机器学习（西瓜书）

注　解

（第 4 章　决策树）

https://blog.csdn.net/jbb0523

前言

经常听人说南大周老师所著的《机器学习》（以下统称为西瓜书）是一本入门教材，是一本科普性质的教科书。在该书第十次印刷之际，周老师在“如何使用本书”中也提到“这是一本入门级教科书”。然而，本人读起来却感觉该书远不止“科普”“入门”那么简单，书中的很多公式需要思考良久方能推导，很多概念需要反复咀嚼才能消化。边读边想着是不是应该将自己学习时遇到的一些知识难点的解析分享出来，以帮助更多的人入门。自己的确也随手做过一些笔记，但由于怀疑这仅是自己的个别现象，毕竟读书期间，思考更多的是如何使用单片机、DSP、ARM、FGPA 等，而这些基本是不需要推导任何公式的，因此作罢。偶然间在周老师的新浪微博看到如下对话：

此时方知，可能“读不懂”并不是个别现象。因此决定写一本“西瓜书注解”或者称为“西瓜书读书笔记”，对自己研读西瓜书时遇到的“台阶”进行解释和推导，以帮助更多的人能够更快地进入到这个领域。另外，近期越来越强地意识到，扎扎实实地推导一些基础算法的公式，无论是对于理解算法本身机理还是进行学术研究，都是非常有必要的。

自己会根据个人学习进度和研究需要按章发布，不知道能不能坚持写完，加油！

毕竟自己也是一名初学者，所以可能一些概念解释并不完整、一些公式推导并不优美，甚至会存在错误，这是不可避免的，不接受漫骂，但欢迎将问题反馈给我，共同学习进步！

（网盘链接：https://pan.baidu.com/s/1QtEiNnk8jMzmbs0KPBN-_w）

第 4 章目录

第 4 章 决策树

正如前言所述，从本章至第 10 章，介绍一些经典而常用的机器学习方法。这些方法大部分为分类方法，也穿插了少量回归算法（例如 6.5 节的支持向量回归）；大部分为监督学习方法，少量为无监督学习方法（第 9 章的聚类和第 10 章的降维）。

4.1 基本流程

作为本章的开篇，首先要明白决策树在做什么。正如图 4.1 所示的决策过程，决策树就是不断根据某属性进行划分的过程（注意：每次决策时考虑范围是在上次决策结果的限定范围之内的），即“If…Elseif…Else……”的决定过程。

但是，划分到什么时候，就停止划分呢？这就是图 4.2 中的三个“**return**”代表的递归返回。以下解释图 4.2 中的三种递归返回（可参考表 4.1 西瓜数据集 2.0）：

首先，应该明白决策树的基本流程是根据某种原则（即图 4.2 第 8 行）每次选择一个属性按属性的取值（例如西瓜的“触感”包括硬滑和软粘）进行划分（将所有“触感”为硬滑西瓜的分到一起，将所有“触感”为软粘的西瓜分到一起；根据该属性划分后，产生的各子集内部样本在该属性上的取值分别相同），依此再对各划分子集进行递归划分。然而有几种特殊情况：

(I)**划分期望得到的结果是各子集只包含同一类别的样本**，例如好瓜（或坏瓜），若递归划分过程中某子集已经只含有某一类别的样本（例如只含好瓜），那么此时划分的目的已经达到了，无需再进行划分，此即为**递归返回的情形(1)**；

(II)递归划分时每次选择一个属性，并且划分依据属性不能重复使用，例如本次根据“触感”对当前样本子集D进行划分，那么后面对该子集D划分成的子集（及子集的子集……）再次进行递归划分时不能再使用“触感”（图 4.2 第 14 行的$A \setminus \{a_*\}$表示从候选依据属性中将当前使用的依据属性去除，这是因为根据某属性划分之后，产生的各个子集在该属性上的取值相同）；但样本的属性是有限的，因此划分次数不超过属性个数；若所有属性均已被作为划分依据，此时子集中仍含有多类样本（例如仍然同时含有好瓜和坏瓜），但是因已无属性可用作划分依据（即子集中样本在各属性上取值均相同，但却无法达到**子集只包含同一类别的样本**），此时只能少数服从多数，以此子集中样本数最多的类为标记，此即为**递归返回的情形(2)**的$A = \varnothing$；

(III)每递归一次，候选的属性个数就减少一个；假设现在还剩一个属性可用（比如“触感”），而且子集中仍含有多类样本（例如仍然同时含有好瓜和坏瓜），由于只剩一个属性可用，我们也只能在这个属性上划分，但是子集所有样本在该属性上的取值都相同（例如“触感”都是硬滑），这就尴尬了，没办法划分了（大家的“触感”都一样，怎么划分？其实剩余多个属性也类似的，划分就是根据属性的不同取值将当前样本集合分为多个子集，若当前样本集合在任何候选属性上的取值相同，则无法划分），此时也只能少数服从多数，以此子集中样本数最多的类为标记，此即为**递归返回的情形(2)**的“D中样本在A上取值相同”；

(IV)根据某属性进行划分时，若该属性多个属性值中的某个属性值不包含任何样本（这可能是训练集的样本不够充分造成的），例如对当前子集D以“纹理”属性来划分，“纹理”共有三个值：清晰、稍糊、模糊，但发现当前子集D中并无样本“纹理”属性取值为模糊（即子集D中样本或为清晰，或为稍糊），此时对于取值为清晰的子集和取值为稍糊的子集继续递归，而对于取值为模糊的分支，因为无样本落入，将其标记为叶结点，其类别标记为D中

样本最多的类（即把父结点的样本分布作为当前结点的先验分布）；注意，此分支必须保留，因为测试时，可能会有样本落入该分支。此即**递归返回的情形(3)。**

如此用文字叙述还是比较抽象，接下来使用在表 4.1 西瓜数据集 2.0 上基于信息增益生成的图 4.4 决策树来具体说明：

(I) 根据“纹理”属性划分后，发现“纹理”为模糊的 3 个样本全部是坏瓜，即**递归返回的情形(1)**：当前结点包含的样本全属于同一类别，无需划分；

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
8	乌黑	稍蜷	浊响	**清晰**	稍凹	硬滑	是
10	青绿	硬挺	清脆	**清晰**	平坦	软粘	否
15	乌黑	稍蜷	浊响	**清晰**	稍凹	软粘	否
7	乌黑	稍蜷	浊响	**稍糊**	稍凹	软粘	是
9	乌黑	稍蜷	沉闷	**稍糊**	稍凹	硬滑	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
17	青绿	蜷缩	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
16	浅白	蜷缩	浊响	**模糊**	平坦	硬滑	否

同理，接下来对“纹理”为清晰的样本，再次根据“根蒂”属性划分后，发现“根蒂”为蜷缩的样本全部是好瓜，“根蒂”为硬挺的样本全部是坏瓜，即**递归返回的情形(1)**；此时只有“根蒂”为稍蜷的样本同时含有好瓜和坏瓜，因此需要继续进行划分：

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
8	乌黑	**稍蜷**	浊响	**清晰**	稍凹	硬滑	是
15	乌黑	**稍蜷**	浊响	**清晰**	稍凹	软粘	否
10	青绿	**硬挺**	清脆	**清晰**	平坦	软粘	否

(II)接下来对“根蒂”为稍蜷的样本，再次根据“色泽”属性进行划分后，发现“色泽”为青绿的样本全部是好瓜，即**递归返回的情形(1)**；发现没有“色泽”为浅白的样本，即**递归返回情形(3)**：当前结点包含的样本集合为空，不能划分；因此同样把当前结点标记为叶结点，将其类别设定为其父结点所含样本最多的类别，即好瓜（如下表所示，其父结点包含三个样本，其中两个为好瓜，一个不是好瓜）：

编号	**色泽**	**根蒂**	敲声	**纹理**	脐部	触感	好瓜
6	**青绿**	**稍蜷**	浊响	**清晰**	稍凹	软粘	是
8	**乌黑**	**稍蜷**	浊响	**清晰**	稍凹	硬滑	是
15	**乌黑**	**稍蜷**	浊响	**清晰**	稍凹	软粘	否

(III)注意：图 4.4 的决策树中没有出现递归返回情形(2)，这是因为表 4.1 的西瓜数据集 2.0 样本数太少。假设表 4.1 的数据集只包含“纹理”和“根蒂”两个属性，则根据“纹理”划分之后，再对“纹理”为清晰的样本根据“根蒂”划分，发现两个属性都已使用，但“根蒂”为稍蜷的样本子集中仍同时包含好瓜和坏瓜，即**递归返回的情形(2)**的$A = \varnothing$：

编号	**根蒂**	**纹理**	好瓜
1	**蜷缩**	**清晰**	是
2	**蜷缩**	**清晰**	是
3	**蜷缩**	**清晰**	是
4	**蜷缩**	**清晰**	是
5	**蜷缩**	**清晰**	是
6	**稍蜷**	**清晰**	是
8	**稍蜷**	**清晰**	是
15	**稍蜷**	**清晰**	否
10	**硬挺**	**清晰**	否

此时将“根蒂”为稍蜷的结点标记为叶结点，并将其类别设定为该结点所含样本最多的类别，即“好瓜”（结点包含三个样本，其中两个为好瓜，一个不是好瓜）。

(IV)对于**递归返回情形(2)**的“D中样本在A上取值相同”，如下表所示（非表 4.1 数据，纯为解释该递归返回情形而编造的）：

编号	**根蒂**	**纹理**	脐部	触感	好瓜
1	**蜷缩**	**清晰**	凹陷	硬滑	是
2	**蜷缩**	**清晰**	凹陷	硬滑	是
3	**蜷缩**	**清晰**	凹陷	硬滑	是
4	**蜷缩**	**清晰**	凹陷	硬滑	否
5	**蜷缩**	**清晰**	凹陷	硬滑	否

此时已根据“纹理”和“根蒂”两个属性进行划分，样本子集中仍同时包含好瓜和坏瓜，仍有“脐部”和“触感”两个属性可以使用，但当前子集中的样本在剩余的“脐部”和“触感”两个属性上的取值均相等（均为凹陷和硬滑），因此无法进行划分，因此将该结点标记为叶结点，并将其类别设定为该结点所含样本最多的类别，即“好瓜”（结点包含五个样本，其中三个为好瓜，两个不是好瓜）。

总结：(a)划分期望得到的结果是各子集只包含同一类别的样本（暂不考虑 4.3 节讨论的过拟合问题）；(b)使用过的属性在后续各子集继续划分时不能再使用，因为各子集内的样本在该属性上取值相同（仅针对离散属性，若是连续属性，则可以重复使用，如表 4.3 中的“密度”和“含糖率”）；(c)对于递归返回情形(1)，虽有可用属性但样本类别却已相同，可以理解为这些属性对当前划分不起作用；(d)对于**递归返回情形(2)**，当前子集样本在所有属性上取值相同，但样本类别却不同，可以理解为**样本集含的标记含有噪声**（特征相同的样本，被标记的类别却不相同，应该是标记数据时弄错了），**或应增加新特征**（即当前的几种特征属性不足以描述样本）；(e)决策树并不一定是二叉树，分枝个数等于属性取值个数。

4.2 划分选择

本节介绍的三种划分选择方法，即信息增益、增益率、基尼指数分别对应著名的 ID3、C4.5 和 CART 三种决策树学习算法。

1、式(4.1)的解释

该式即为信息论中的信息熵定义公式，以下仅解释其最大值和最小值。

根据信息论的知识，当各类样本所占比例相同时信息熵$\mathrm{Ent}(D)$最大：

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

当样本集合D中只含某一类样本时信息熵$\mathrm{Ent}(D)$最小：（若$p=0$，则$p\log_2 p=0$）

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k=-1\cdot\log_2 1-0\cdot\log_2 0-\ldots-0\cdot\log_2 0=0$$

由以上两种极限，可以更好的理解式(4.1)下面一行的话：“$\mathrm{Ent}(D)$的值越小，则D的纯度越高”，若D中只包含一类样本，则$\mathrm{Ent}(D)=0$，此时纯度最高。

2、式(4.2)的解释

我们希望经过依据某属性划分后，划分所得各子集的纯度越高越好，即各子集的信息熵$\mathrm{Ent}(D^v)$越小越好（若是各子集样本的类别相同，则所有子集的$\mathrm{Ent}(D^v)=0$），也就是信息增益$\mathrm{Gain}(D,a)$越大越好。

划分所得各子集包含的样本个数有多有少，故以其所占比例$|D^v|\,/\,|D|$为权重进行加权。

3、式(4.4)的解释

为了理解该式的“固有值”的概念，可以将式(4.4)与式(4.1)比较一下。

式(4.1)可重写为：

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

其中$p_k=\left|D^k\right|\,/\,|D|$，为第$k$类样本所占的比例。

与式(4.4)的表达式作一下对比：

$$\mathrm{IV}(a)=-\sum_{v=1}^{V}\frac{|\mathcal{D}^v|}{|\mathcal{D}|}\log_2\frac{|\mathcal{D}^v|}{|\mathcal{D}|}$$

其中$p_v=|D^v|\,/\,|D|$，为属性a等于第v个属性值的样本所占的比例。

即式(4.1)是按样本集合D的类别标记计算的信息熵，而式(4.4)是按属性a的取值计算的信息熵（相当于将属性a作为类别）。

考虑一种特殊情况：若属性a的取值与类别标记一一对应，则$\mathrm{Ent}(D)=\mathrm{IV}(a)$，式(4.2)中的第二项等于零，此时式(4.3)的增益率等于 1。

4、式(4.5)的推导

假设数据集D中的样例标记种类共有三类，每类样本所占比例分别为p_1,p_2,p_3。现从数

据集D中随机抽取两个样本，如下表所示

	p_1	p_2	p_3
p_1	p_1p_1	p_1p_2	p_1p_3
p_2	p_2p_1	p_2p_2	p_2p_3
p_3	p_3p_1	p_3p_2	p_3p_3

两个样本类别标记正好一致的概率为（即对角线元素之和）：

$$p_1p_1+p_2p_2+p_3p_3=\sum_{k=1}^{|\mathcal{Y}|=3}p_k^2$$

两个样本类别标记不一致的概率为（即“基尼值”）：

$$\mathrm{Gini}(D)=p_1p_2+p_1p_3+p_2p_1+p_2p_3+p_3p_1+p_3p_2=\sum_{k=1}^{|\mathcal{Y}|=3}\sum_{k'\neq k}p_kp_{k'}$$

很容易证明（提公因式即可，注意$p_1+p_2+p_3=1$）：

$$\begin{aligned}&(p_1p_1+p_2p_2+p_3p_3)+(p_1p_2+p_1p_3+p_2p_1+p_2p_3+p_3p_1+p_3p_2)\\=&(p_1p_1+p_1p_2+p_1p_3)+(p_2p_1+p_2p_2+p_2p_3)+(p_3p_1+p_3p_2+p_3p_3)\\=&p_1(p_1+p_2+p_3)+p_2(p_1+p_2+p_3)+p_3(p_1+p_2+p_3)\\=&p_1+p_2+p_3=1\end{aligned}$$

即得式(4.5)：

$$\mathrm{Gini}(D)=\sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k}p_kp_{k'}=1-\sum_{k=1}^{|\mathcal{Y}|}p_k^2$$

从一个数据集D中任取两个样本，类别标记一致的概率越大表示其纯度越高（即大部分样本属于同一类）；类别标记不一致的概率（即基尼值）越大表示纯度越差；因此选择使各划分子集的基尼值尽可能小的属性作为划分属性，基尼指数即各子集基尼值的加权平均。

4.3 剪枝处理

本节内容通俗易懂，几乎不需要什么注解。

以下仅结合图 4.5 继续讨论一下图 4.1 中的递归返回条件。图 4.5 与图 4.4 均是基于信息增益生成的决策树，不同在于图 4.4 基于表 4.1，而图 4.5 基于表 4.2 的训练集。

结点③包含训练集“脐部”为稍凹的样本（编号 6、7、15、17），当根据“根蒂”再次进行划分时不含有“根蒂”为硬挺的样本（递归返回情形(3)），而恰巧四个样本（编号 6、7、15、17）含两个好瓜和两个坏瓜，因此叶结点硬挺的类别**随机**从类别好瓜和坏瓜中选择其一；

结点⑤包含训练集“脐部”为稍凹且“根蒂”为稍蜷的样本（编号 6、7、15），当根据“色泽”再次进行划分时不含有“色泽”为浅白的样本（递归返回情形(3)），因此叶结点浅白类别标记为好瓜（编号 6、7、15 样本中，前两个为好瓜，最后一个为坏瓜）；

结点⑥包含训练集“脐部”为稍凹、“根蒂”为稍蜷、“色泽”为乌黑的样本（编号 7、15），当根据“纹理”再次进行划分时不含有“纹理”为模糊的样本（递归返回情形(3)），而恰巧两个样本（编号 7、15）含好瓜和坏瓜各一个，因此叶结点模糊的类别**随机**从类别好瓜和坏瓜中选择其一；

图 4.5 两次**随机**选择均选为好瓜，实际上表示了一种归纳偏好（1.4 节）。

4.4 连续与缺失值

有些分类器只能使用离散属性，当遇到连续属性时则需要特殊处理。专门有人研究“属性离散化”技术，有兴趣可以参考[Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 734-750.]。若结合第 11 章 11.2 节至 11.4 节分别介绍的“过滤式(filter)”算法、“包裹式(wrapper)”算法、“嵌入式(embedding)”算法的概念，若先使用某个属性离散化算法对连续属性离散化后再调用 C4.5 决策树生成算法，则是一种过滤式算法；若如 4.4.1 节所述，则应该属于嵌入式算法（并没有以学习器的预测结果准确率为评价标准，而是与决策树生成过程融为一体，因此不应该划入包裹式算法）。

类似地，有些分类器不能使用含有缺失值的样本，需要进行预处理。常用的缺失值填充方法是：对于连续属性，采用已该属性的均值进行填充；对于离散属性，采用属性值个数最多的样本进行填充；这实际上假设了数据集样本是基于独立同分布采样得到的。

特别地，一般缺失值仅指特征属性含有缺失，若类别含有缺失，一般会直接抛弃该样本。当然，也可以根据 11.6 节的式(11.24)，在低秩假设下对数据集缺失值进行填充。

1、式(4.8)的解释

该式就是简单遍历所有可能的候选划分点集合T_a，选出使信息增益最大的划分点。

2、式(4.12)的解释

该式括号内与式(4.2)基本一样，区别在于式(4.2)中的$|D^v|/|D|$改为式(4.11)的$\tilde{r}_v$，在根据式(4.1)计算信息熵时第k类样本所占的比例改为式(4.10)的$\tilde{p}_k$；所有计算结束后再乘以式(4.9)的ρ。

有关式(4.9) (4.10) (4.11)中的权重w_x，初始化为 1；以图 4.9 为例，在根据“纹理”进行划分时，除编号为 8,10 的两个样本在此属性缺失之外，其余样本根据自身在该属性上的取值分别划入稍糊、清晰、模糊三个子集，而编号为 8,10 的两个样本则要按比例同时划入三个子集；具体来说，稍糊子集包含样本 7, 9, 13, 14, 17 共 5 个样本，清晰子集包含样本 1, 2, 3, 4, 5, 6, 15 共 7 个样本，模糊子集包含样本 10, 11, 16 共 3 个样本，总共 15 个在该属性不含缺失值的样本，而此时各样本的权重w_x为初始化 1，因此编号为 8,10 的两个样本分到稍糊、清晰、模糊三个子集的权重分别为$\frac{5}{15}$, $\frac{7}{15}$和$\frac{3}{15}$。

4.5 多变量决策树

本节通俗易懂，只要理解了图 4.11 的“轴平行”边界的来历，一切就都明白了。

1、图 4.10 的解释

只想用该图强调一下，离散属性不可以重复使用，但连续属性是可以重复使用的。

2、图 4.11 的解释

对应着图 4.10 的决策树，来看一看图 4.11 的边界是怎么画出来的：

在下图中，黑色正斜杠(/)阴影部分表示已确定标记为坏瓜的样例，红色反斜杠(\)阴影部分表示已确定标记为好瓜的样例，空白部分表示需要进一步划分的样例。

第一次划分条件是“含糖率≤0.126？”，满足此条件的样例直接被标记为坏瓜（如图(a)黑色阴影部分所示），而不满足此条件的样例还需要进一步划分（如图(a)空白部分所示）；

在第一次划分的基础上对图(a)空白部分继续进行划分，第二次划分条件是“密度≤0.381？”，满足此条件的样例直接被标记为坏瓜（如图(b)新增黑色阴影部分所示），而不满足此条件的样例还需要进一步划分（如图(b)空白部分所示）；

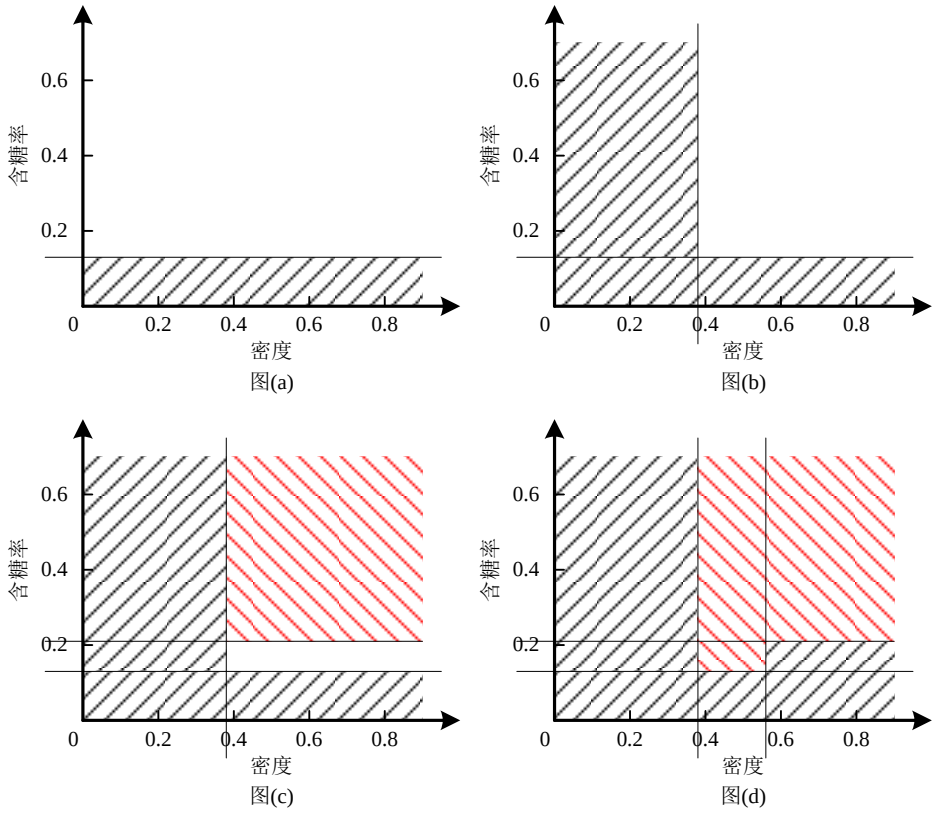
在第二次划分的基础上对图(b)空白部分继续进行划分，第三次划分条件是“含糖率≤0.205？”，不满足此条件的样例直接标记为好瓜（如图(c)新增红色阴影部分所示），而满足此条件的样例还需进一步划分（如图(c)空白部分所示）；

在第三次划分的基础上对图(c)空白部分继续进行划分，第四次划分的条件是“密度≤0.560？”，满足此条件的样例直接标记为好瓜（如图(d)新增红色阴影部分所示），而不满足此条件的样例直接标记为坏瓜（如图(d)新增黑色阴影部分所示）。

经过四次划分已无空白部分，表示决策树生成完毕。

从图(d)中可以清晰地看出好瓜与坏瓜的分类边界。

如书中所注，分类边界的每一段都是与坐标轴平行的，其实更严格地说应该是分类边界的每一段都是与坐标轴垂直的，话虽一样，但道理不同。如图 4.11 所示，由“含糖率≤0.126？”确定的第一段边界实际是函数“含糖率=0.126”这个常函数的图像，而此类常函数的图像是坐标轴“含糖率”本身垂直的。



3、图 4.14 的解释

在图 4.14 中共有两条分类边界：

左边一条为一次函数“$-0.800\times$密度$-0.044\times$含糖率$=-0.313$”；

右边一条为一次函数“$-0.365\times$密度$-0.366\times$含糖率$=-0.158$”；

如果仍不明白，可将含糖率替换为y，将密度替换为x，得到一次函数的表达式就清楚了。

关于判断直线的哪一侧大于零，哪一侧小于零，简单来说法向量方向是大于零的区域。例如直线“$-0.800\times$密度$-0.044\times$含糖率$+0.313=0$”，法向量为$(-0.800,-0.044)$，指向第三象限，为“$-0.800\times$密度$-0.044\times$含糖率$+0.313>0$”区域，因此图 4.14 左边直线的左侧为“坏瓜”，这是因为图 4.13 显示，当“$-0.800\times$密度$-0.044\times$含糖率$+0.313\leqslant 0$”不成立时（即大于零）为坏瓜。

注意，无论是图 4.11 还是图 4.14，所要划分的区域仅是第一象限，因为两个属性“含糖率”和“密度”取值范围均为正值。

4.6 本章小节

本章作为西瓜书“介绍经典而常用的机器学习方法”的开篇，通篇以“西瓜”为例进行讲解，通俗易懂。个人感觉最可能会让人产生困惑的知识点是决策树三种递归返回情形（参见 4.1 节的详细解释）、缺失值处理的具体细节（参见式(4.12)的解释，具体还得自己琢磨一下 4.4.2 节的例子）、决策树分类边界的轴平行特点（参见图 4.11 的解释）。

值得一提的是，决策树不仅可以用于分类任务，也可以用于回归任务，常见的做法是用叶结点所有样本的平均值作为预测回归值。

其它常见的与决策树密切相关模型有：随机森林(Random Forest, RF)和梯度提升树(Gradient Boosted Decision Trees, GBDT)，但这两个模型实际上属于集成学习(参见第 8 章)的范畴；其中随机森林参见西瓜书 8.3.2 节，GBDT 参见博客园刘建平 Pinard 所写的《梯度提升树(GBDT)原理小结》(https://www.cnblogs.com/pinard/p/6140514.html)。