03 k近邻算法

从邻居投票到 kd 树检索

本讲学习目标

- 解释 k 近邻怎样依据距离、邻居数量和多数投票形成分类结果
- 区分 k 近邻的预测规则与 kd 树的检索职责
- 借助示意图说明 kd 树的空间组织，跟随例子说明下降、候选更新、回溯和剪枝
- 认识特征尺度、邻居数量与数据条件对预测和检索的影响

1 k 近邻预测

距离、邻居数量与投票规则，以及完整分类结果

k 近邻分类规则

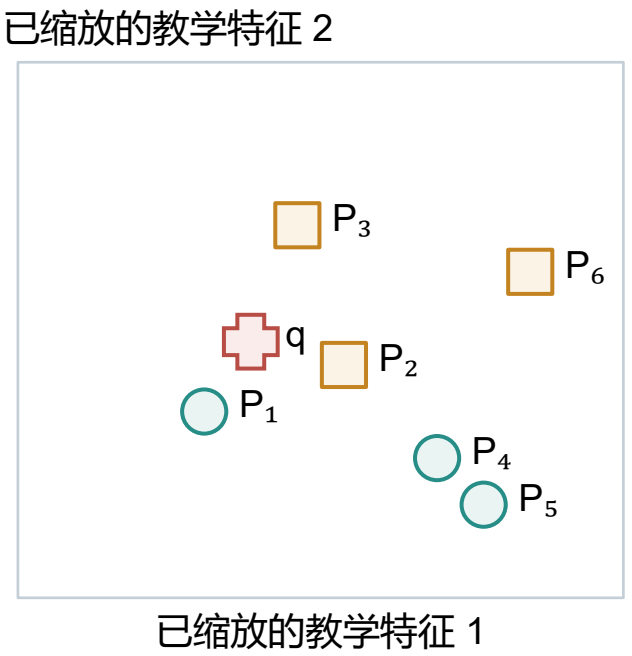
- 所用特征与距离应反映任务相关的相似性，使邻近样本的标签对查询点具有参考意义
  - 特征尺度影响距离及邻居排序，需采用适合任务的尺度
- 保留带标签训练样本，查询时选择邻居并用其标签预测类别；本讲使用欧氏距离和等权多数投票

1. **衡量距离**：计算查询点到各训练样本的距离
   - 欧氏距离表示特征空间中两点之间的直线距离
2. **选择邻居**：按距离从小到大排序，取最近的 k 个训练样本
   - k 是参与投票的邻居数量
3. **多数投票**：每个邻居投一票，以票数最多的类别作为预测
   - 平票需预先约定处理方式

训练样本与查询点

六个训练点已有猫或狗的标签，坐标表示两个已缩放的抽象教学特征

点	坐标	类别
P_1	$(2, 3)$	猫
P_2	$(5, 4)$	狗
P_3	$(4, 7)$	狗
P_4	$(7, 2)$	猫
P_5	$(8, 1)$	猫
P_6	$(9, 6)$	狗



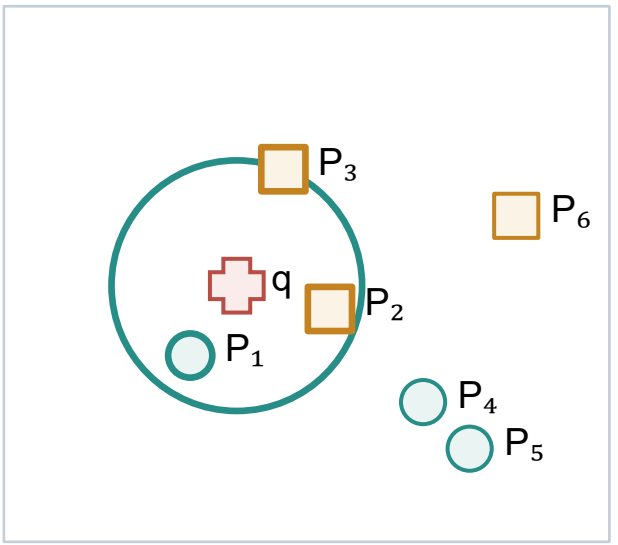
查询点 $q = (3,4.5)$ 的类别未知，需要利用训练样本预测

本例使用欧氏距离，选择 $k = 3$ 个邻居，按标签等权多数投票；没有等距或投票平局，点的坐标与类别均为教学构造

近邻选择与多数投票

按查询点 q 到六个训练点的欧氏距离排序，取前三个作为 $k = 3$ 的邻居

顺序	训练点	到 q 的距离（约）	类别	参与投票
1	P_1	1.80	猫	是
2	P_2	2.06	狗	是
3	P_3	2.69	狗	是
4	P_4	4.72	猫	否
5	P_5	6.10	猫	否
6	P_6	6.18	狗	否

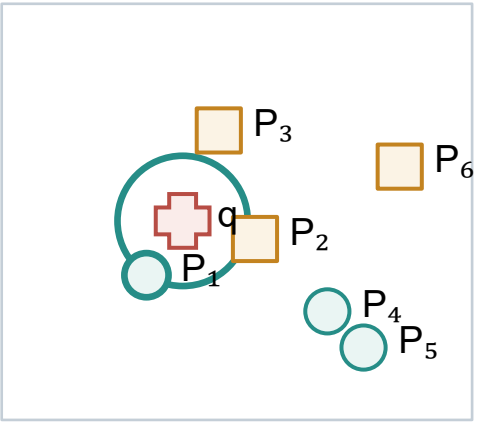


- 三个最近邻按标签投票，预测类别为**狗**
  - 邻居为 P_1, P_2, P_3，标签为猫、狗、狗
  - 狗得到两票，猫得到一票
  - 这是依据邻居标签得到的预测，查询点的真实类别仍未知

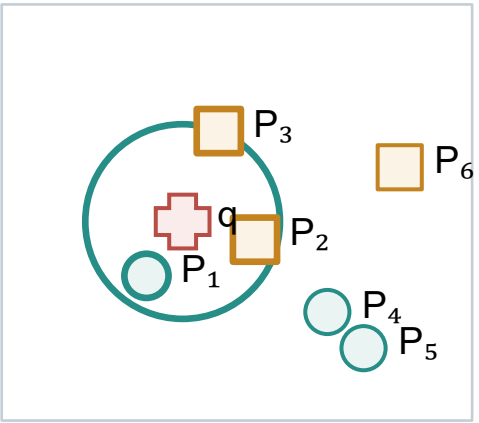
邻居数量与预测结果

训练样本、查询点、特征尺度、欧氏距离和等权投票规则保持不变，只改变邻居数量 k

邻居数量	参与投票的邻居	类别票数	预测类别
$k = 1$	P_1	猫 1 票，狗 0 票	猫
$k = 3$	P_1, P_2, P_3	猫 1 票，狗 2 票	狗



k = 1



k = 3

- 距离排序保持不变，k 改变参与投票的邻居集合
  - 本例增加的两个邻居均为狗，投票结果随之改变
- **邻居数量的取值**
  - 较小的 k 更依赖局部个别样本，较大的 k 汇总更广邻域
  - 适合的取值依赖当前数据与任务

2 kd 树空间组织与查询

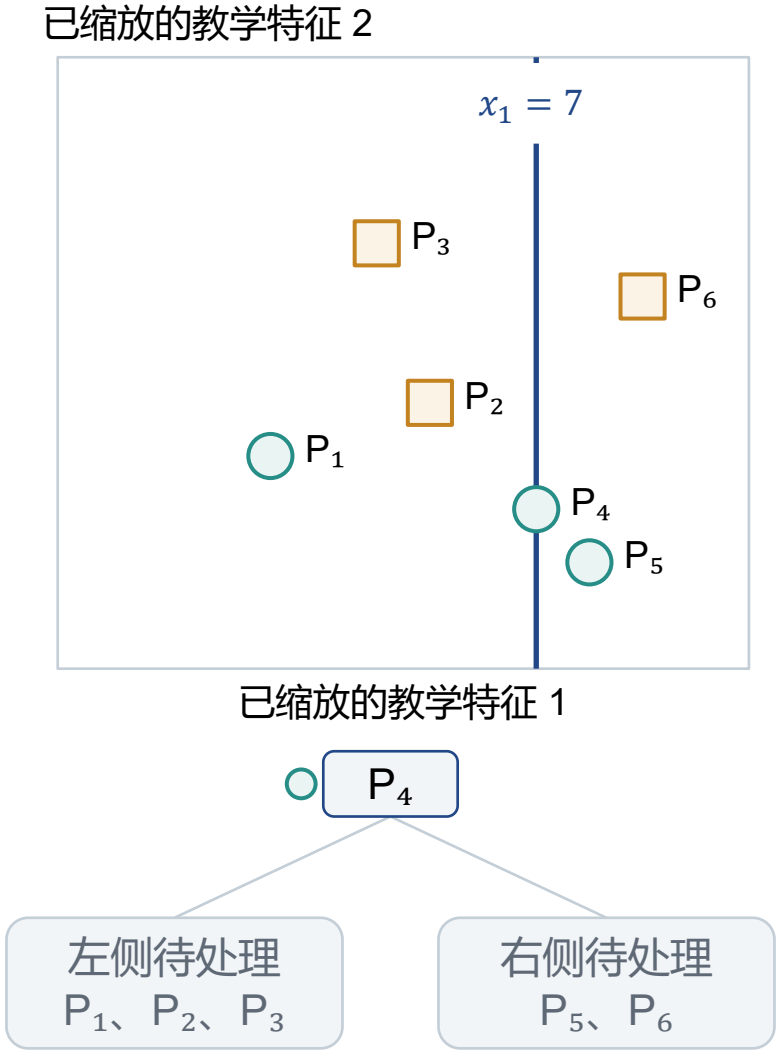
检索与预测的接口、空间组织与精确近邻查询

近邻检索与分类预测

- **近邻检索**：已知最近的 k 个邻居即可投票；训练集增大后，需要高效找到这些邻居
  - **直接搜索**：逐点比较全部训练样本，样本增多时，每次查询的距离计算量随之增加
  - **kd 树检索**：按坐标维度递归切分空间，将训练点组织为二叉树
    - 区域距离下界表示其中任意点到查询点的距离至少有多大；下界足以排除更近点时，可跳过整个区域
- **分类预测**：邻居确定后，仍按其标签等权多数投票
  - 相同表示、距离和歧义约定下，精确检索返回相同邻居；检索方式改变，邻居目标与投票规则保持不变



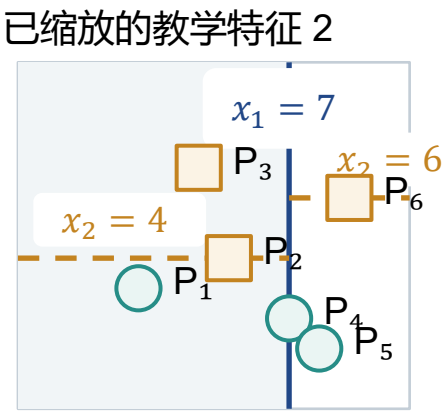
kd 树的递归空间划分

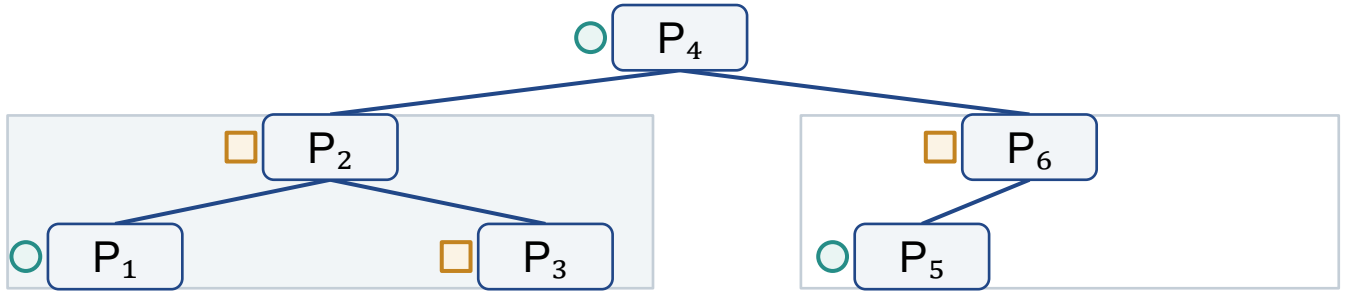


- 按空间位置组织训练点，为近邻查询建立区域与子树
  1. 选择坐标维度，以一个训练点的坐标切分当前区域
  2. 该点保存为节点，其余点按坐标大小分到两侧子树
  3. 在子区域递归切分，形成树结构
- **本例构造约定**
  - **切分维度**：交替使用 x_1, x_2
  - **切分点**：按当前维度排序后取上中位数；偶数个点取两个居中点中坐标较大者
    - 本例根为 P_4，切分线为 $x_1 = 7$

空间划分与树结构的对应

- 每个节点保存一个训练点，切分边界只作用于该节点负责的区域；左右子树保存两侧的点





左区域对应左子树；右区域对应右子树

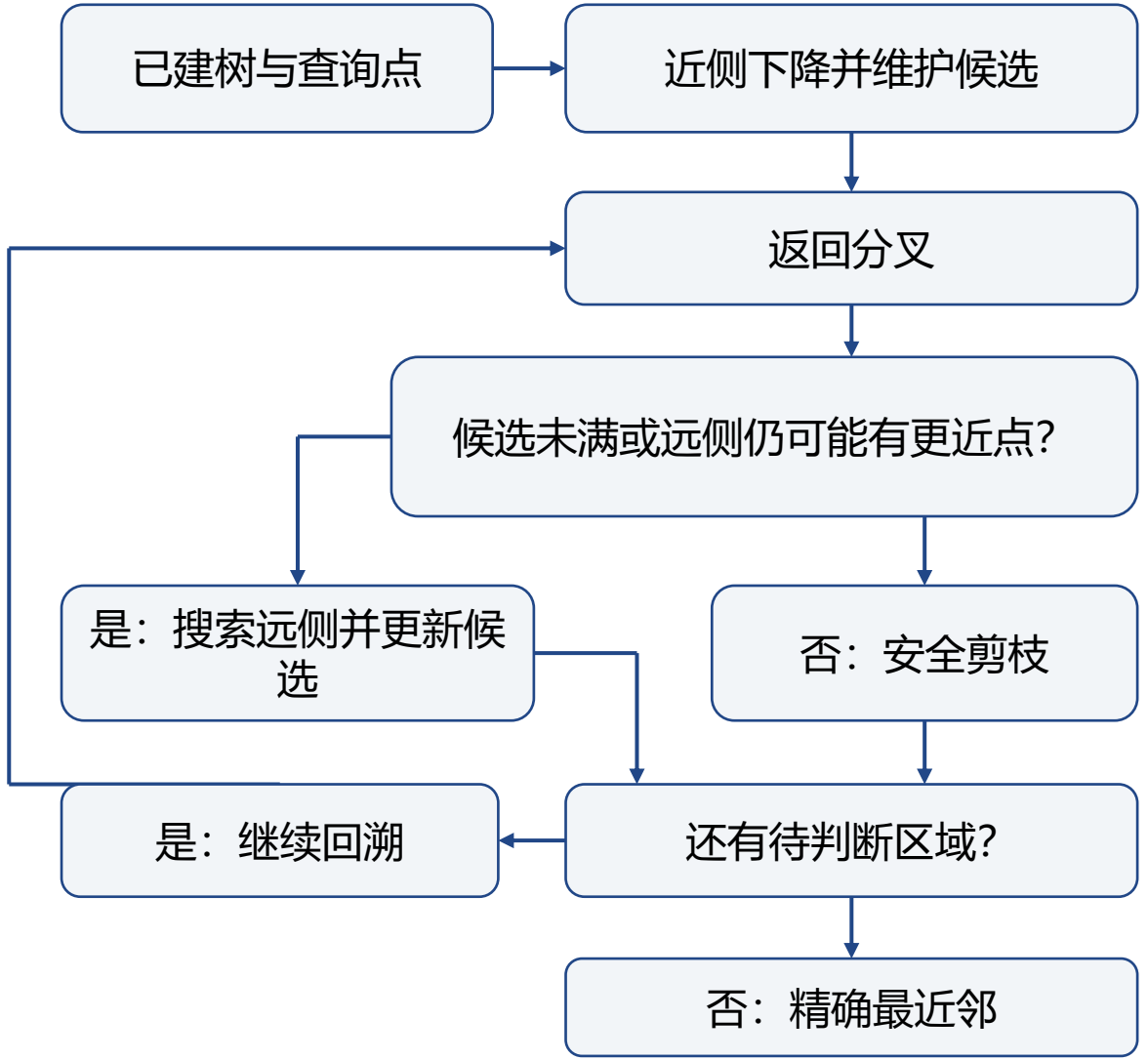
- **根节点** P_4：以 $x_1 = 7$ 分开左右区域，对应两棵子树
- **子节点** P_2, P_6：分别在左、右区域内按 x_2 切分，形成各自的子树
- **叶节点** P_1, P_3, P_5：没有待分配的剩余点，不再产生子树

- 这棵树保留全部六个训练点；查询时根据节点的切分维度与边界，选择下降方向并检查其他区域

候选建立与区域判断

子树对应区域；当前切分线两侧，查询点所在一侧称**近侧**，另一侧称**远侧**（本例不在切分线上）

1. **近侧下降**：优先检查近侧，访问节点时保留已检查点中最近的至多 k 个作为候选
2. **回溯判断**：返回分叉处；候选未满或远侧仍可能有更近点时继续搜索，否则安全剪枝
3. **查询完成**：必要区域均已搜索或安全排除，候选才成为精确最近邻



候选维护与剪枝条件

- **候选维护**：记录已找到哪些点及其最远距离
  - **候选集合**：已访问点中保留的至多 k 个最近点
  - **候选更新**：访问节点时计算距离；不足 k 个则加入，已满则仅用更近的新点替换最远候选
  - **搜索半径** r：候选已满时，最远候选到查询点的距离；未满时不能据有限半径排除区域
- **远侧检查**：判断另一侧是否仍可能改善候选
  - **距离下界** δ：查询点到切分线的垂直距离；到达远侧任意点都必须跨过切分线，因此其欧氏距离至少为 δ
  - **继续搜索**：候选未满，或已满且 $\delta \leq r$ 时，搜索另一侧
  - **安全剪枝**：候选已满且 $\delta > r$ 时，可跳过另一侧子树
- 先下降到近侧，再沿访问路径返回并检查远侧；返回检查的过程称为**回溯**
  - 所有必要分支处理完毕，候选成为最终最近邻；目标仍是距查询点最近的 k 个训练点

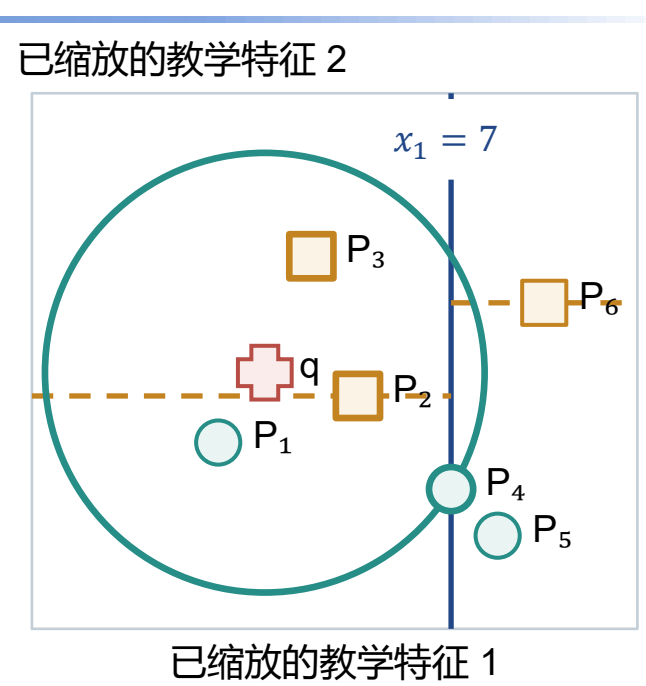
近侧下降与候选建立

在已建树上按欧氏距离查询 $q = (3,4.5)$ 的 $k = 3$ 个近邻，初始候选为空

每次访问节点，先更新候选，再选择近侧

访问节点	到 q 的距离（约）	更新后的候选集合	近侧选择
P_4	4.72	$\{P_4\}$	x_1：$3 < 7$，进入左子树
P_2	2.06	$\{P_4, P_2\}$	x_2：$4.5 > 4$，进入右子树
P_3	2.69	$\{P_4, P_2, P_3\}$	x_1：$3 < 4$，进入空左子树

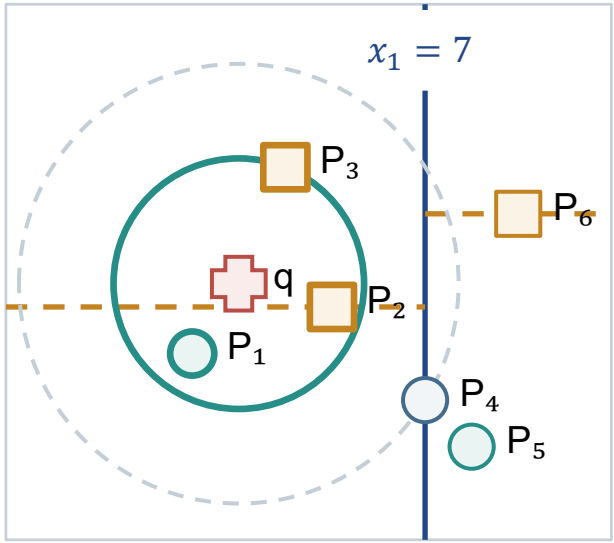
- 候选达到三个，最远点为 P_4，$r \approx 4.72$
  - 搜索圆以 q 为圆心、r 为半径，表示候选的最远距离范围
- P_3 两侧均为空，返回 P_2
  - 祖先的远侧仍待检查，候选尚非最终最近邻



回溯检查与候选更新

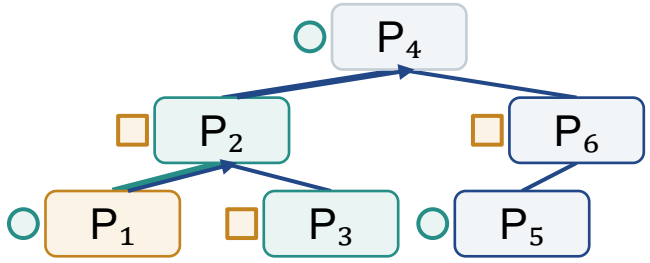
- 返回 P_2，判断是否搜索左子树
  - 候选为 $\{P_4, P_2, P_3\}$，$r \approx 4.72$
  - 右侧已处理，左侧待检查
- 切分线为 $x_2 = 4$，$\delta = |4.5 - 4| = 0.5$
  - $0.5 < 4.72$，远侧仍可能有更近点，需搜索左子树
  - 旧搜索圆跨过切分线；判断依据是下界与半径的比较
- 访问 P_1，更新候选
  - 其距离约 1.80，小于最远候选 P_4 的 4.72
  - P_1 替换 P_4，候选变为 $\{P_1, P_2, P_3\}$
  - 最远点改为 P_3，$r \approx 2.69$
- 沿已访问路径继续回溯
  - P_1 两侧为空；返回 P_2，其两侧已处理，再返回根 P_4，检查右子树

已缩放的教学特征 2



已缩放的教学特征 1

虚线：旧半径 4.72
实线：新半径 2.69



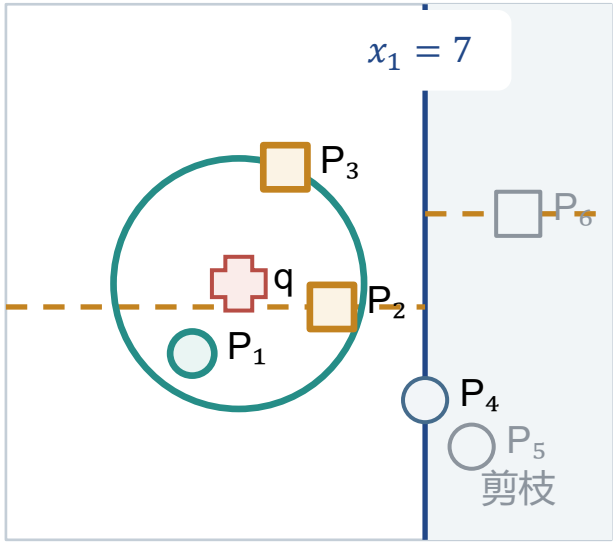
回溯：$P_2 \to P_1 \to P_2 \to P_4$

安全剪枝与查询结果

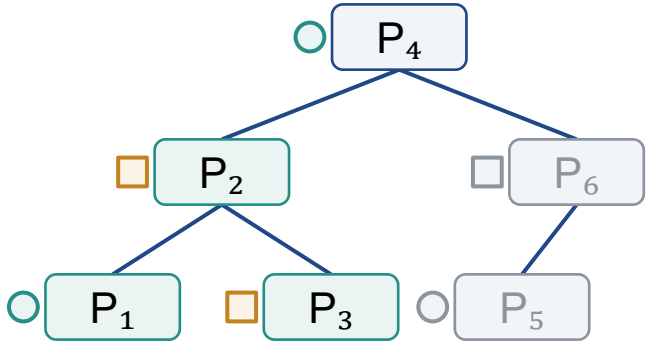
- 返回根 P_4，判断是否搜索右子树
  - 候选为 $\{P_1, P_2, P_3\}$，$r \approx 2.69$
  - 左子树已处理，右子树待检查
- 切分线为 $x_1 = 7$，$\delta = |3 - 7| = 4$
  - 右侧任意点到 q 的距离至少为 4，超过当前半径 2.69
  - 可跳过含 P_6, P_5 的右子树；搜索圆未触及切分线
  - 本次剪枝跳过两个点的距离检查，保持精确结果
- 根的两侧均已搜索或安全排除，查询结束
  - 最近邻为 P_1, P_2, P_3，与前面直接搜索的结果相同
  - 标签为猫、狗、狗，狗两票、猫一票，预测为**狗**；真实类别仍未知

其他数据上的检索效率仍取决于数据条件

已缩放的教学特征 2



已缩放的教学特征 1



3 方法条件与总结

预测条件、检索条件与已建立认识

k 近邻与 kd 树的使用条件

k 近邻预测

- **特征与尺度**：邻近标签是否有参考意义，取决于特征和任务；尺度差异可能使某些特征主导欧氏距离，应采用适合任务的尺度。本例特征已缩放
- **邻居数量**：k 改变投票范围；较小的 k 更依赖局部个别样本，较大的 k 汇总更广邻域，取值依赖数据与任务
- **样本访问**：预测时需要保留或访问带标签训练样本

kd 树检索

- 按前述规则维护候选、执行必要回溯与安全剪枝，得到与直接搜索**相同的最近邻**；投票规则保持不变
- **检索效率**：可排除的区域取决于维度与数据分布；高维或切分难以排除候选时，仍可能检查大量节点，甚至慢于直接搜索

本讲总结

- **k 近邻预测**：以距离选择最近的 k 个邻居，由标签投票得到预测类别
- **kd 树检索**：递归划分空间，使子树对应区域；下降建立候选，回溯时比较距离下界与搜索半径，完成必要搜索或安全剪枝，得到精确邻居
- kd 树改变寻找邻居的过程；相同邻居按相同规则投票，**预测结果不变**
- 特征、尺度和 k 影响预测依据；维度与数据分布影响检索效率