③ Entropy

$$H(q) = -\sum_{i=1}^{K} P_{qK} \log_2 P_{qK}$$

binary split:

$$H(q) = -P\log_2 P - (1-P)\log_2(1-P) \quad ; \quad 0 < P < 1$$

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 $P = 1/2$

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

Also, $\Rightarrow$ $\lim_{P \to 1/2^{\pm}} (-P\log P - (1-P)\log(1-P)) = 1 \leq 1$ ✓

$\therefore$ 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:

$$-\sum_{i=1}^{n} \left(\frac{1}{n}\right)\log_2\left(\frac{1}{n}\right) \stackrel{?}{=} \log_2(n)$$

$$= -\sum_{i=1}^{n} \frac{1}{n}\left(\cancel{\log_2(1)}^{\,0} - \log_2(n)\right)$$

$$= +\sum_{i=1}^{n} +\frac{1}{n}\log_2(n)$$

$$= n\left(\frac{1}{n}\log(n)\right)$$

$$= \log_2(n) \stackrel{\checkmark}{=} \log_2(n).$$

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

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

$\iota(q)$ is always positive or 0, so to maximize Gain(),

we must minimize $\sum_{i=1}^{|V|} \frac{N_i}{N_q} \iota(i)$, which is subtracted

from $\iota(q)$.

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

→ $\frac{N_i}{N_q}$ → 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.

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

(5) $$\text{Gini}(q) = \sum_{k \neq k'} p_k p_{k'}$$

$$= \sum_{k=1}^{M} \Big(p_k \times \underbrace{\sum_{k' \neq k} p_{k'}}_{(1-p_k)} \Big)$$

$$= \sum_{k=1}^{M} \big(p_k \times (1 - p_k) \big)$$

$$= \sum_{k=1}^{M} p_k (1 - p_k) \quad \checkmark$$

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