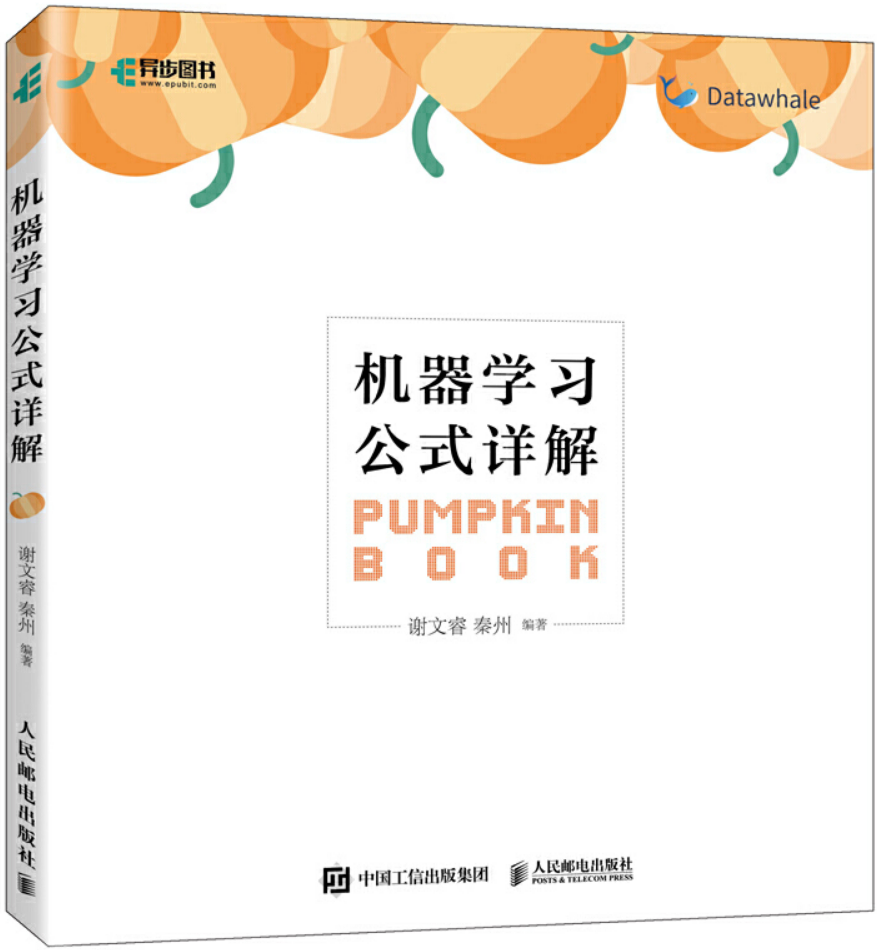

《机器学习公式详解》（南瓜书）

第4章 决策树

本节主讲：谢文睿

本节大纲

西瓜书对应章节：4.1、4.2

1. 算法原理
2. ID3决策树
3. C4.5决策树
4. CART决策树

算法原理

从逻辑角度，一堆if else语句的组合
从几何角度，根据某种准则划分特征空间

最终目的：将样本越分越“纯”

ID3决策树

自信息：

$$I(X) = -\log_b p(x)$$

当$b = 2$时单位为bit，当$b = e$时单位为nat

信息熵（自信息的期望）：度量随机变量X的不确定性，信息熵越大越不确定

$$H(X) = E[I(X)] = -\sum_x p(x)\log_b p(x) \quad （此处以离散型为例）$$

计算信息熵时约定：若$p(x) = 0$，则$p(x)\log_b p(x) = 0$。当X的某个取值的概率为1时信息熵最小（最确定），其值为0；当X的各个取值的概率均等时信息熵最大（最不确定），其值为$\log_b |X|$，其中$|X|$表示X可能取值的个数。详细证明过程参见《机器学习公式详解》（南瓜书）式（4.1）的解析~

ID3决策树

将样本类别标记y视作随机变量，各个类别在样本集合D中的占比$p_k(k=1,2,...,|\mathcal{Y}|)$视作各个类别取值的概率，则样本集合$D$（随机变量$y$）的信息熵（底数$b$取2）为

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

此时的信息熵所代表的“不确定性”可以转换理解为集合内样本的“纯度”（举个栗子）

ID3决策树

条件熵（Y的信息熵关于概率分布X的期望）：在已知X后Y的不确定性

$$H(Y|X)=\sum_{x}p(x)H(Y|X=x)$$

从单个属性（特征）a的角度来看，假设其可能取值为$\{a^1,a^2,...,a^V\}$，D^v表示属性a取值为$a^v\in\{a^1,a^2,...,a^V\}$的样本集合，$\frac{|D^v|}{D}$表示占比，那么在已知属性$a$的取值后，样本集合$D$的条件熵为

$$\sum_{v=1}^{V}\frac{|D^v|}{|D|}\operatorname{Ent}\left(D^v\right)$$

ID3决策树

信息增益：在已知属性（特征）a的取值后y的不确定性减少的量，也即纯度的提升

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}\left(D^v\right)$$

ID3决策树：以信息增益为准则来选择划分属性的决策树

$$a_*=\underset{a\in A}{\arg\max}\,\mathrm{Gain}(D,a)$$

C4.5决策树

信息增益准则对可能取值数目较多的属性有所偏好（例如“编号”这个较为极端的例子，不过其本质原因不是取值数目过多，而是每个取值里面所包含的样本量太少），为减少这种偏好可能带来的不利影响，C4.5决策树选择使用“增益率”代替“信息增益”，增益率定义为

$$\text{Gain\_ratio}(D,a)=\frac{\text{Gain}(D,a)}{\text{IV}(a)}$$

其中

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

称为属性a的“固有值”，a的可能取值个数V越大，通常其固有值$\text{IV}(a)$也越大。但是，增益率对可能取值数目较少的属性有所偏好。

C4.5决策树

因此，C4.5决策树并未完全使用“增益率”代替“信息增益”，而是采用一种启发式的方法：先选出信息增益高于平均水平的属性，然后再从中选择增益率最高的。

CART决策树

基尼值：从样本集合D中随机抽取两个样本，其类别标记不一致的概率。因此，基尼值越小，碰到异类的概率就越小，纯度自然就越高。

$$
\begin{aligned}
\operatorname{Gini}(D) &= \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} \\
&= \sum_{k=1}^{|\mathcal{Y}|} p_k \left(1 - p_k\right) \\
&= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2
\end{aligned}
$$

CART决策树

属性a的基尼指数（类比信息熵和条件熵）：

$$\text{Gini\_index}(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\text{Gini}\left(D^v\right)$$

CART决策树：选择基尼指数最小的属性作为最优划分属性

$$a_*=\underset{a\in A}{\arg\min}\ \text{Gini\_index}\left(D,a\right)$$

那具体的划分点呢?

CART决策树

CART决策树的实际构造算法如下：

- 首先，对每个属性a的每个可能取值v，将数据集D分为$a = v$和$a \neq v$两部分来计算基尼指数，即

$$\text{Gini\_index}(D, a) = \frac{|D^{a=v}|}{|D|}\text{Gini}(D^{a=v}) + \frac{|D^{a\neq v}|}{|D|}\text{Gini}(D^{a\neq v})$$

- 然后，选择基尼指数最小的属性及其对应取值作为最优划分属性和最优划分点；
- 最后，重复以上两步，直至满足停止条件。

具体例子参见《机器学习公式详解》（南瓜书）式（4.6）的解析~

预告

下一节：神经网络

西瓜书对应章节：5.1、5.2、5.3

结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 关注公众号【Datawhale】，发送【南瓜书】三个字获取机器人“小豚”的微信二维码
2. 添加“小豚”为微信好友，然后对“小豚”发送【南瓜书】三个字即可自动邀请进群、