Arbres de décision : Formules pour les calculs

1- Mesure de l'information

L'entropie de Shannon exprime la quantité d'information : Le nombre de bits nécessaires pour coder l'information.

- P bits permettent de coder 2^P informations.
- $\log_2(N)$ bits permettent de coder N informations.

Si on a n classes (C_1, C_2,.., C_n) de probabilités respectives p_1, p_2,.., p_n, la quantité d'information relative à la connaissance de la classe est définie par **l'entropie d'information** :

$$I = \sum_{i=1..n} -p_i \log_2 p_i$$

- **I = 0** quand $\exists$ i/ $p_i = 1$ (une seule classe).
- **I est maximale** quand $\forall$ i/ $p_i = 1/n$ (classes équiprobables).

2- Gain d'information (ID3)

- **freq(T, C_j) :** Nombre d'objets de T appartenant à la classe C_j.
- L'information relative à T est définie :

$$Info(T) = -\sum_{j=1}^{n} \frac{freq(T, C_j)}{|T|} \log_2 \frac{freq(T, C_j)}{|T|}$$

- Une mesure similaire de T après partition selon l'attribut A (contenant n valeurs) est :

$$Info_A(T) = \sum_{i \in D_A} \frac{|T_i|}{|T|} Info(T_i)$$

D_A =Domaine de valeurs de l'attribut A.

- Le gain d'information mesure le gain obtenu suite au partitionnement selon l'attribut A :

$$Gain(T, A) = Info(T) - Info_A(T)$$

3- Ratio de gain (C4,5)

- Une mesure de l'information contenue dans l'attribut A (mesure de dispersion) est définie :

$$Split\ Info(T, A) = -\sum_{i \in D_A} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}$$

- Le ratio de gain mesure le gain calibré par Split Info :

$$Gain\ Ratio(T, A) = \frac{Gain(T, A)}{Split\ Info(T, A)}$$

Quantité d'information générée par T et utile pour la classification