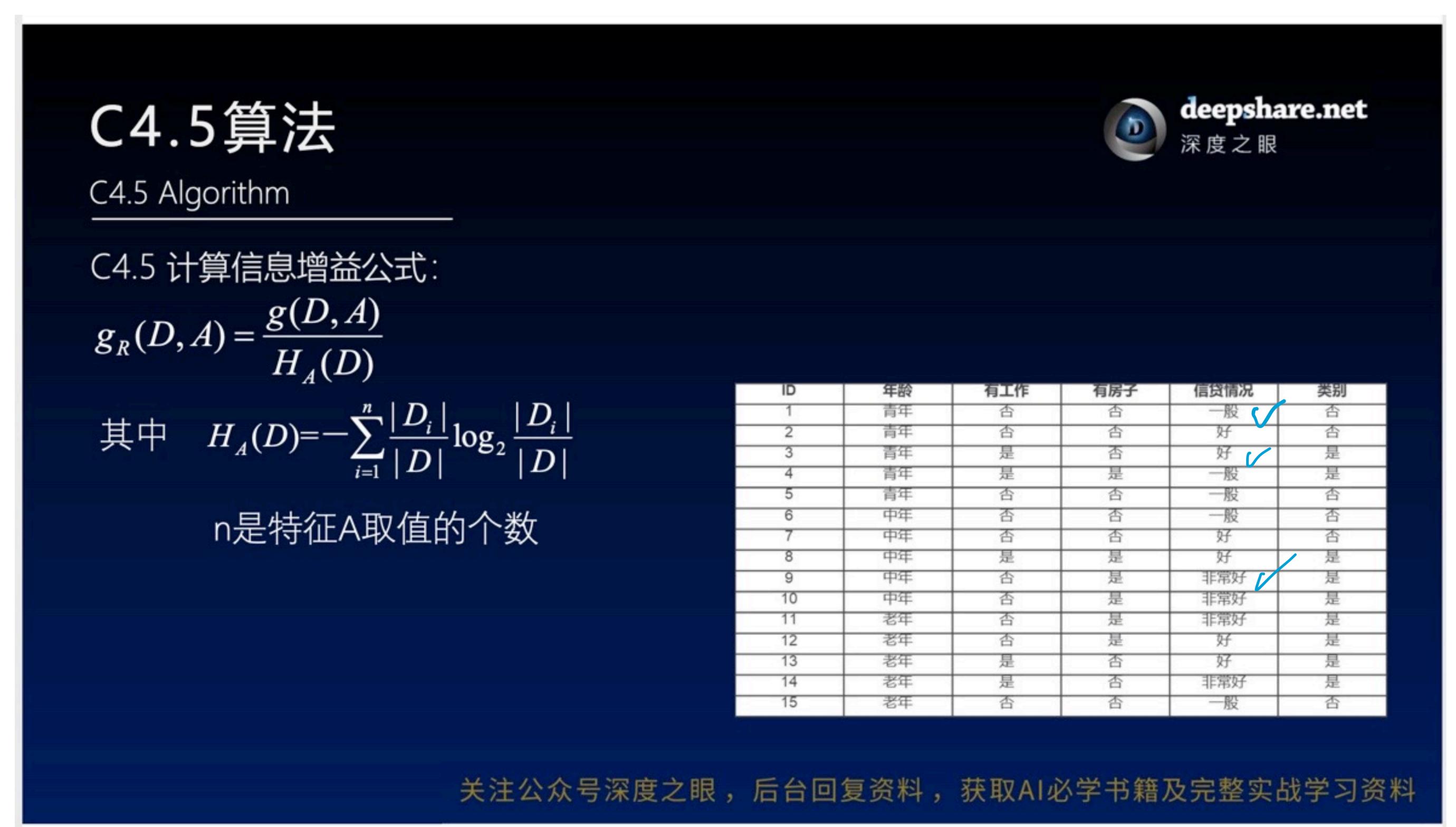
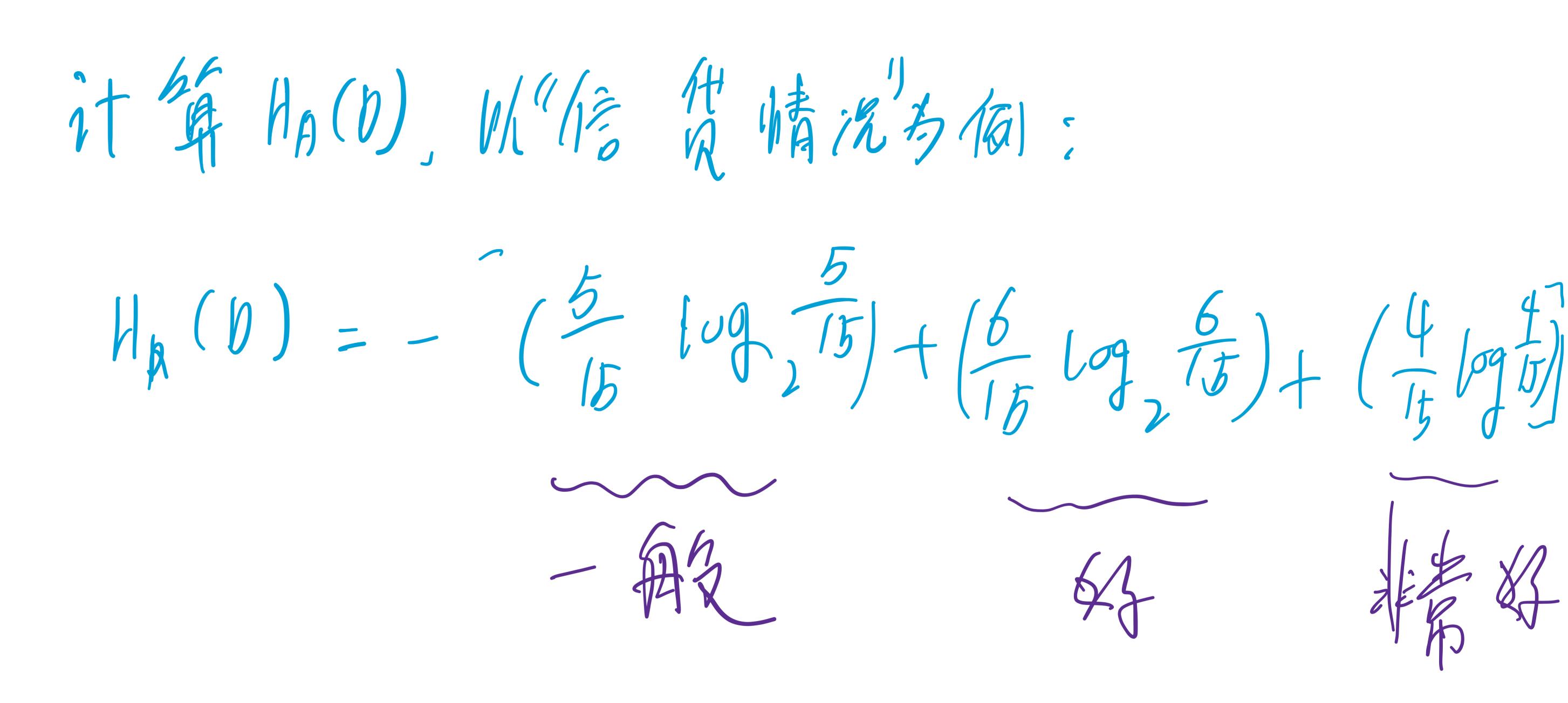
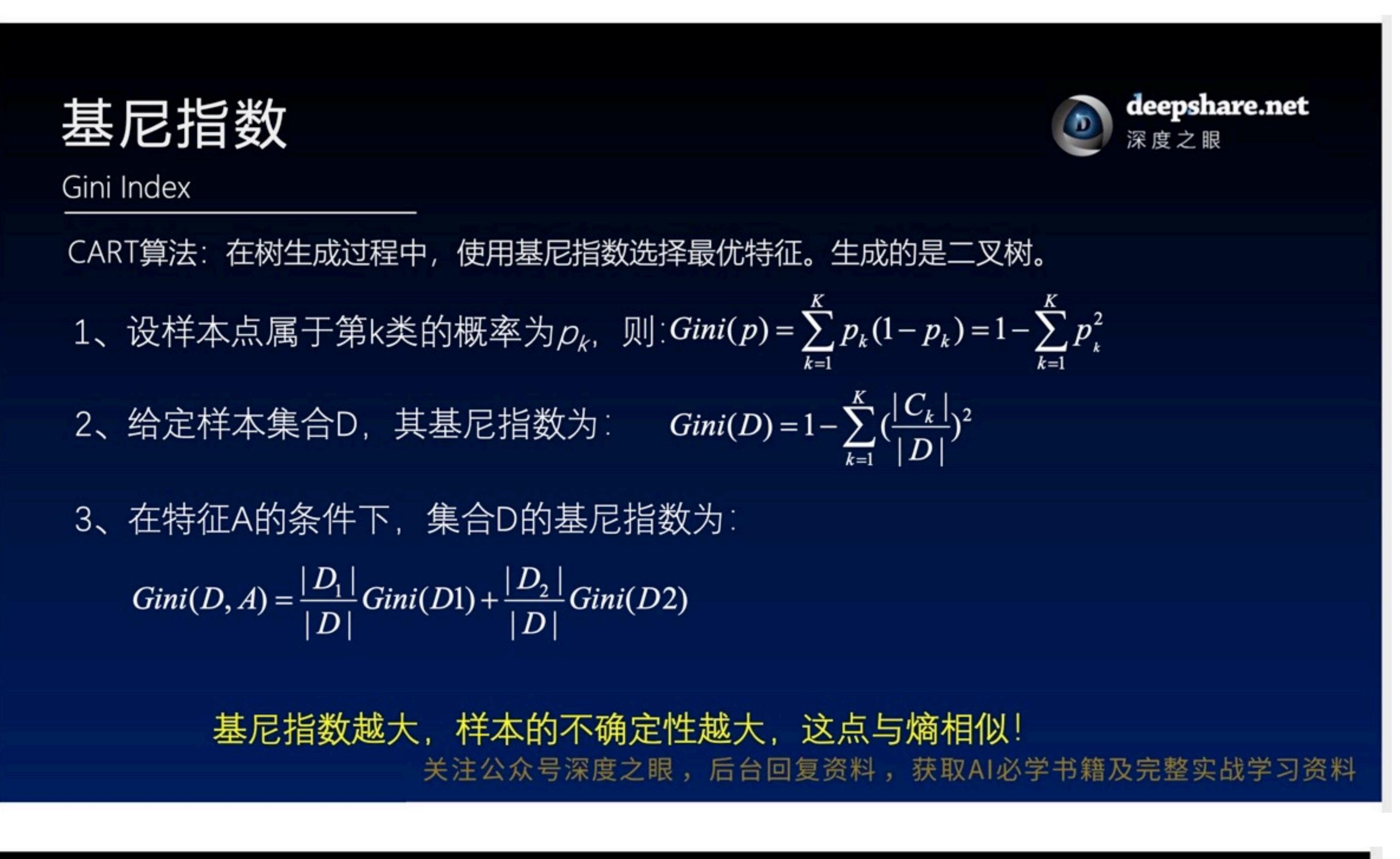


在计算特征的每个取值下的熵的时候,需要估计每个类别k的概率!显然我们是直接用频率去近似的概率。然而根据大数定律,只有当样本数足够多的时候,频率才可以准确的近似概率。样本数越少,对概率的估计结果的方差就会越大(想象一下做抛硬币实验来近似正面向上的概率,如果只抛两次,那么得到的正面向上的概率可能会非常离谱。而如果抛1万次,不论何时何地几乎总能得到近似0.5的概率)而方差大会导致什么结果呢?显然就会导致该取值下的类别错估计为非均匀分布呀!而非均匀分布,不就是说导致该取值下的熵更小了。所以说,ID3决策树,或者说信息增益,或者说条件熵,并不是说一定会偏向于取值多的特征,而是数据集的不充足以及客观存在的大数定律导致取值多的特征在计算条件熵时容易估计出偏小的条件熵。





为什么信息指数 HA(D)可以起到能调的作用?



基尼指数 $Gini(D)=1-\sum\limits_{k=1}^{K}(\frac{ C_k }{ D })^2$ $Gini(D,A)=\frac{ D_1 }{ D }Gini(D1)+\frac{ D_2 }{ D }Gini(D2)$ deepshare.net $\Re g \geq \mathbb{R}$						
Gini Index						
$Gini(D, A_1=1) = \frac{5}{15}(1 - ((\frac{2}{5})^2 + (\frac{3}{5})^2)) + \frac{10}{15}(1 - ((\frac{3}{10})^2 + (\frac{7}{10})^2)) = 0.44$						
$Gini(D, A_1 = 2) = 0.48$	ID	年齡	有工作	有房子	信贷情况	类别
	1	青年	否	否	一般	否
$Gini(D, A_1 = 3) = 0.44$	2	青年	否	否	好	否
	3	青年	是	否	好	是
$Gini(D, A_2 = 1) = 0.32$	4	青年	是	是	一般	是し
	5	青年	否	否	一般	否
$Gini(D, A_3 = 1) = 0.27$	6	中年	否	否	一般	否
0,113	7	中年	否	否	好	- 否
$Gini(D, A_4 = 1) = 0.36$	9	中年	是不	是	好 非常好	是
, , , , , , , , , , , , , , , , , , , ,	10	中年	否	是	非常好	是是
$Gini(D, A_4 = 2) = 0.47$	11	老年	否	是	非常好	是
	12	老年	否	是	好	是
$Gini(D, A_4 = 3) = 0.32$	13	老年	是	否	好	是
	73 min 14 mm	一人老年一一	是大田立	A 1 N 西 +> かか	非常好一	224 一是如此
因此选A ₃ 为最优特征,A ₃ =1为最优切分点	不	老年	了小十 ,否 3人 中人	不	及元胺夫山	() 一层八十

为什么基尼指数越大,混乱维度越大?