四、决策树

主讲教师：周志华

剪枝

为了尽可能正确分类训练样本，有可能造成分支过多 → 过拟合

可通过主动去掉一些分支来降低过拟合的风险

基本策略：

- 预剪枝 (pre-pruning)：提前终止某些分支的生长
- 后剪枝 (post-pruning)：生成一棵完全树，再“回头”剪枝

剪枝过程中需评估剪枝前后决策树的优劣 ⟶ 第 2 章

现在我们假定使用“留出法”

数据集

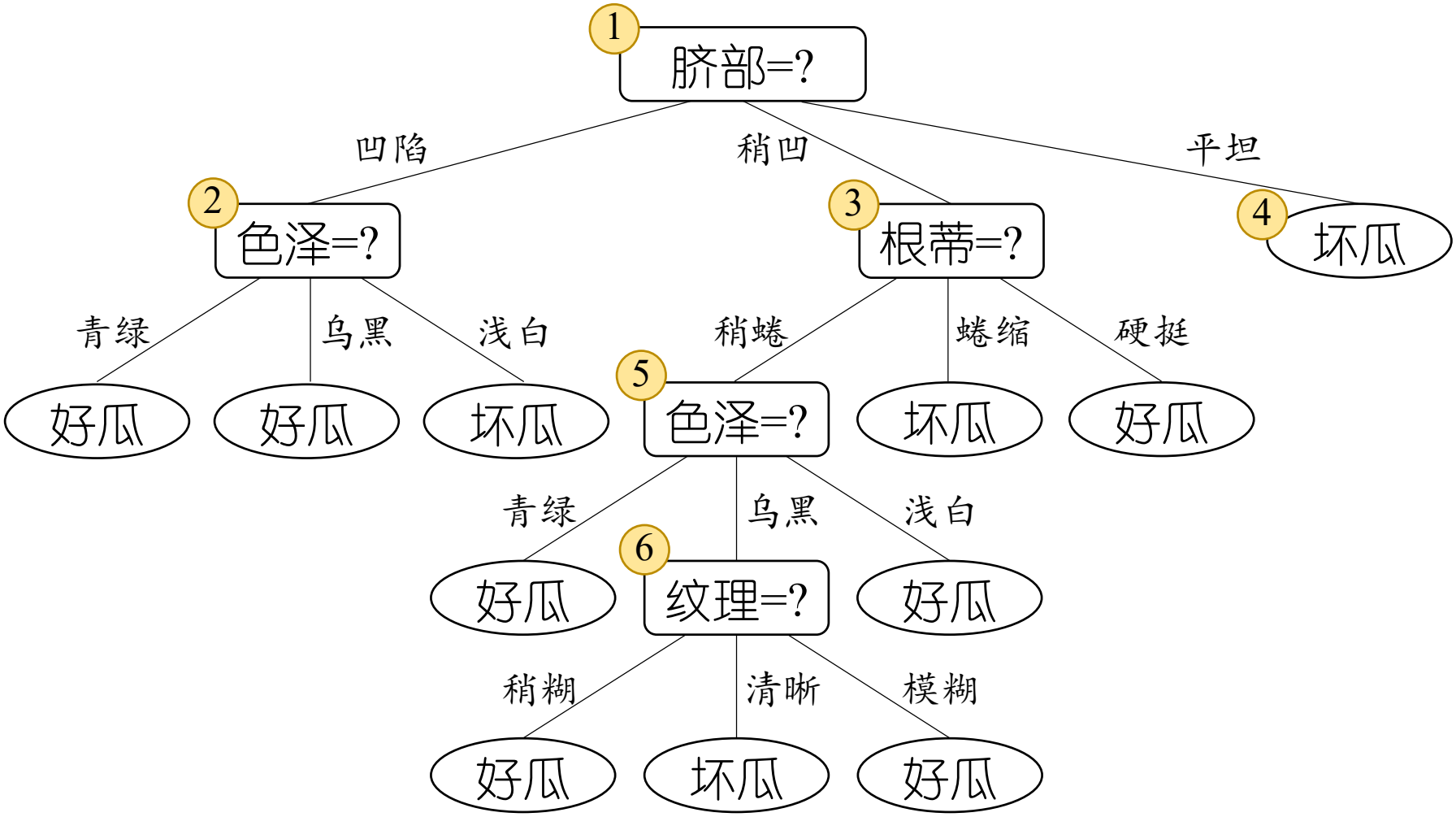
表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

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
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
10	青绿	硬挺	清脆	清晰	平坦	软粘	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
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
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
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
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

训练集

验证集

未剪枝决策树

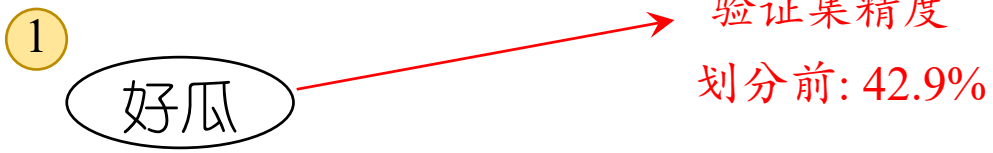

1
脐部=?
凹陷
稍凹
平坦
2
色泽=?
3
根蒂=?
4
坏瓜
青绿
乌黑
浅白
稍蜷
蜷缩
硬挺
好瓜
好瓜
坏瓜
5
色泽=?
坏瓜
好瓜
青绿
乌黑
浅白
6
好瓜
纹理=?
好瓜
稍糊
清晰
模糊
好瓜
坏瓜
好瓜

预剪枝

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
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
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
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

结点1：若不划分，则根结点为叶结点，类别标记为训练样例最多的类别，若选“好瓜”，则验证集中{4,5,8}被分类正确，验证集精度为 3/7 x 100% = 42.9%

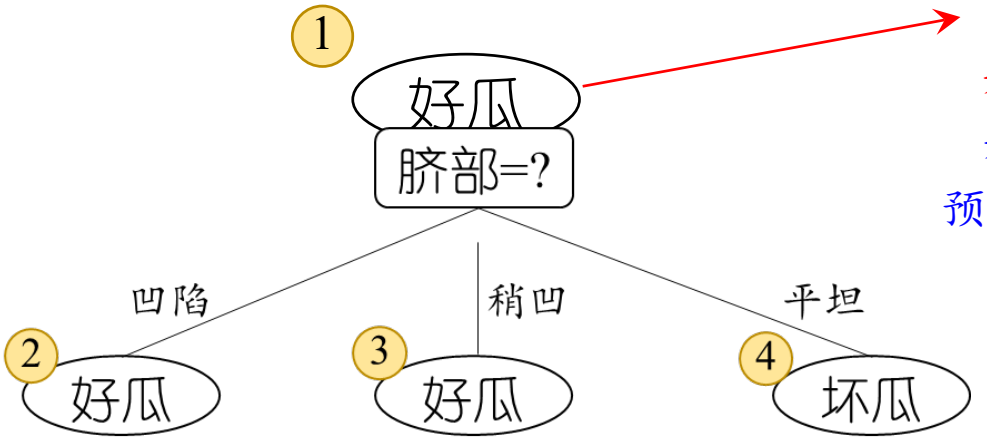


预剪枝

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
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
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
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

结点1：若不划分，则根结点为叶结点，类别标记为训练样例最多的类别，若选“好瓜”，则验证集中{4,5,8}被分类正确，验证集精度为 3/7 x 100% = 42.9%



验证集精度

划分前: 42.9%

划分后: 71.4%

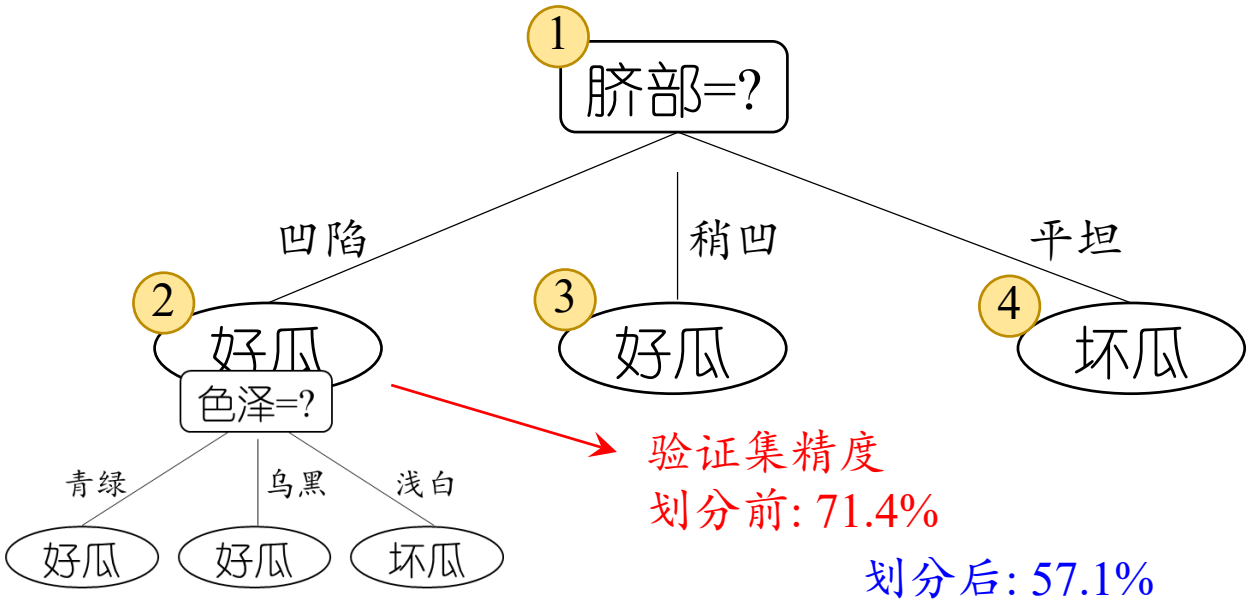
预剪枝决策: **划分**

结点1若划分，则根据划分后结点②③④ 的训练样例，它们将分别标记为“好瓜”“好瓜”“坏瓜”。此时，验证集中编号为 {4,5,8,11,12}的样例被划分正确，验证集精度为 5/7 x 100% = 71.4%

预剪枝

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
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
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
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



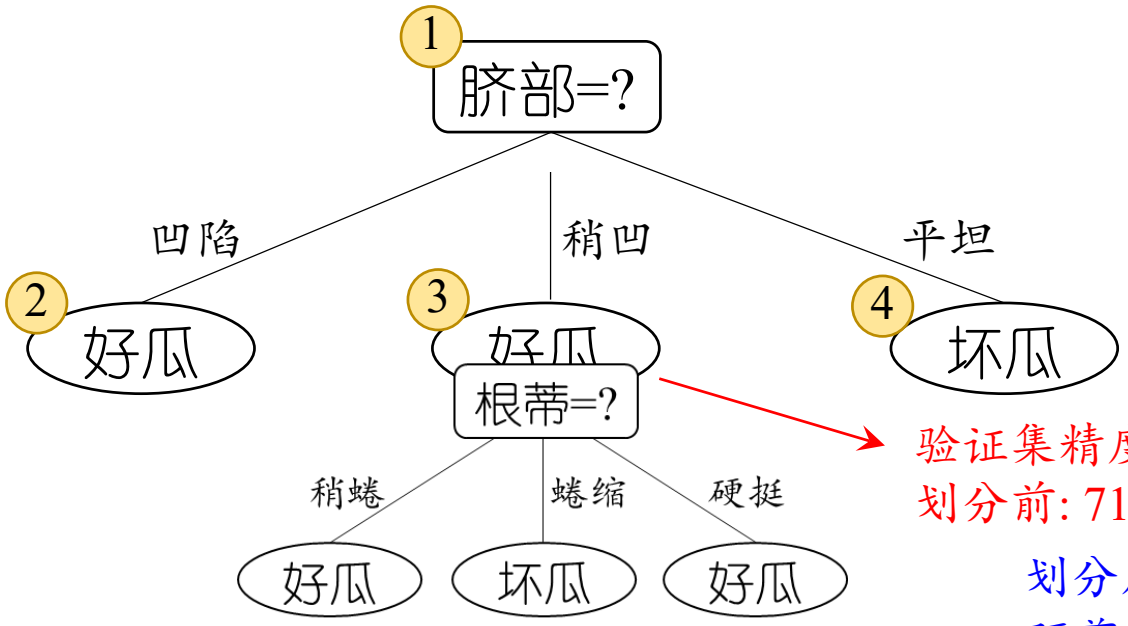
验证集精度
划分前: 71.4%
划分后: 57.1%
预剪枝决策: **禁止划分**

结点2：若划分，则验证集中{4,8,11,12} 被分类正确，验证集精度为 4/7 x100% = 57.1%

预剪枝

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
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
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
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



验证集精度
划分前: 71.4%

划分后: 71.4%

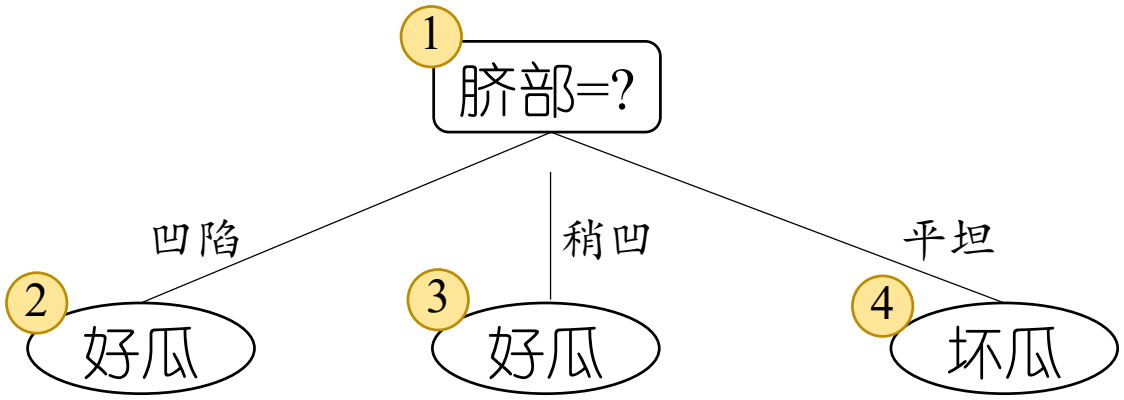
预剪枝决策: **禁止划分**

结点3：若划分，则验证集中{4,5,8,11,12} 被分类正确，验证集精度为 5/7 x100% = 71.4%

预剪枝

验证集

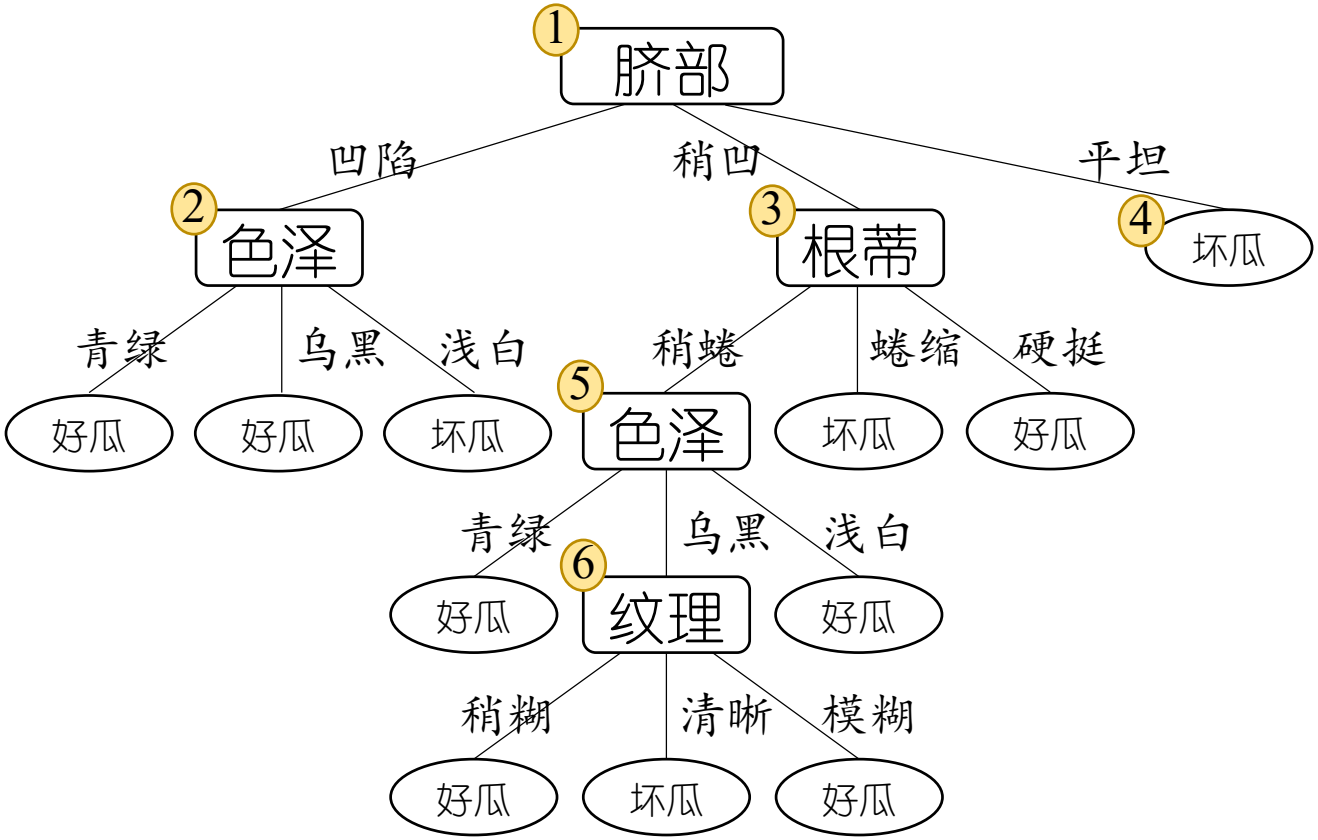
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否



最终，预剪枝的得到的决策树

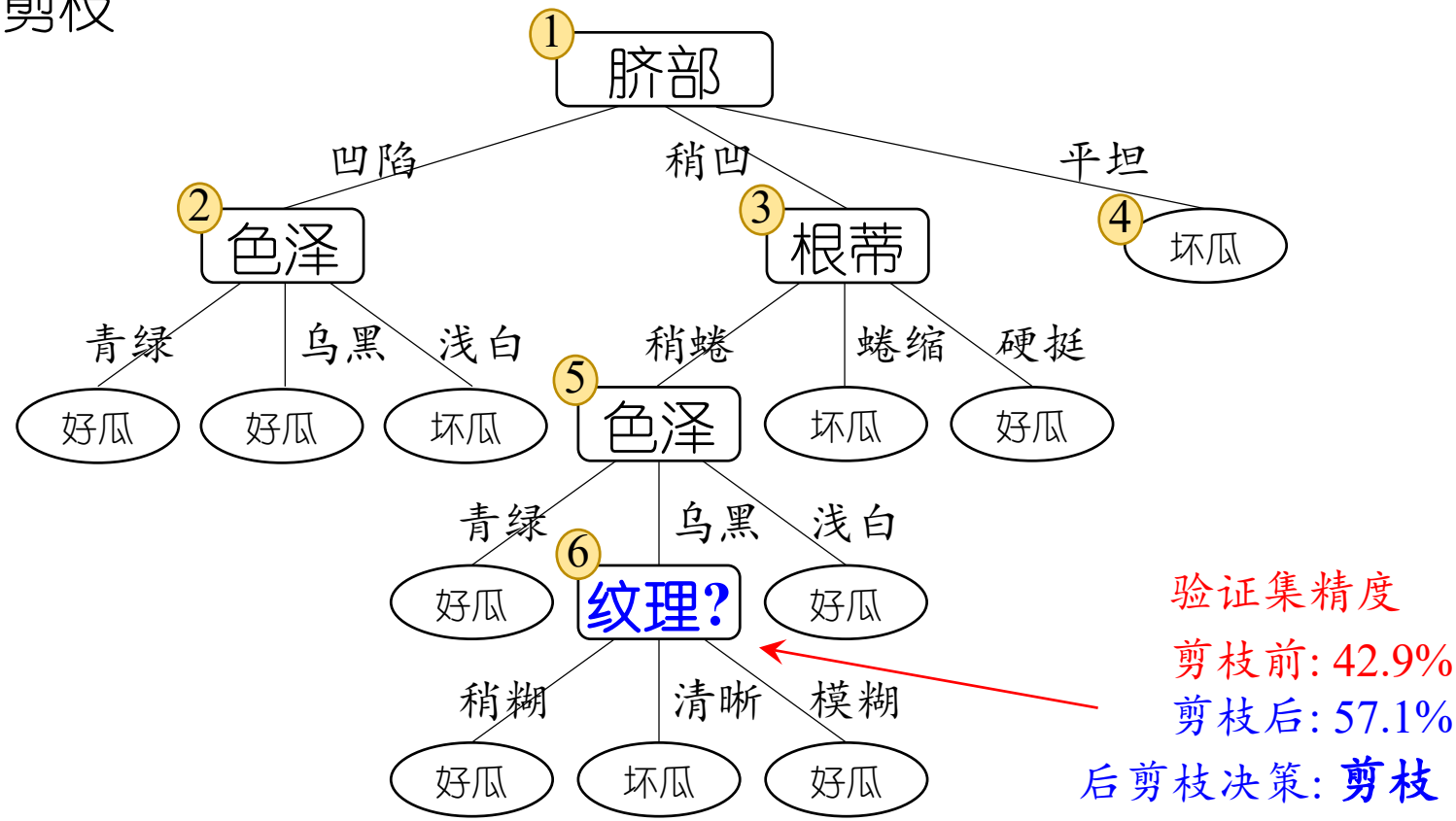
后剪枝

先生成一棵完整的决策树，其验证集精度测得为 42.9%



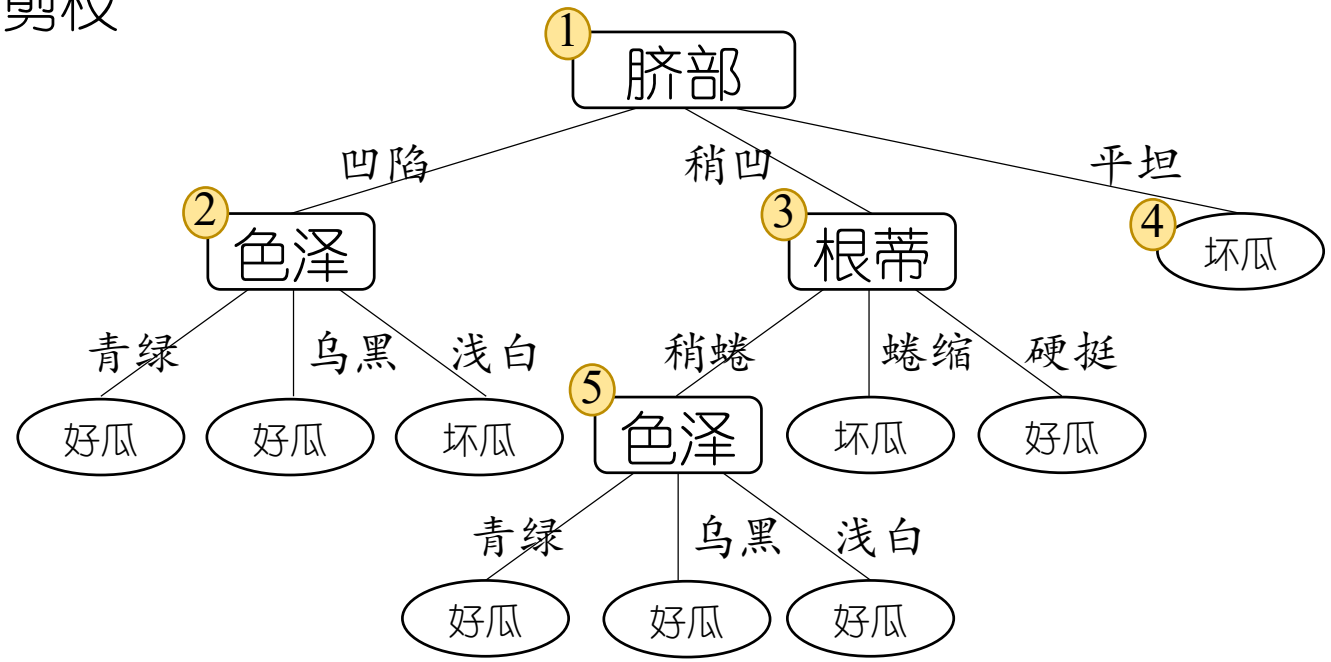
后剪枝 (续)

首先考虑结点⑥，若将其替换为叶结点，根据落在其上的训练样例 $\{7, 15\}$ 将其标记为“好瓜”，测得验证集精度提高至 57.1%，于是决定剪枝



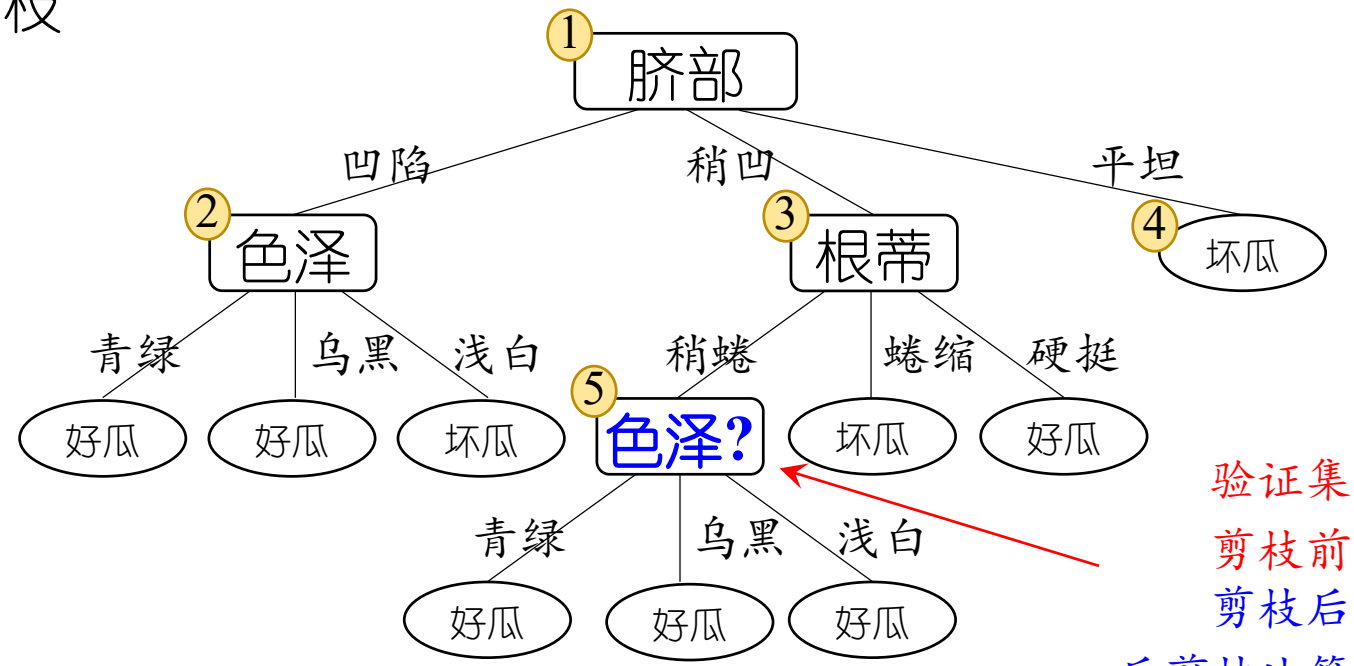
后剪枝 (续)

首先考虑结点⑥，若将其替换为叶结点，根据落在其上的训练样例 $\{7, 15\}$ 将其标记为“好瓜”，测得验证集精度提高至 57.1%，于是决定剪枝



后剪枝 (续)

然后考虑结点⑤，若将其替换为叶结点，根据落在其上的训练样例 $\{6, 7, 15\}$ 将其标记为“好瓜”，测得验证集精度仍为 57.1%，可以不剪枝



验证集精度

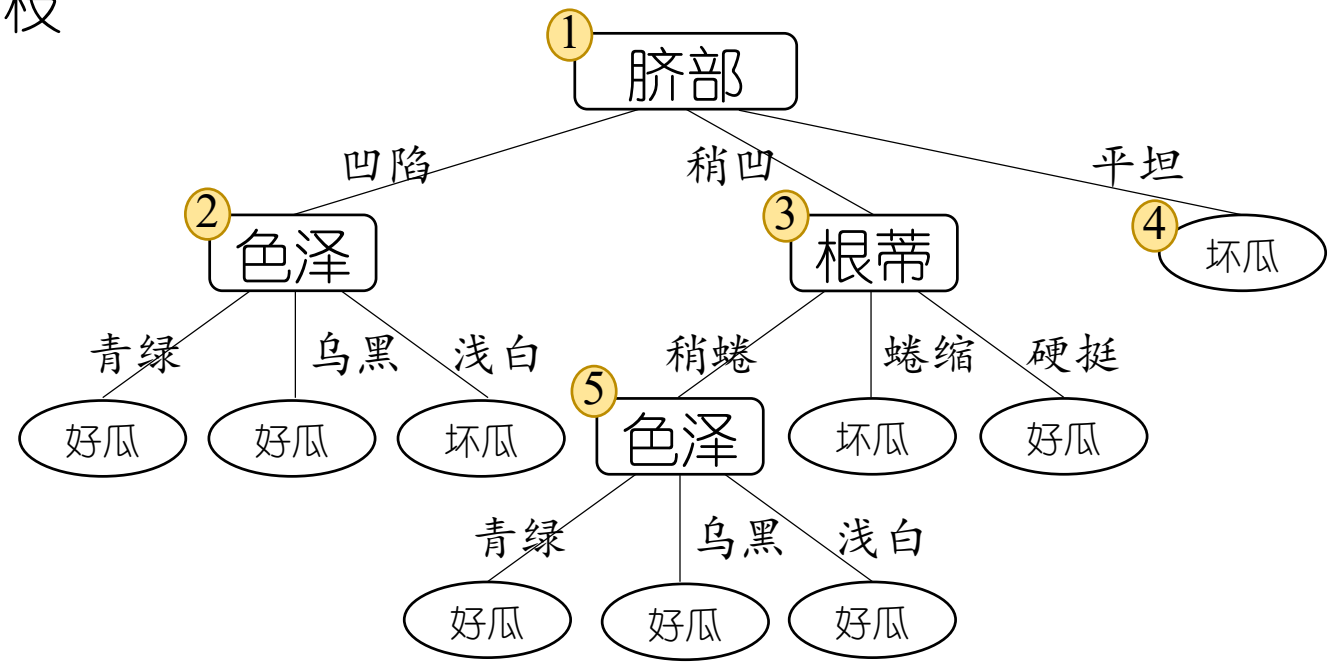
剪枝前: 57.1%

剪枝后: 57.1%

后剪枝决策: **不剪枝**

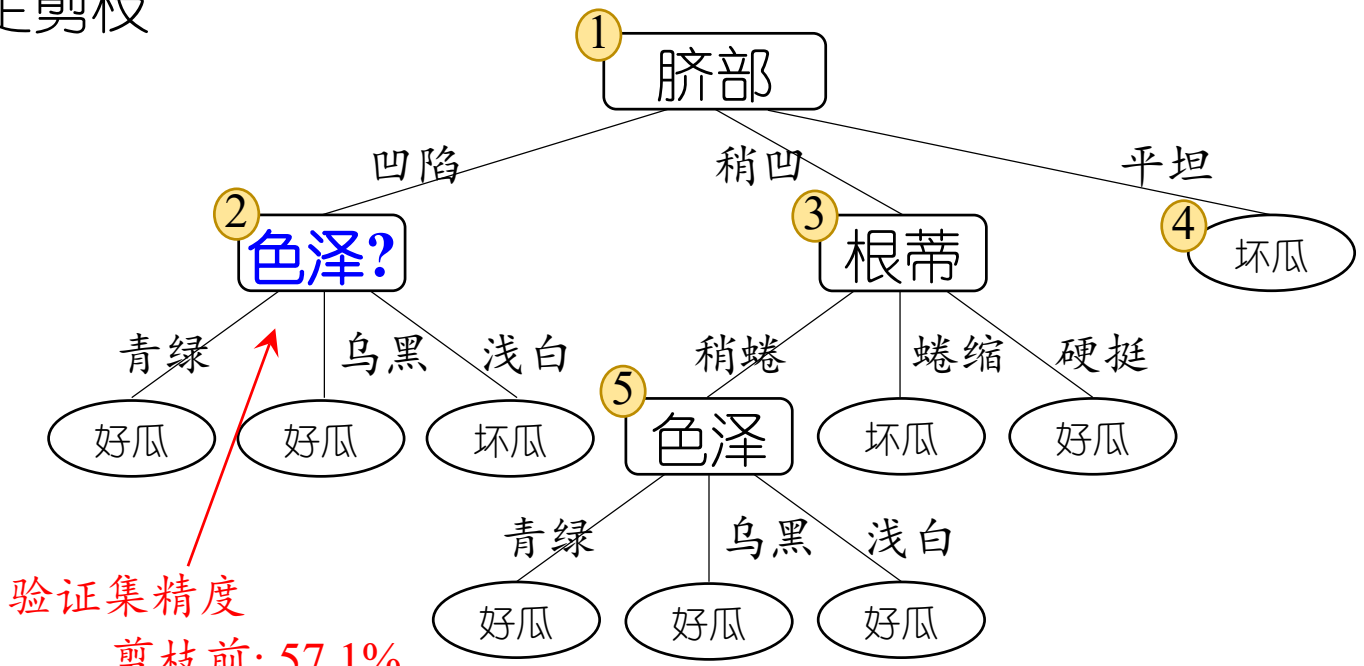
后剪枝 (续)

然后考虑结点⑤，若将其替换为叶结点，根据落在其上的训练样例 $\{6, 7, 15\}$ 将其标记为“好瓜”，测得验证集精度仍为 57.1%，可以不剪枝



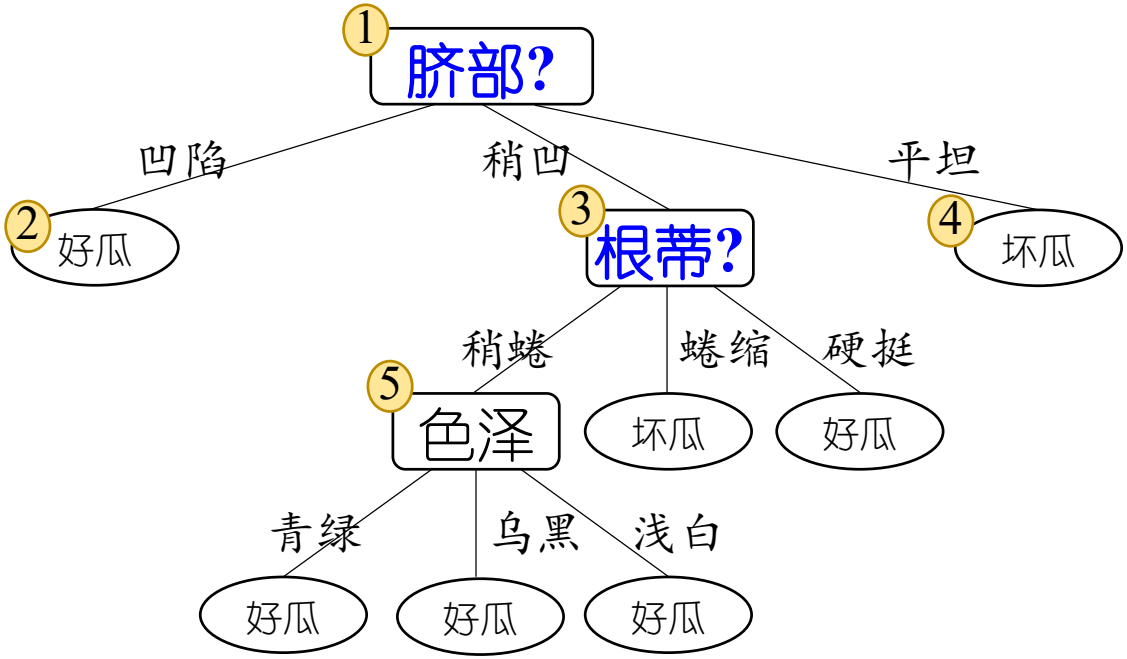
后剪枝 (续)

对结点②，若将其替换为叶结点，根据落在其上的训练样例 $\{1, 2, 3, 14\}$，将其标记为“好瓜”，测得验证集精度提升至 71.4%，决定剪枝



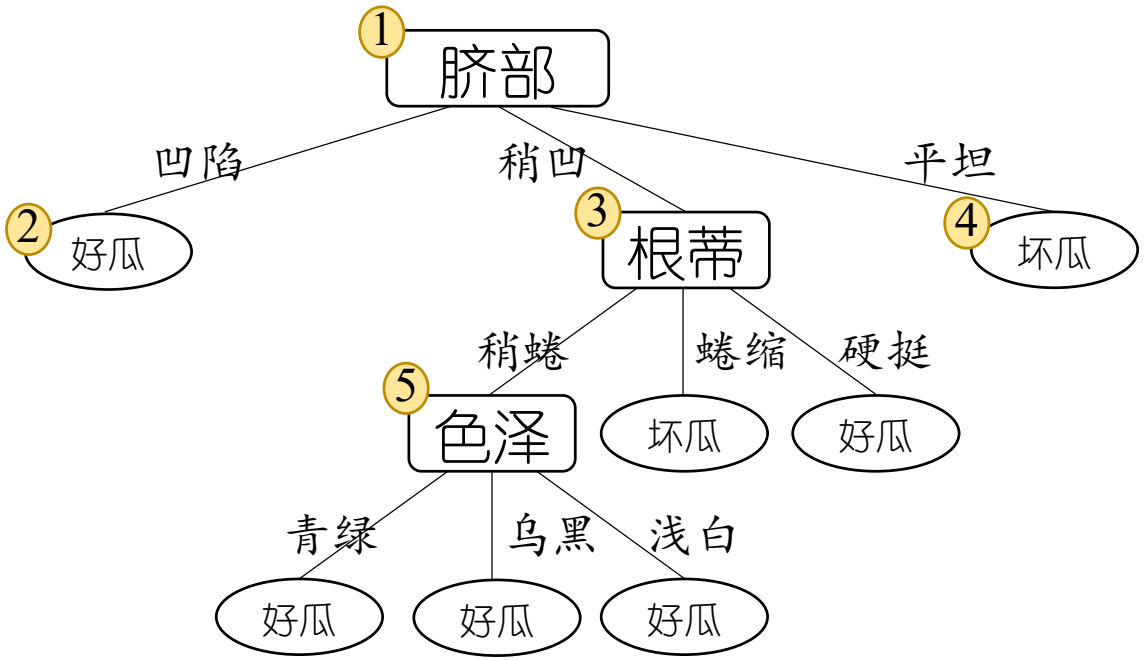
后剪枝 (续)

对结点③和①，先后替换为叶结点，均未测得验证集精度提升，于是不剪枝



后剪枝 (续)

最终，后剪枝得到的决策树：



预剪枝 vs. 后剪枝

- 时间开销：

- 预剪枝：测试时间开销降低，训练时间开销降低
- 后剪枝：测试时间开销降低，训练时间开销增加

- 过/欠拟合风险：

- 预剪枝：过拟合风险降低，欠拟合风险增加
- 后剪枝：过拟合风险降低，欠拟合风险基本不变

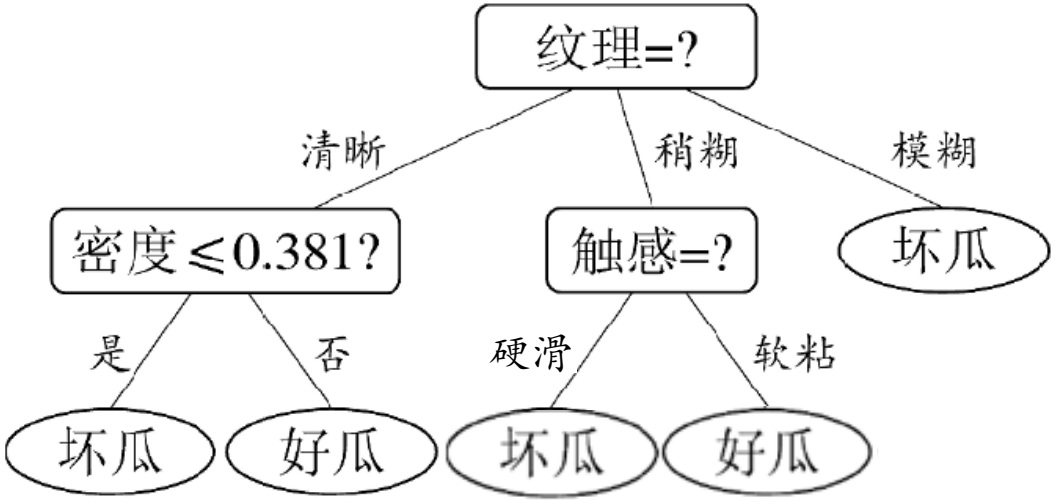
- 泛化性能：后剪枝 通常优于 预剪枝

连续值

基本思路：连续属性离散化

常见做法：二分法 (bi-partition)

- n 个属性值可形成 n-1 个候选划分
- 然后即可将它们当做 n-1 个离散属性值处理



缺失值

现实应用中，经常会遇到属性值“缺失”(missing)现象

仅使用无缺失的样例？ → 对数据的极大浪费

使用带缺失值的样例，需解决：

Q1：如何进行划分属性选择？

Q2：给定划分属性，若样本在该属性上的值缺失，如何进行划分？

基本思路：样本赋权，权重划分

一个例子

仅通过无缺失值的样例来判断划分属性的优劣

学习开始时，根结点包含样例集 D 中全部17个样例，权重均为 1

表 4.4 西瓜数据集 2.0α

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

以属性“色泽”为例，该属性上无缺失值的样例子集 $\tilde{D}$ 包含 14 个样例，信息熵为

$$\mathrm{Ent}(\tilde{D}) = -\sum_{k=1}^{2} \tilde{p}_k \log_2 \tilde{p}_k = -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985$$

一个例子

令 $\tilde{D}^1$, $\tilde{D}^2$, $\tilde{D}^3$ 分别表示在属性“色泽”上取值为“青绿”“乌黑”以及“浅白”的样本子集，有

$$\mathrm{Ent}(\tilde{D}^1) = -\left(\tfrac{2}{4}\log_2\tfrac{2}{4} + \tfrac{2}{4}\log_2\tfrac{2}{4}\right) = 1.000 \quad \mathrm{Ent}(\tilde{D}^2) = -\left(\tfrac{4}{6}\log_2\tfrac{4}{6} + \tfrac{2}{6}\log_2\tfrac{2}{6}\right) = 0.918$$

$$\mathrm{Ent}(\tilde{D}^3) = -\left(\tfrac{0}{4}\log_2\tfrac{0}{4} + \tfrac{4}{4}\log_2\tfrac{4}{4}\right) = 0.000$$

因此，样本子集 $\tilde{D}$ 上属性“色泽”的信息增益为

$$\begin{aligned}\mathrm{Gain}(\tilde{D}, \text{色泽}) &= \mathrm{Ent}(\tilde{D}) - \sum_{v=1}^{3} \tilde{r}_v \mathrm{Ent}(\tilde{D}^v) \\ &= 0.985 - \left(\tfrac{4}{14} \times 1.000 + \tfrac{6}{14} \times 0.918 + \tfrac{4}{14} \times 0.000\right) \\ &= 0.306\end{aligned}$$

无缺失值样例中
属性 a 上取值为 v 的样例占比

于是，样本集 D 上属性“色泽”的信息增益为

$$\mathrm{Gain}(D, \text{色泽}) = \rho \times \mathrm{Gain}(\tilde{D}, \text{色泽}) = \tfrac{14}{17} \times 0.306 = 0.252$$

样本集中 属性 a 上无缺失值的样例占比

一个例子

类似地可计算出所有属性在数据集上的信息增益

$$\text{Gain}(D, \text{色泽}) = 0.252 \qquad \text{Gain}(D, \text{根蒂}) = 0.171$$

$$\text{Gain}(D, \text{敲声}) = 0.145 \qquad \text{Gain}(D, \text{纹理}) = 0.424$$

$$\text{Gain}(D, \text{脐部}) = 0.289 \qquad \text{Gain}(D, \text{触感}) = 0.006$$

进入“纹理=清晰”分支

进入“纹理=稍糊”分支

进入“纹理=模糊”分支

样本权重在各子结点仍为1

在“纹理”上出现缺失值，样本 8, 10 同时进入三个分支，三分支上的权重分别为 7/15，5/15，3/15

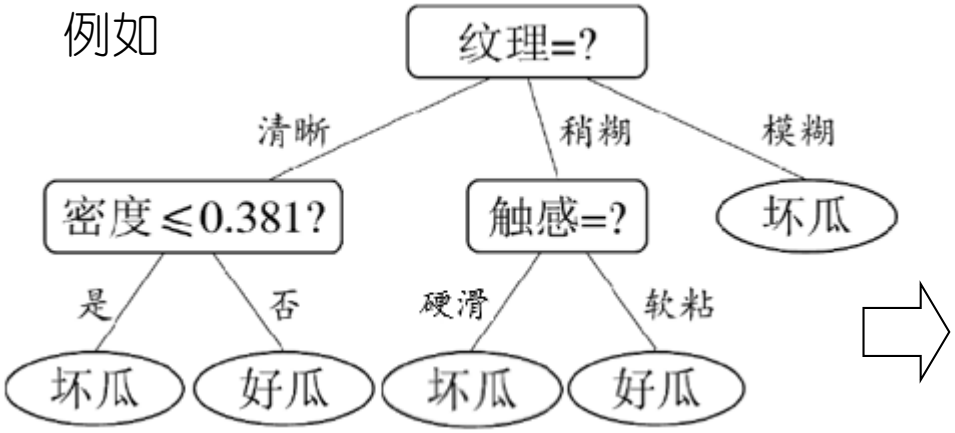
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	–	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	–	是
3	乌黑	蜷缩	–	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	–	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	–	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	–	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	–	否
12	浅白	蜷缩	–	模糊	平坦	软粘	否
13	–	稍蜷	浊响	稍糊	凹陷	硬滑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

权重划分

从“树”到“规则”

- 一棵决策树对应于一个“规则集”
- 每个从根结点到叶结点的分支路径对应于一条规则

例如



- IF (纹理=清晰) ∧ (密度≤0.381) THEN 坏瓜
- IF (纹理=清晰) ∧ (密度>0.381) THEN 好瓜
- IF (纹理=稍糊) ∧ (触感=硬滑) THEN 坏瓜
- IF (纹理=稍糊) ∧ (触感=软粘) THEN 好瓜
- IF (纹理=模糊) THEN 坏瓜

好处：

- 改善可理解性
- 进一步提升泛化能力

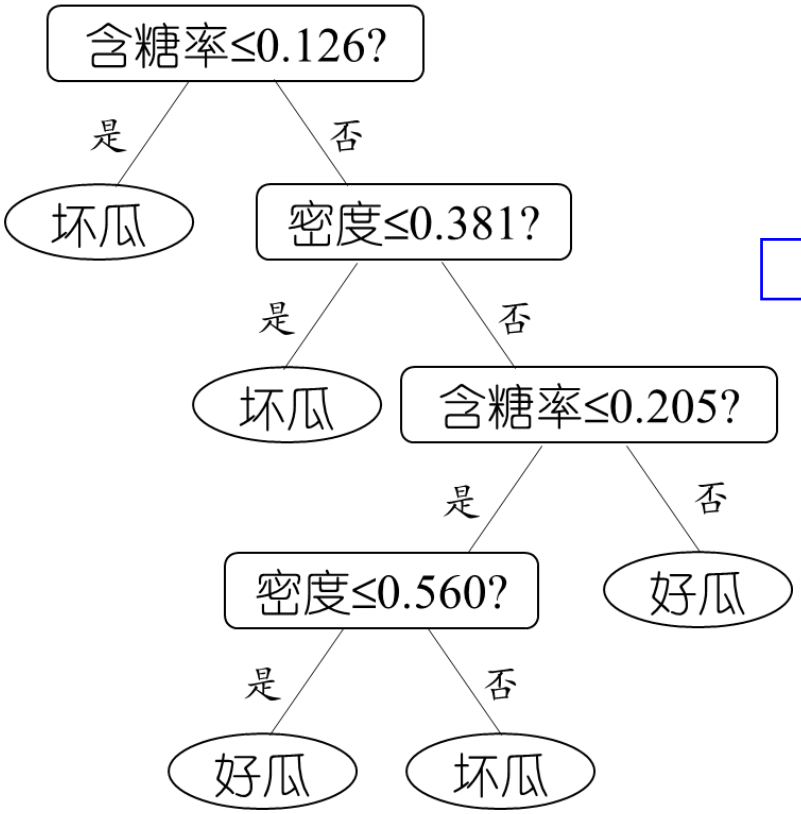
由于转化过程中通常会进行前件合并、泛化等操作

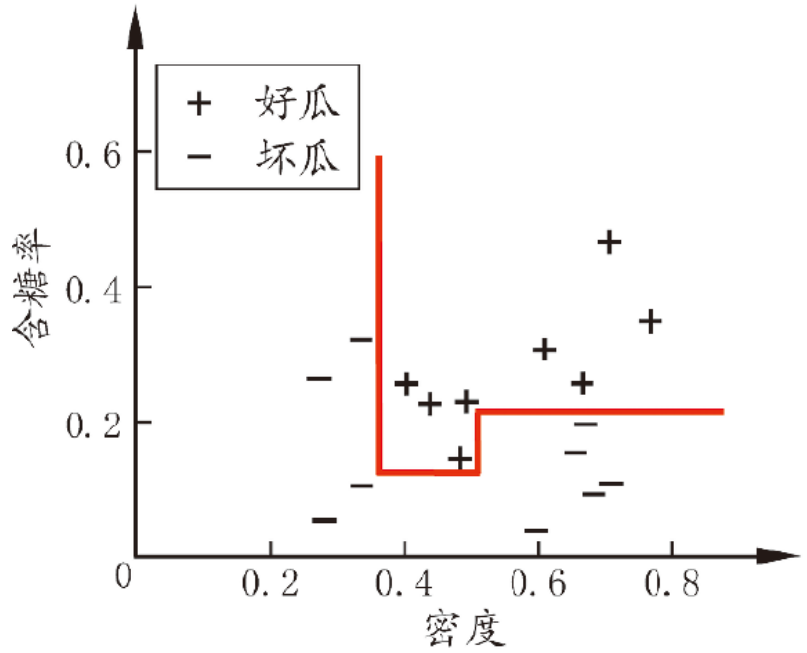
例如 C4.5Rule 的泛化能力通常优于 C4.5决策树

轴平行划分

单变量决策树：在每个非叶结点仅考虑一个划分属性

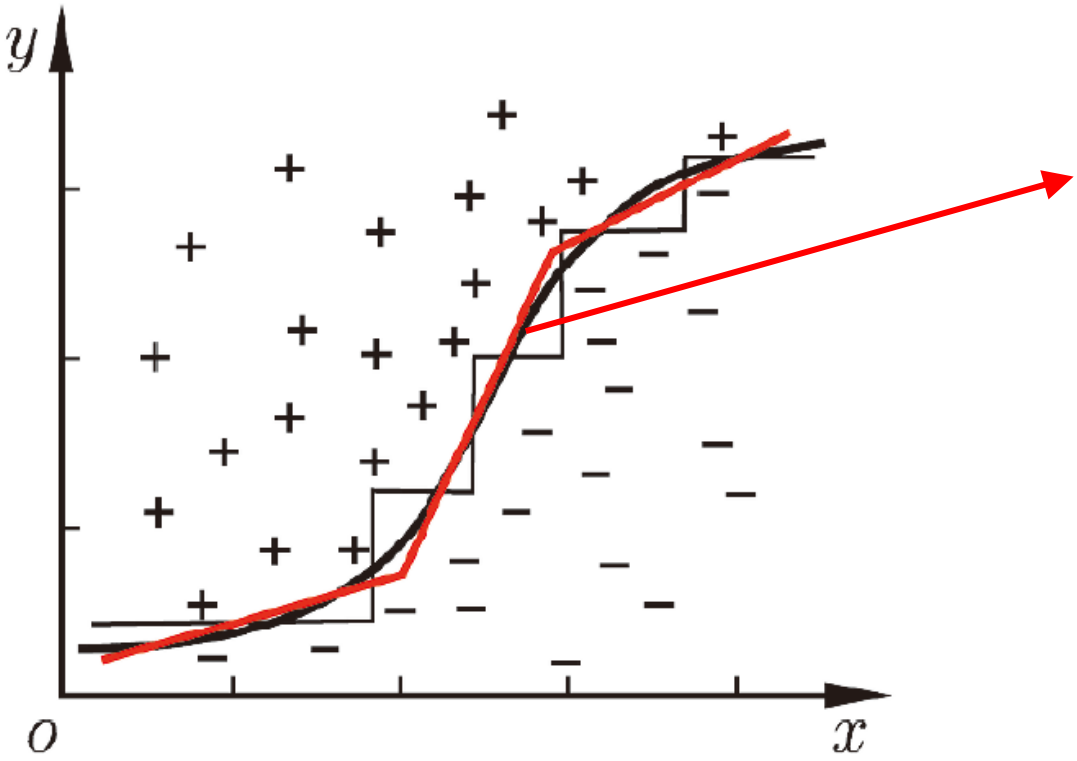
产生“轴平行”分类面





轴平行 vs. 倾斜

当学习任务所对应的分类边界很复杂时，需要非常多段划分才能获得较好的近似

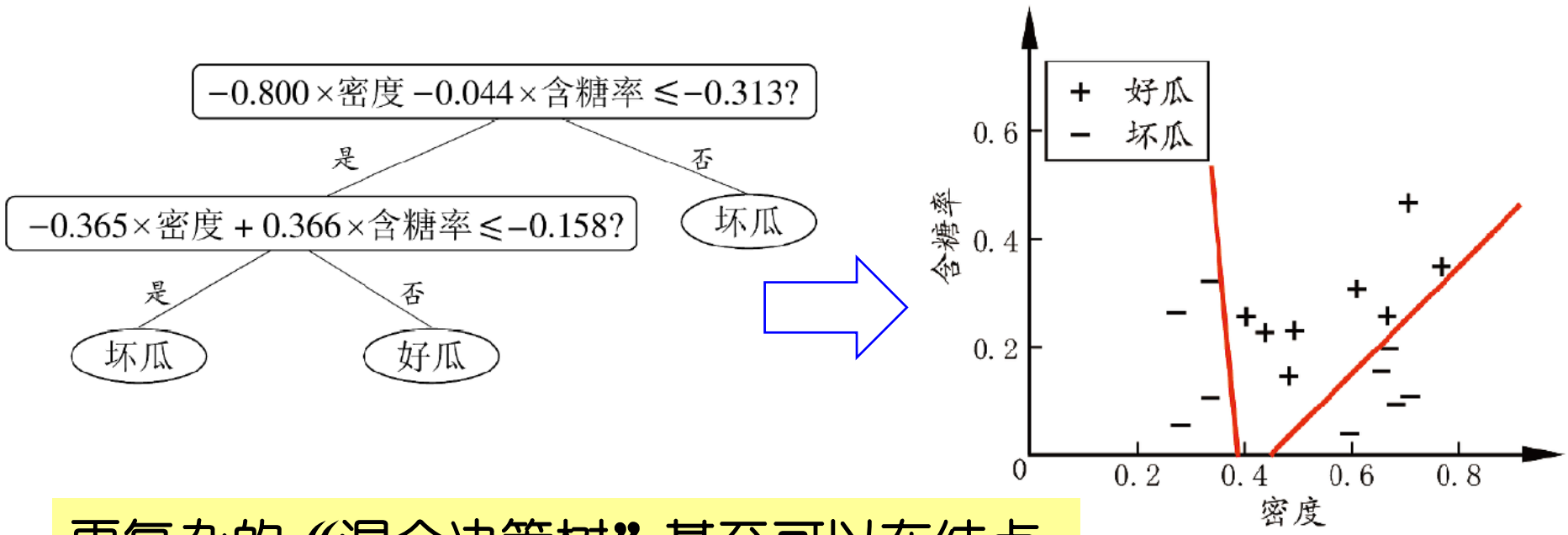


能否产生这样的分类边界？

多变量(multivariate)决策树

多变量决策树：每个非叶结点不仅考虑一个属性

例如“斜决策树” (oblique decision tree) 不是为每个非叶结点寻找最优划分属性，而是建立一个线性分类器



更复杂的“混合决策树”甚至可以在结点嵌入神经网络或其他非线性模型

前往第五站……

五、神经网络

主讲教师：周志华

什么是神经网络？

“神经网络是由具有适应性的简单单元组成的广泛并行互连的网络，它的组织能够模拟生物神经系统对真实世界物体所作出的交互反应”

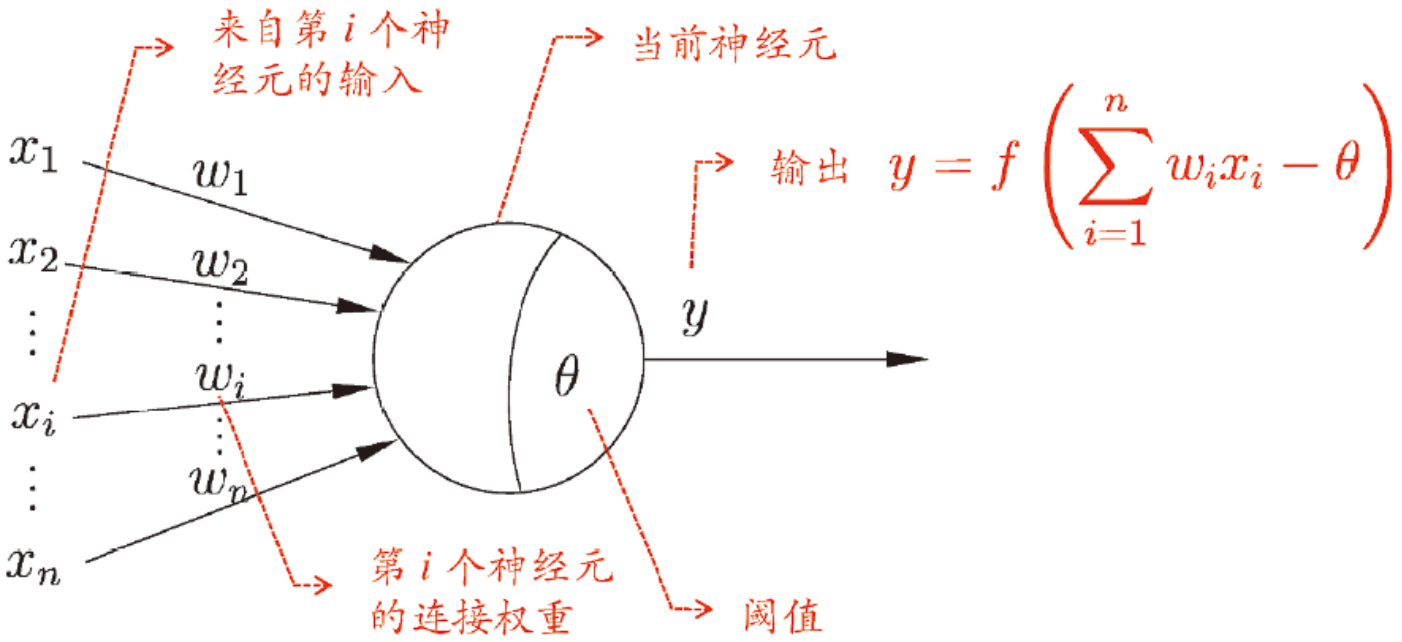
[T. Kohonen, 1988，*Neural Networks* 创刊号]

神经网络是一个很大的学科领域，本课程仅讨论神经网络与机器学习的交集，即“神经网络学习”

亦称“连接主义(connectionism)”学习

“简单单元”：神经元模型

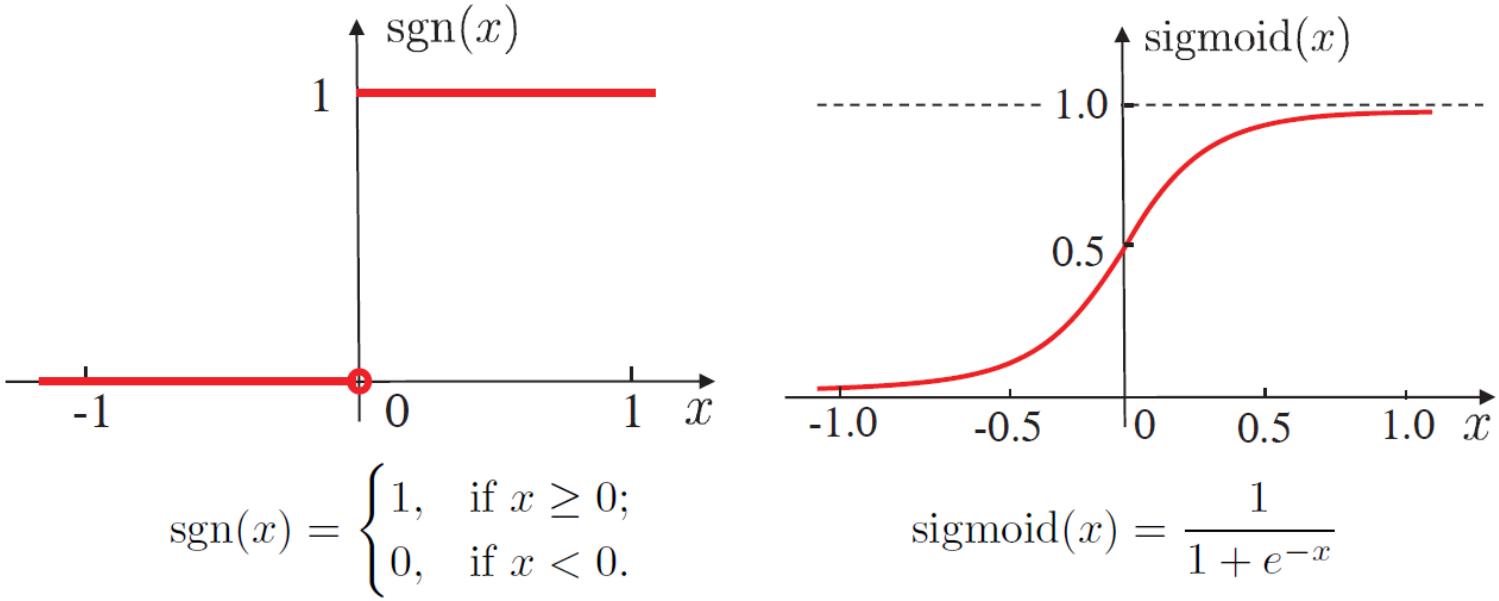
M-P 神经元模型 [McCulloch and Pitts, 1943]



神经网络学得的知识蕴含在连接权与阈值中

神经元的“激活函数”

- 理想激活函数是阶跃函数，0表示抑制神经元而1表示激活神经元
- 阶跃函数具有不连续、不光滑等不好的性质，常用的是 Sigmoid 函数



$$\mathrm{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases}$$

$$\mathrm{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

(a) 阶跃函数　　(b) Sigmoid 函数

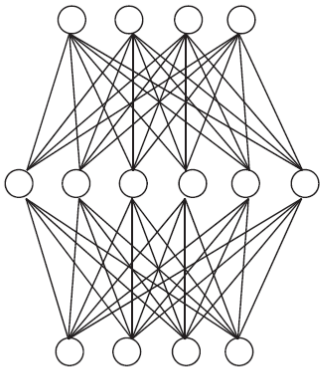
图 5.2　典型的神经元激活函数

多层前馈网络结构

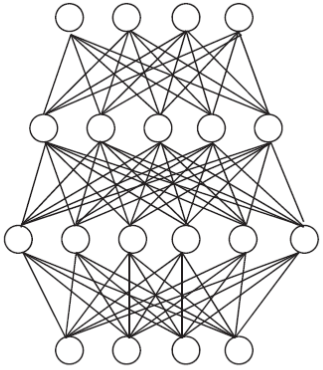
多层网络：包含隐层的网络

前馈网络：神经元之间不存在同层连接也不存在跨层连接

隐层和输出层神经元亦称“功能单元”(functional unit)



(a) 单隐层前馈网络



(b) 双隐层前馈网络

多层前馈网络有强大的表示能力（”万有逼近性”）

仅需一个包含足够多神经元的隐层，多层前馈神经网络就能以任意精度逼近任意复杂度的连续函数　　[Hornik et al., 1989]

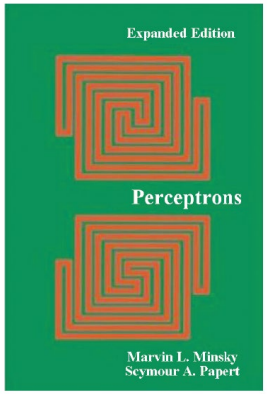
但是，如何设置隐层神经元数是未决问题(Open Problem)．实际常用“试错法”

神经网络发展回顾

1940年代-萌芽期：M-P模型 (1943), Hebb 学习规则 (1945)

1956左右-1969左右~繁荣期：感知机 (1958), Adaline (1960), …

1969年：Minsky & Papert “Perceptrons”



马文·闵斯基
(1927–2016)
1969年图灵奖

冰河期

1984左右 -1997左右~繁荣期：Hopfield (1983), BP (1986), …

1997年左右：SVM文本分类成功 及 统计学习 兴起

沉寂期

2012-至今~繁荣期：深度学习

交替模式：
热十三？年
冷十五？年

2019年3月27日，ACM宣布：

Geoffrey Hinton, Yann LeCun, Yoshua Bengio

因对深度学习的卓越贡献获得图灵奖

科学的发展总是
“螺旋式上升”

三十年河东
三十年河西

坚持才能有结果