# 决策树

以表中的西瓜数据集2.0为例，该数据集包含17个训练样例，用以学习一棵能预测没剖开的是不是好瓜的决策树.

| 编号 | 色泽 | 根蒂 | 敲声 | 纹理 | 脐部 | 触感 | 好瓜 |
|---|---|---|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 是 |
| 2 | 乌黑 | 蜷缩 | 沉闷 | 清晰 | 凹陷 | 硬滑 | 是 |
| 3 | 乌黑 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 是 |
| 4 | 青绿 | 蜷缩 | 沉闷 | 清晰 | 凹陷 | 硬滑 | 是 |
| 5 | 浅白 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 是 |
| 6 | 青绿 | 稍蜷 | 浊响 | 清晰 | 稍凹 | 软粘 | 是 |
| 7 | 乌黑 | 稍蜷 | 浊响 | 稍糊 | 稍凹 | 软粘 | 是 |
| 8 | 乌黑 | 稍蜷 | 浊响 | 清晰 | 稍凹 | 硬滑 | 是 |
| 9 | 乌黑 | 稍蜷 | 沉闷 | 稍糊 | 稍凹 | 硬滑 | 否 |
| 10 | 青绿 | 硬挺 | 清脆 | 清晰 | 平坦 | 软粘 | 否 |
| 11 | 浅白 | 硬挺 | 清脆 | 模糊 | 平坦 | 硬滑 | 否 |
| 12 | 浅白 | 蜷缩 | 浊响 | 模糊 | 平坦 | 软粘 | 否 |
| 13 | 青绿 | 稍蜷 | 浊响 | 稍糊 | 凹陷 | 硬滑 | 否 |
| 14 | 浅白 | 稍蜷 | 沉闷 | 稍糊 | 凹陷 | 硬滑 | 否 |
| 15 | 乌黑 | 稍蜷 | 浊响 | 清晰 | 稍凹 | 软粘 | 否 |
| 16 | 浅白 | 蜷缩 | 浊响 | 模糊 | 平坦 | 硬滑 | 否 |
| 17 | 青绿 | 蜷缩 | 沉闷 | 稍糊 | 稍凹 | 硬滑 | 否 |

可计算出根结点的信息熵为：

$$\mathrm{Ent}(D)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{8}{17}\log_2\frac{8}{17}+\frac{9}{17}\log_2\frac{9}{17}\right)=0.998$$

◇◇◇

我们要计算出当前属性集合{色泽，根蒂、敲声，纹理，脐部，触感}中每个属性的信息增益.以属性"色泽"为例，它有3个可能的取值：{青绿，乌黑，浅白}. 若使用该属性对D进行划分，则可得到3个子集：分别记为：D1（色泽=青绿），D2（色泽=乌黑），D3（色泽=浅白）.

子集D1包含编号为{1，4，6，10，13，17}的6个样例，D2包含编号为{2，3，7，8，9，15}的6个样例，D3包含编号为{5，11，12，14，16}的5个样例，根据公式：

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k$$

可计算出用"色泽"划分之后所获得的3 个分支结点的信息熵为：

$$\mathrm{Ent}(D^1)=-\left(\frac{3}{6}\log_2\frac{3}{6}+\frac{3}{6}\log_2\frac{3}{6}\right)=1.000$$
$$\mathrm{Ent}(D^2)=-\left(\frac{4}{6}\log_2\frac{4}{6}+\frac{2}{6}\log_2\frac{2}{6}\right)=0.918$$
$$\mathrm{Ent}(D^3)=-\left(\frac{1}{5}\log_2\frac{1}{5}+\frac{4}{5}\log_2\frac{4}{5}\right)=0.722$$

◇◇◇

于是，可计算出属性"色泽"的信息增益为

$$\begin{aligned}\mathrm{Gain}(D,\text{色泽})&=\mathrm{Ent}(D)-\sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\\&=0.998-\left(\frac{6}{17}\times 1.000+\frac{6}{17}\times 0.918+\frac{5}{17}\times 0.722\right)\\&=0.109\ .\end{aligned}$$

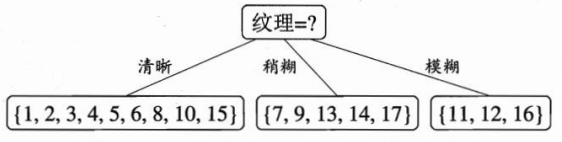
类似的，我们可计算出其他属性的信息增益：

$$\mathrm{Gain}(D,\text{根蒂})=0.143;\quad \mathrm{Gain}(D,\text{敲声})=0.141;$$
$$\mathrm{Gain}(D,\text{纹理})=0.381;\quad \mathrm{Gain}(D,\text{脐部})=0.289;$$
$$\mathrm{Gain}(D,\text{触感})=0.006.$$

显然，属性"纹理"的信息增益最大，于是它被选为划分属性

◇◇◇

如图给出了基于"纹理"对根结点进行划分的结果，各分支结点所包含的样例子集显示在结点中.



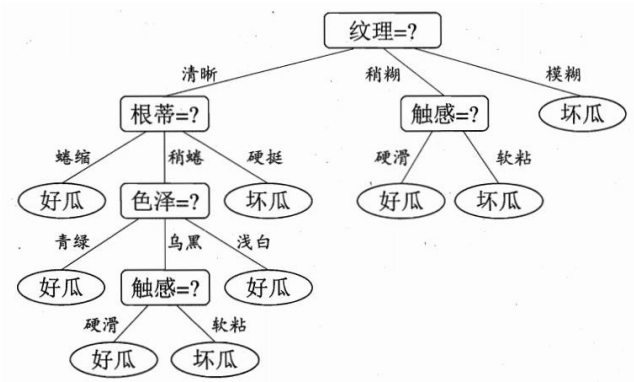

然后，决策树学习算法将对每个分支结点做进一步划分.以图中第一个分支结点"纹理=清晰"为例，该结点包含的样例集合 D1中有编号为{1，2，3，4，5，6，8，10，15}的9个样例，可用属性集合为{色泽，根蒂，敲声，脐部，触感}（"纹理"不再作为候选）. 基于DI计算出各属性的信息增益：

$$\mathrm{Gain}(D^1,\text{色泽})=0.043;\quad \mathrm{Gain}(D^1,\text{根蒂})=0.458;$$
$$\mathrm{Gain}(D^1,\text{敲声})=0.331;\quad \mathrm{Gain}(D^1,\text{脐部})=0.458;$$
$$\mathrm{Gain}(D^1,\text{触感})=0.458.$$

"根蒂"、"脐部"、"触感"3个属性均取得了最大的信息增益，可任选其中之一作为划分属性.

◇◇◇

对每个分支结点进行上述操作，最终得到的决策树如图所示：




## 朴素贝叶斯

下面我们用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例"测1"进行分类

首先估计类先验概率 P(c)，显然有:

$$P(\text{好瓜}=\text{是}) = \frac{8}{17} \approx 0.471\ ,$$
$$P(\text{好瓜}=\text{否}) = \frac{9}{17} \approx 0.529\ .$$

| 编号 | 色泽 | 根蒂 | 敲声 | 纹理 | 脐部 | 触感 | 密度 | 含糖率 | 好瓜 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 0.697 | 0.46 | 是 |
| 2 | 乌黑 | 蜷缩 | 沉闷 | 清晰 | 凹陷 | 硬滑 | 0.774 | 0.376 | 是 |
| 3 | 乌黑 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 0.634 | 0.264 | 是 |
| 4 | 青绿 | 蜷缩 | 沉闷 | 清晰 | 凹陷 | 硬滑 | 0.608 | 0.318 | 是 |
| 5 | 浅白 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 0.556 | 0.215 | 是 |
| 6 | 青绿 | 稍蜷 | 浊响 | 清晰 | 稍凹 | 软粘 | 0.403 | 0.237 | 是 |
| 7 | 乌黑 | 稍蜷 | 浊响 | 稍糊 | 稍凹 | 软粘 | 0.481 | 0.149 | 是 |
| 8 | 乌黑 | 稍蜷 | 浊响 | 清晰 | 稍凹 | 硬滑 | 0.437 | 0.211 | 是 |
| 9 | 乌黑 | 稍蜷 | 沉闷 | 稍糊 | 稍凹 | 硬滑 | 0.666 | 0.091 | 否 |
| 10 | 青绿 | 硬挺 | 清脆 | 清晰 | 平坦 | 软粘 | 0.243 | 0.267 | 否 |
| 11 | 浅白 | 硬挺 | 清脆 | 模糊 | 平坦 | 硬滑 | 0.245 | 0.057 | 否 |
| 12 | 浅白 | 蜷缩 | 浊响 | 模糊 | 平坦 | 软粘 | 0.343 | 0.099 | 否 |
| 13 | 青绿 | 稍蜷 | 浊响 | 稍糊 | 凹陷 | 硬滑 | 0.639 | 0.161 | 否 |
| 14 | 浅白 | 稍蜷 | 沉闷 | 稍糊 | 凹陷 | 硬滑 | 0.657 | 0.198 | 否 |
| 15 | 乌黑 | 稍蜷 | 浊响 | 清晰 | 稍凹 | 软粘 | 0.36 | 0.37 | 否 |
| 16 | 浅白 | 蜷缩 | 浊响 | 模糊 | 平坦 | 硬滑 | 0.593 | 0.042 | 否 |
| 17 | 青绿 | 蜷缩 | 沉闷 | 稍糊 | 稍凹 | 硬滑 | 0.719 | 0.103 | 否 |

| 编号 | 色泽 | 根蒂 | 敲声 | 纹理 | 脐部 | 触感 | 密度 | 含糖率 | 好瓜 |
|---|---|---|---|---|---|---|---|---|---|
| 测 1 | 青绿 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 0.697 | 0.460 | ? |

然后，为每个属性估计条件概率 $P(x_i \mid c)$

$$P_{\text{青绿}|\text{是}} = P(\text{色泽}=\text{青绿} \mid \text{好瓜}=\text{是}) = \frac{3}{8} = 0.375\ ,\ P_{\text{清晰}|\text{是}} = P(\text{纹理}=\text{清晰} \mid \text{好瓜}=\text{是}) = \frac{7}{8} = 0.875\ ,$$
$$P_{\text{青绿}|\text{否}} = P(\text{色泽}=\text{青绿} \mid \text{好瓜}=\text{否}) = \frac{3}{9} \approx 0.333\ ,\ P_{\text{清晰}|\text{否}} = P(\text{纹理}=\text{清晰} \mid \text{好瓜}=\text{否}) = \frac{2}{9} \approx 0.222\ ,$$
$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂}=\text{蜷缩} \mid \text{好瓜}=\text{是}) = \frac{5}{8} = 0.375\ ,\ P_{\text{凹陷}|\text{是}} = P(\text{脐部}=\text{凹陷} \mid \text{好瓜}=\text{是}) = \frac{6}{8} = 0.750\ ,$$
$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂}=\text{蜷缩} \mid \text{好瓜}=\text{否}) = \frac{3}{9} \approx 0.333\ ,\ P_{\text{凹陷}|\text{否}} = P(\text{脐部}=\text{凹陷} \mid \text{好瓜}=\text{否}) = \frac{2}{9} \approx 0.222\ ,$$
$$P_{\text{浊响}|\text{是}} = P(\text{敲声}=\text{浊响} \mid \text{好瓜}=\text{是}) = \frac{6}{8} = 0.750\ ,\ P_{\text{硬滑}|\text{是}} = P(\text{触感}=\text{硬滑} \mid \text{好瓜}=\text{是}) = \frac{6}{8} = 0.750\ ,$$
$$P_{\text{浊响}|\text{否}} = P(\text{敲声}=\text{浊响} \mid \text{好瓜}=\text{否}) = \frac{4}{9} \approx 0.444\ ,\ P_{\text{硬滑}|\text{否}} = P(\text{触感}=\text{硬滑} \mid \text{好瓜}=\text{否}) = \frac{6}{9} \approx 0.667\ ,$$

对连续属性可考虑概率密度函数，假定 $P(x_i \mid c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$ 其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第c类样本在第i个属性上取值的均值和方差，则有

$$p(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

$$\begin{aligned} p_{\text{密度: 0.697}|\text{是}} &= p(\text{密度}=0.697 \mid \text{好瓜}=\text{是}) \\ &= \frac{1}{\sqrt{2\pi}\cdot 0.129} \exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right) \approx 1.959\ , \\ p_{\text{密度: 0.697}|\text{否}} &= p(\text{密度}=0.697 \mid \text{好瓜}=\text{否}) \\ &= \frac{1}{\sqrt{2\pi}\cdot 0.195} \exp\left(-\frac{(0.697-0.496)^2}{2\cdot 0.195^2}\right) \approx 1.203\ , \end{aligned}$$

$$p_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi}\cdot 0.101}\exp\left(-\frac{(0.460-0.279)^2}{2\cdot 0.101^2}\right) \approx 0.788,$$

$$p_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.108}\exp\left(-\frac{(0.460-0.154)^2}{2\cdot 0.108^2}\right) \approx 0.066.$$

于是，有:

$$P(\text{好瓜}=\text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.038,$$

$$P(\text{好瓜}=\text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80\times 10^{-5}.$$

由于0.038>6.80×10-5，因此，朴素贝叶斯分类器将测试样本"测1"判别为"好瓜".

## k均值

下面西瓜数据集4.0为例来演示k均值算法的学习过程.

| 编号 | 密度 | 含糖率 | 编号 | 密度 | 含糖率 | 编号 | 密度 | 含糖率 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.697 | 0.460 | 11 | 0.245 | 0.057 | 21 | 0.748 | 0.232 |
| 2 | 0.774 | 0.376 | 12 | 0.343 | 0.099 | 22 | 0.714 | 0.346 |
| 3 | 0.634 | 0.264 | 13 | 0.639 | 0.161 | 23 | 0.483 | 0.312 |
| 4 | 0.608 | 0.318 | 14 | 0.657 | 0.198 | 24 | 0.478 | 0.437 |
| 5 | 0.556 | 0.215 | 15 | 0.360 | 0.370 | 25 | 0.525 | 0.369 |
| 6 | 0.403 | 0.237 | 16 | 0.593 | 0.042 | 26 | 0.751 | 0.489 |
| 7 | 0.481 | 0.149 | 17 | 0.719 | 0.103 | 27 | 0.532 | 0.472 |
| 8 | 0.437 | 0.211 | 18 | 0.359 | 0.188 | 28 | 0.473 | 0.376 |
| 9 | 0.666 | 0.091 | 19 | 0.339 | 0.241 | 29 | 0.725 | 0.445 |
| 10 | 0.243 | 0.267 | 20 | 0.282 | 0.257 | 30 | 0.446 | 0.459 |

假定聚类簇数k=3，算法开始时随机选取三个样本 $x_6, x_{12}, x_{24}$ 作为初始均值向量，即

$$\boldsymbol{\mu}_1 = (0.403; 0.237),\ \boldsymbol{\mu}_2 = (0.343; 0.099),\ \boldsymbol{\mu}_3 = (0.532; 0.472)$$

考察样本 $x_1$=(0.697;0.460)，它与当前均值向量 $\mu_1, \mu_2, \mu_3$的距离分别为0.369，0.506，0.166，因此 $x_1$ 将被划入簇 $C_3$中.类似的，对数据集中的所有样本考察一遍后，可得当前簇划分为:

$$C_1 = \{\boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}, \boldsymbol{x}_{13}, \boldsymbol{x}_{14}, \boldsymbol{x}_{15}, \boldsymbol{x}_{17}, \boldsymbol{x}_{18}, \boldsymbol{x}_{19}, \boldsymbol{x}_{20}, \boldsymbol{x}_{23}\};$$

$$C_2 = \{\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{16}\};$$

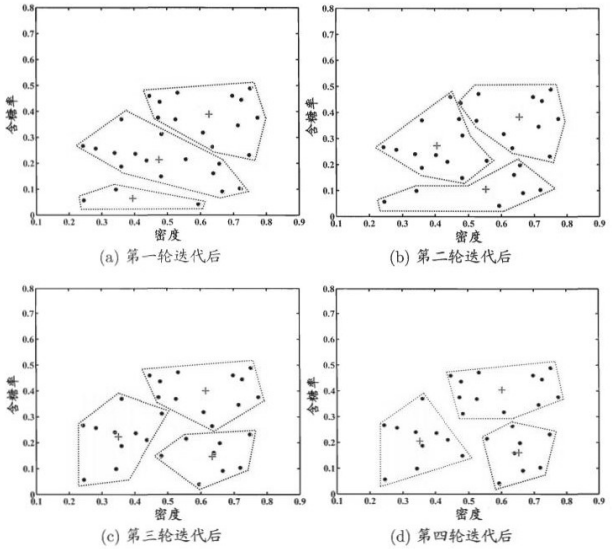
$$C_3 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \boldsymbol{x}_{24}, \boldsymbol{x}_{25}, \boldsymbol{x}_{26}, \boldsymbol{x}_{27}, \boldsymbol{x}_{28}, \boldsymbol{x}_{29}, \boldsymbol{x}_{30}\}.$$

根据公式： $\boldsymbol{\mu}_i' = \frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$ 可从 $C_1$、$C_2$、$C_3$分别求出新的均值向量:

$$\boldsymbol{\mu}_1' = (0.473; 0.214),\ \boldsymbol{\mu}_2' = (0.394; 0.066),\ \boldsymbol{\mu}_3' = (0.623; 0.388)$$

更新当前均值向量后，不断重复上述过程。

如图所示，第五轮迭代产生的结果与第四轮迭代相同、于是算法停止，得到最终的簇划分。




(a) 第一轮迭代后　(b) 第二轮迭代后

(c) 第三轮迭代后　(d) 第四轮迭代后

## 感知机

如图所示的训练数据集，其正实例点是 $x_1=(3,3)^T$, $x_2=(4,3)^T$ 负实例点是 $x_3=(1,1)^T$，试用感知机学习算法的原始形式求感知机模型 $f(x)=sign(w^Tx+b)$。

解 构建最优化问题：

$$\min_{w,b} L(w,b)=-\sum_{x_i\in M} y_i(w\cdot x+b)$$

求解w，b， 取学习率η =1。

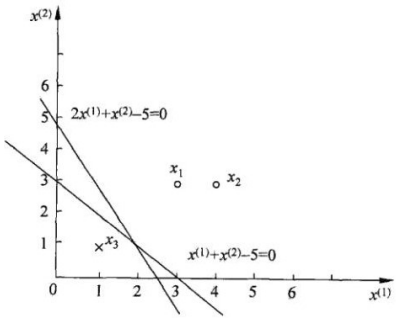
（1）取初值 $w_0=0, b_0=0$

（2）对 $x_1=(3,3)^T$，$y_1(w_0\cdot x_1+b_0)=0$ 未能被正确分类，更新 w，b：

$$w_1=w_0+y_1x_1=(3,3)^T,\quad b_1=b_0+y_1=1$$

得到线性模型：

$$w_1\cdot x+b_1=3x^{(1)}+3x^{(2)}+1$$




（3）对 $x_1, x_2$，显然，$y_i(w_1\cdot x_i+b_1)>0$，被正确分类，不修改 w， b； 对 $x_3=(1,1)^T$， $y_3(w_1\cdot x_3+b_1)<0$ 被误分类，更新 w，b。

$$w_2=w_1+y_3x_3=(2,2)^T,\quad b_2=b_1+y_3=0$$

得到线性模型：

$$w_2\cdot x+b_2=2x^{(1)}+2x^{(2)}$$

如此继续下去，直到：

$$w_7=(1,1)^T,\quad b_7=-3$$

$$w_7\cdot x+b_7=x^{(1)}+x^{(2)}-3$$

对所有数据点 $y_i(w_7\cdot x_i+b_7)>0$，没有误分类点，损失函数达到极小.

分离超平面为：$x^{(1)}+x^{(2)}-3=0$

感知机模型为：$f(x)=\text{sign}(x^{(1)}+x^{(2)}-3)$

迭代过程如下表所示

| 迭代次数 | 误分类点 | $w$ | $b$ | $w\cdot x+b$ |
|---|---|---|---|---|
| 0 | | 0 | 0 | 0 |
| 1 | $x_1$ | $(3,3)^T$ | 1 | $3x^{(1)}+3x^{(2)}+1$ |
| 2 | $x_3$ | $(2,2)^T$ | 0 | $2x^{(1)}+2x^{(2)}$ |
| 3 | $x_3$ | $(1,1)^T$ | −1 | $x^{(1)}+x^{(2)}-1$ |
| 4 | $x_3$ | $(0,0)^T$ | −2 | $-2$ |
| 5 | $x_1$ | $(3,3)^T$ | −1 | $3x^{(1)}+3x^{(2)}-1$ |
| 6 | $x_3$ | $(2,2)^T$ | −2 | $2x^{(1)}+2x^{(2)}-2$ |
| 7 | $x_3$ | $(1,1)^T$ | −3 | $x^{(1)}+x^{(2)}-3$ |
| 8 | 0 | $(1,1)^T$ | −3 | $x^{(1)}+x^{(2)}-3$ |

## KNN

给定一个二维空间的数据集：

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

构造一个平衡kd树。

解：根结点对应包含数据集T的矩形，选择 $x^{(1)}$ 轴，6个数据点的 $x^{(1)}$ 坐标的中位数是7，以平面 $x^{(1)}=7$ 将空间分为左、右两个子矩形（子结点）；接着，左矩形以 $x^{(2)}=4$ 分为两个子矩形，右矩形以 $x^{(2)}=6$ 分为两个子矩形，如此递归，最后得到下图所示的特征空间划分和kd树。