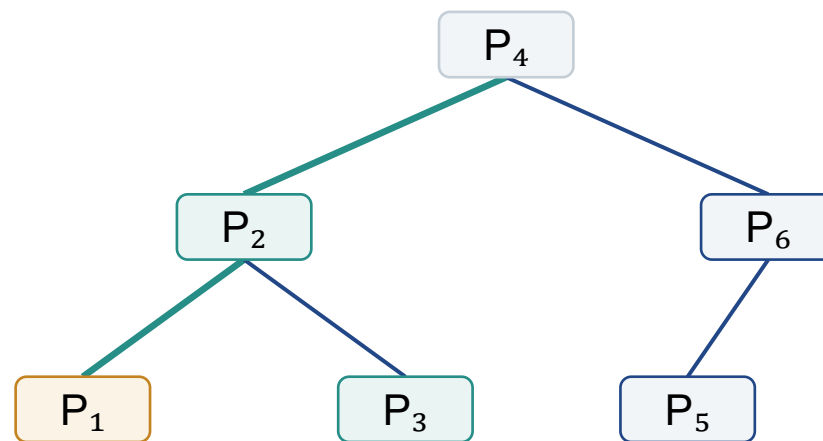
- 到达空子树不表示查询已经结束, 仍需回溯检查

回溯：搜索圆跨过切分线

已缩放的教学特征 2



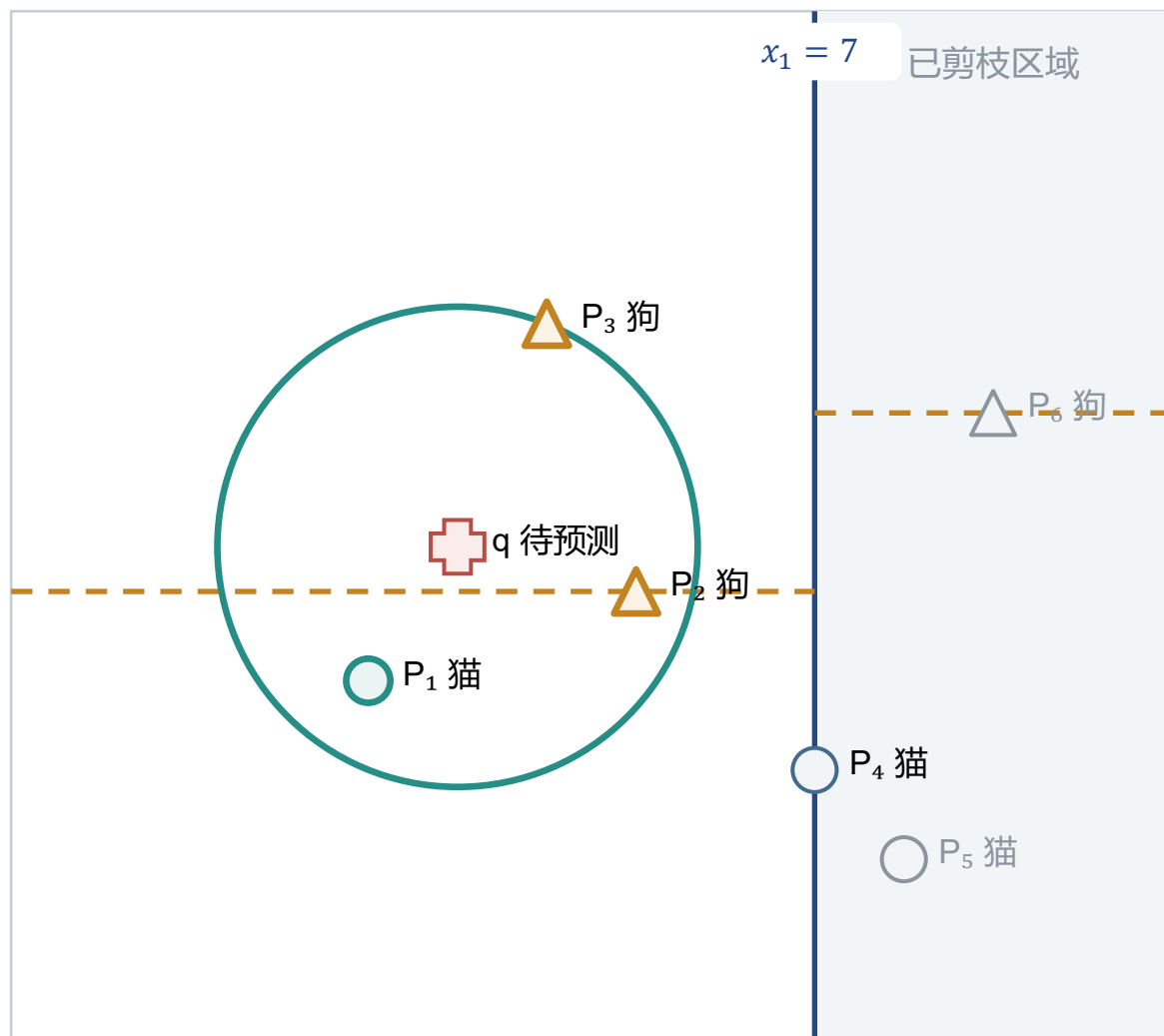
已缩放的教学特征 1



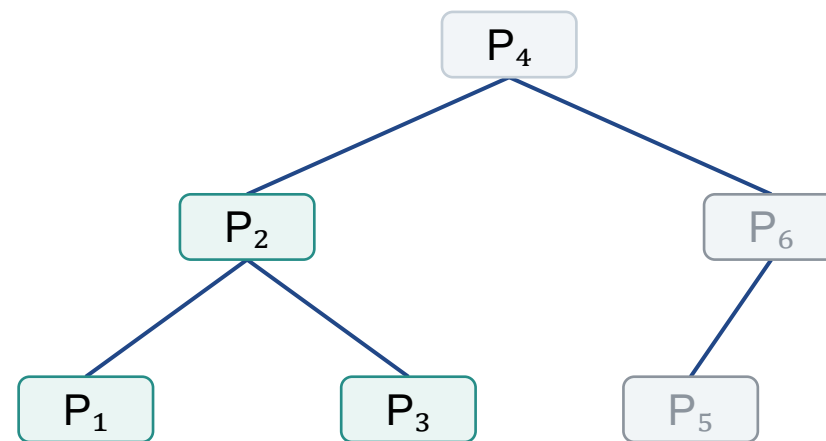
- **边界检查** 候选替换： P_4 被 P_1 替换
 - 回到 P_2 的水平切分边界
 - 边界距离 $0.5 < 4.72$ ，另一侧仍可能有更近点
 - 边界距离是到达另一侧任意点必须先跨过的距离下界
- **候选更新**
 - 访问 P_1 ，用 P_1 替换 P_4
 - 当前候选变为 P_1 、 P_2 、 P_3
 - 搜索半径缩小为约 2.69

剪枝：搜索圆未触及切分线

已缩放的教学特征 2



已缩放的教学特征 1



边界检查

- 回到根节点 P_4 的竖直切分边界
- 边界距离 $4 > 2.69$ ，右侧不可能出现更近点
- 距离下界已超过当前最远候选，因此剪枝不会改变精确邻近结果

剪枝与结果

- 跳过包含 P_5 、 P_6 的右子树
- 最终邻居为 P_1 、 P_2 、 P_3
- 多数投票结果仍为狗

3 小结: k 近邻与 kd 树的作用边界

■ k 近邻预测

- 特征和尺度决定距离的含义
- k 影响局部范围与投票结果
- 预测时需要保留或访问训练样本

■ kd 树检索

- 只改变寻找邻居的过程
- 是否能减少搜索取决于维度和数据分布
- 完整回溯是精确查询的一部分

3.1 共同结论与学习边界

■ 精确检索的结果一致

- 精确 kd 树查询与直接搜索返回相同的邻居

■ 学习边界

- 掌握：k 近邻预测规则，kd 树建树与检索思路
- 自行展开：实现、复杂度、回归和其他近邻索引