Choosing attributes

Plurality value → return the most frequent value.

Entropy of dataset → $H(Goal) = B\left(\dfrac{p}{p+n}\right)$

feature → $\sum\limits_{k} P(v_k) \log_2 \dfrac{1}{P(v_k)}$

$$= -\sum\limits_{k} P(v_k) \log_2 P(v_k)$$

Information gain → $G(y, x) = E(y) - E(y|x)$

- $\nearrow$ target
- $\nearrow$ target given x
- $\searrow$ test for x

$Remainder(A) = \sum\limits_{k=1}^{d} \dfrac{p_k + n_k}{p+n} \; B\left(\dfrac{p_k}{p_k + n_k}\right)$

- $\nearrow$ proportion
- $\nearrow$ entropi
- $\searrow$ A'nın aldığı değerler
- $\searrow$ her value için hesaplayıp topla

Root node'da → dataset'in entropisi — feature'in entropisi

$IG = $ Parent entropy $-$ Feature'in entropisi