221900180 田永铭 AI导论 决策树 Information Gain 非负的证明

证明：由Gain定义：$Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$.

由Ent定义：$Ent(D)=-\sum_{k=1}^{n}p_k\log_2 p_k=-\sum_{k=1}^{n}\frac{|D^k|}{|D|}\log_2\frac{|D^k|}{|D|}$.

其中：n为类别个数.

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

$$=-\sum_{k=1}^{n}\frac{|D^k|}{|D|}\log_2\frac{|D^k|}{|D|}+\sum_{v=1}^{V}\frac{|D^v|}{|D|}\sum_{k=1}^{n}\frac{|D^{vk}|}{|D^v|}\log_2\frac{|D^{vk}|}{|D^v|}$$

（这两项均表示权重，且形式为乘积，因此可交换，仍表示两权重之积.）

换字母顺序 $=-\sum_{k=1}^{n}\frac{|D^k|}{|D|}\log_2\frac{|D^k|}{|D|}+\sum_{k=1}^{n}\frac{|D^k|}{|D|}\sum_{v=1}^{V}\frac{|D^{vk}|}{|D^k|}\log_2\frac{|D^{vk}|}{|D^v|}$

提公因式 $=\sum_{k=1}^{n}\frac{|D^k|}{|D|}\left(\sum_{v=1}^{V}\frac{|D^{vk}|}{|D^k|}\log_2\frac{|D^{vk}|}{|D^k|}-\log_2\frac{|D^k|}{|D|}\right)$.

故欲证 $Gain(D,a)\geq 0$，

只需证：$\sum_{v=1}^{V}\frac{|D^{vk}|}{|D^k|}\log_2\frac{|D^{vk}|}{|D^k|}\geq\log_2\frac{|D^k|}{|D|}$ for $\forall k\in\{1,2,\dots n\}$ 成立.

~~注意到：$|D|$与$|D^k|$均表示样本总量，即$|D|=|D^k|$~~

~~∴即证 $\sum_{v=1}^{V}\frac{|D^{vk}|}{|D|}\log_2\frac{|D^{vk}|}{|D|}\geq\log_2\frac{|D^k|}{|D|}$ for $\forall k\in\{1,2,\dots n\}$ 成立.~~

~~由于 $\log_2(x)$ 为凹函数(二阶导<0)，∴ $\forall\theta_1,\theta_2\dots\theta_k\geq 0$, subject to $\theta_1+\theta_2+\dots+\theta_k=1$,~~

~~有 $f(\theta_1x_1+\theta_2x_2+\dots+\theta_kx_k)\geq\theta_1f(x_1)+\theta_2f(x_2)+\dots+\theta_kf(x_k)$.~~

~~∴取 $\theta_1=\frac{|D^{vk}|}{|D|}$, $\theta_2=|D^{k}|$~~

~~取 $\theta_i=\frac{|D^{ik}|}{|D|}$ $(i=1,2\dots V)$，取~~

即证 $\sum_{v=1}^{V}\frac{|D^{vk}|}{|D^k|}\left(-\log_2\frac{|D^k|}{|D^{vk}|}\right)\geq\log_2\frac{|D^k|}{|D|}$ for $\forall k\in\{1,2,3\dots n\}$ 成立.

由 $f(x)=-\log_2(x)$ 为凸函数，∴ $\forall\theta_1,\theta_2\dots\theta_t\geq 0$, subject to $\theta_1+\theta_2+\dots+\theta_t=1$,

有 $f(\theta_1x_1+\theta_2x_2+\dots+\theta_tx_t)\leq\theta_1f(x_1)+\theta_2f(x_2)+\dots+\theta_tf(x_t)$. （有 $\theta_1+\theta_2+\dots+\theta_V=1$）

取 $\theta_i=\frac{|D^{ik}|}{|D^k|}$ $(i=1,2,3\dots V)$，取 $x_i=\frac{|D^k|}{|D^{ik}|}$ $(i=1,2,3\dots V)$.

∴左式 $\geq-\log_2(\underbrace{1+1+\dots+1}_{V\text{个}})=\log_2\frac{1}{V}$.

∴只需证 ~~$V|D|\leq|D^k|$~~ ~~$V|D^k|\leq|D|$~~ $\forall k$ $V|D^k|\leq|D|$ 即 $V\leq\frac{|D|}{|D^k|}$

~~考虑样本实际意义，当且仅当类别(输出种类)个数为1时，$\frac{|D|}{|D^k|}=1$.~~

只需证 $V\leq\left(\frac{|D|}{|D^k|}\right)_{min}$.

上述证明不太能进行下去，不如为利用条件概率定理：

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

$$= -\sum_{k=1}^{n} p_k \log_2 p_k + \sum_{x=1}^{n} p_x \sum_{k=1}^{n} p(k|x) \log_2 (k|x)$$

简化记作

$$= -\sum_{k} p_k \log_2 p_k + \sum_{x} p_x \sum_{k} p(k|x) \log_2 (k|x)$$

由条件概率性质

$$= -\sum_{k} \sum_{x} p(x,k) \log_2 p_k + \sum_{k} \sum_{x} p(k,x) \log_2 (k|x)$$

$$= \sum_{x} \sum_{k} p(x,k) \log_2 \frac{p(k|x)}{p(k)}$$

$$= -\sum_{x} \sum_{k} p(x,k) \log_2 \frac{p(x)p(k)}{p(x,k)}$$

由 $f(x)=\log_2 x$ 凹

$$\geq -\log_2 p(x,k) \frac{p(x)p(k)}{p(x,k)} = -\log_2 p(x)p(k)$$

$\because p(x), p(k) \leq 1, \therefore -\log_2 p(x)p(k) \leq 0$，$\therefore$ 原式 ≥ 0. 得证.

①方法也可以用：

原因：之前写错一个字母。

欲证：$\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D^v|} \geq \log_2 \frac{|D|}{|D|}$.

由 $-\log x$ 为凸函数得：（凸函数性质）

$\sum_{v=1}^{V} \frac{|D^v|}{|D|} \left(-\log_2 \frac{|D^v|}{|D^v|}\right)$ ——> 多校正，目标为1.

$\geq -\log_2 \left(\frac{|D^1|}{|D|} + \frac{|D^2|}{|D|} + \cdots + \frac{|D^v|}{|D|} \right) = \log_2 \frac{|D|}{|D|}$.

二得证。