统计学习第三章

2022年8月10日 星期三 12:50

k近邻法

输入: 训练数据集: $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

输出: $y=\arg\max\limits_{c_j}\sum\limits_{x_i\in N_k(x)} I(y_i=c_j)$, ……（$I$为指示函数, $y_i=c_j$时I为1, 否则为0）

○设特征空间 $\mathcal{X}$ 为n维实数向量空间 R^n, $x_i, x_j\in\mathcal{X}$

$x_i=(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T$

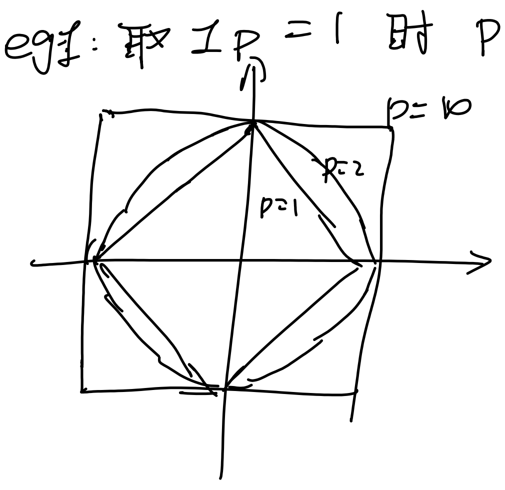
$x_j=(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)})^T$

其中 x_i, x_j 的 L_p 距离定义为:

$$L_p(x_i,x_j)=\left(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^p\right)^{\frac{1}{p}} \quad p\geq 1$$

- $p=2$: 欧氏距离: $L_2(x_i,x_j)=\left(\sum\limits_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^2\right)^{\frac{1}{2}}$
- $p=1$: 曼哈顿距离: $L_1(x_i,x_j)=\sum\limits_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|$
- $p=\infty$: 各坐标距离的max: $L_\infty(x_i,x_j)=\max\limits_{l}|x_i^{(l)}-x_j^{(l)}|$

eg1: 取 $L_p=1$ 时 $p=1$, $p=2$, $p=\infty$ 时的图形:



「解释」:

- p=1 指特征向量中每一个值加起来为1
- p=2 指图形上一点到原点距离为1
- p=∞ 指正方形上一最大值（取了绝对值之后）

eg2: $x_1=(1,1)^T$, $x_2=(5,1)^T$, $x_3=(4,4)^T$ 在p取不同值时 L_p 距离下 x_1 的最近邻点.

1° $p=1$ 时, $L(x_1,x_2)=4$ $L(x_1,x_3)=6$

2° $p=2$ 时, $L(x_1,x_2)=4$, $L(x_1,x_3)=3\sqrt{2}$

3° $p=\infty$ 时, $L(x_1,x_2)=4$, $L(x_1,x_3)=3$.

$p=1,2$时, 最邻近点为x_2, $L=\infty$ 时, 最邻近点为x_3

注: k值较小, 相当于用较小邻域

k值较大, 相当于用较大邻域.

分类决策规则

$f: R^n\to\{c_1,c_2,\cdots,c_K\}$

误分类概率: $P(Y\neq f(X))=1-P(Y=f(X))$

$$\frac{1}{k}\sum_{x_i\in N_k(x)} I(y_i\neq c_j)=1-\frac{1}{k}\sum_{x_i\in N_k(x)} I(y_i=c_j)$$

（I为指示函数, 是为1, 不是为0）

k近邻法的实现: kd树.

○构造kd树: 构造根结点, 将实例保存在相应结点,

算法: 输入: $T=\{x_1,x_2,\cdots,x_n\}$, 其中 $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(k)})^T$

输出: kd 树.

1) 开始: 构造根结点, 对应于包含T的k维空间的超矩形区域
选择 $x^{(1)}$ 为坐标轴, 以T中所有实例 $x^{(1)}$ 坐标的中位数为切分点

2) 重复: 对深度为j的结点, 选择 $x^{(l)}$ 为切分的坐标轴,
$l=j(\bmod k)+1$

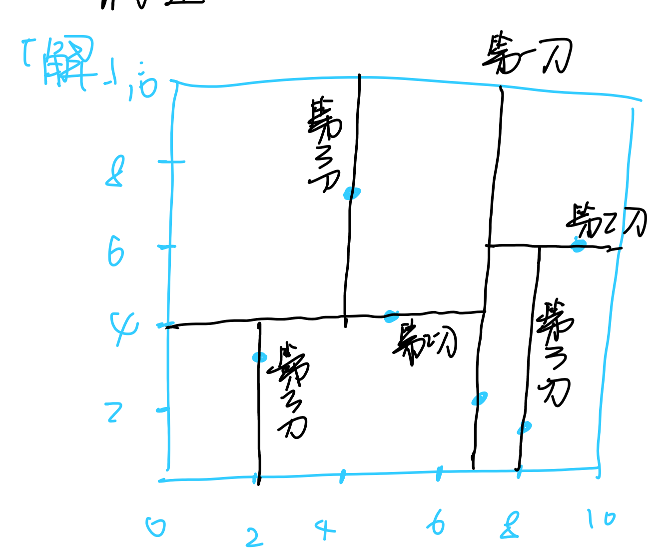
3) 直到两个子区域没有实例存在时停止, 从而形成kd树的区域划分

eg: 给定二维数据集:

$T=\{(2,3)^T, (5,4)^T, (9,6)^T, (4,7)^T, (8,1)^T, (7,2)^T\}$

构造一个平衡kd树.

[解]:

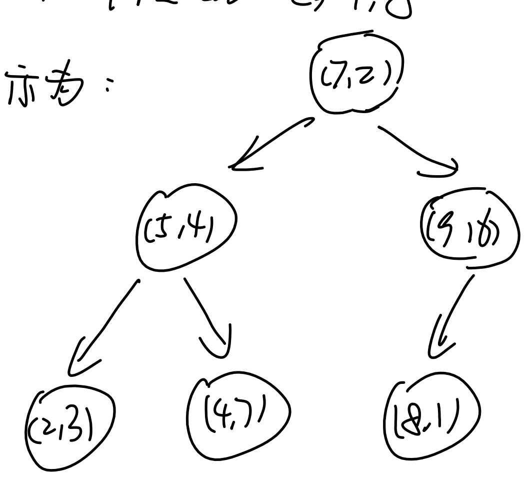


先切 $x^{(1)}$ 中位数为7. （中位数取大不取小）

再切 $x^{(2)}$ 中位数 4, 6

再切 $x^{(1)}$ 中位数 2, 4, 8

又可表示为:



搜索kd树 → 如何与k邻近建立联系?

先找叶结点, 再找父结点, 往上找……

eg. P57

点S与D同为叶结点, 画圆后判定E点距离更近.