

③ Entropy

$$\odot H(\mathcal{E}) = - \sum_{i=1}^K p_{ik} \log_2 p_{ik}$$

binary split:

$$H(\mathcal{E}) = -P \log_2 P - (1-P) \log_2 (1-P) ; 0 < P < 1$$

Need to prove: $-P \log P - (1-P) \log(1-P) \leq 1$?
(assume log base 2)

Balanced probabilities = High entropy (lowest information gain)
Max entropy at $P = 1/2$

$$-(1/2) \log(1/2) - (1/2) \log(1/2) \stackrel{?}{\leq} 1$$

$$\text{Also, } \rightarrow \lim_{P \rightarrow 1/2^{\pm}} (-P \log P - (1-P) \log(1-P)) = 1 \leq 1 \quad \checkmark$$

\therefore This shows that given a binary data split, the MAXIMUM entropy is 1

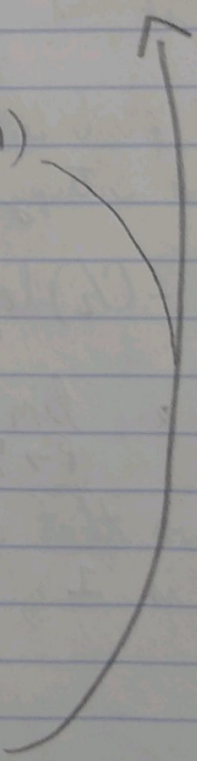
(3)(b) $n = \#$ of branches

n	$\max(H)$
2	1
3	1.58
4	2
8	3
16	4

Same as $\log(n)$

The maximum entropy for multiway branching, given n branches, is $\log(n)$.

Need to prove:

$$\begin{aligned} & - \sum_{i=1}^n \left(\frac{1}{n} \right) \log_2 \left(\frac{1}{n} \right) \stackrel{?}{=} \log_2(n) \\ & = - \sum_{i=1}^n \frac{1}{n} \left(\log_2(1) - \log_2(n) \right) \\ & = + \sum_{i=1}^n + \frac{1}{n} \log_2(n) \\ & = n \left(\frac{1}{n} \log_2(n) \right) \\ & = \log_2(n) \quad \checkmark = \log_2(n). \end{aligned}$$


$$\textcircled{4} \text{ Gain}(q, V) = b(q) - \sum_{i=1}^{|V|} \frac{N_i}{N_q} b(i)$$

Show that maximizing Gain minimizes impurity (b) measure over the $|V|$ children.

$b(q)$ is always positive ^{or 0,} so to maximize $\text{Gain}()$, we must minimize $\sum_{i=1}^{|V|} \frac{N_i}{N_q} b(i)$, which is subtracted from $b(q)$.

$\sum_{i=1}^{|V|} \frac{N_i}{N_q} b(i)$: In order to minimize, $\left(\frac{N_i}{N_q}\right)$ and $(b(i))$ need to be minimized.

$\rightarrow \frac{N_i}{N_q} \rightarrow N_q$ is constant | A feature with N_i values closer to 0 or N_q should be selected for, because this results more homogenous nodes leading to LOWER impurity.

$\rightarrow b(i) \rightarrow$ High purity = Low impurity = low $\sum_{i=1}^{|V|} \frac{N_i}{N_q} b(i)$
so maximizing purity, and therefore Gain, can only occur through Minimizing the Impurity Measure.

$$\begin{aligned}
 \textcircled{5} \quad Gini(q) &= \sum_{k \neq k'} p_k p_{k'} \quad ; \quad N \\
 &= \sum_{k=1}^M \left(p_k \times \underbrace{\sum_{k' \neq k} p_{k'}}_{\rightarrow (1-p_k)} \right)
 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{k=1}^M (p_k \times (1-p_k)) \\
 &= \sum_{k=1}^M p_k (1-p_k) \quad \checkmark
 \end{aligned}$$

$$\therefore \sum_{k=1}^M p_k (1-p_k) = \sum_{k \neq k'} p_k p_{k'} \quad \text{where } M > 2$$