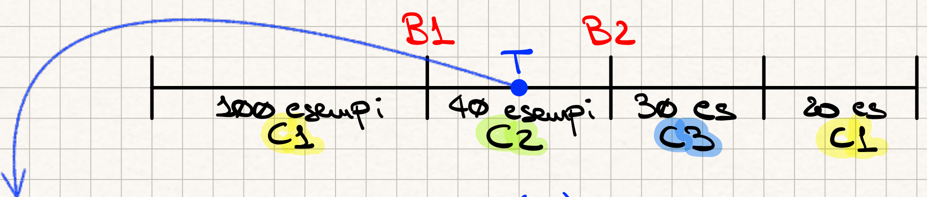
DISCRETIZZARE CON FAYYAD-IRANI

L'obiettivo è trovare una partizione ottima per ogni valore continuo, determinata tramite i punti di taglio.
Gli autori del paper hanno dimostrato che i punti di taglio ottimali stanno sempre fra due esempi di classi diverse, considerando la sequenza dei punti ordinata secondo un certo attributo A.



Potenziale punto di taglio (T)

Se vogliamo definirlo T è un punto di confine b sse considerando la sequenza di punti ordinati secondo l'attributo A:

$$\exists e_1, e_2 : e_1 \in C_i \text{ e } e_2 \in C_j : A(e_1) < T < A(e_2) \text{ e}$$
$$\nexists e' : A(e_1) < A(e') < A(e_2)$$

Dove e_1 e e_2 sono due esempi di classe diversa.

Assumiamo che b sia il punto in mezzo fra $A(e_1)$ e $A(e_2)$.
Sia B_A il set di tutti i punti di frontiera candidati per l'attributo A.
Questo algoritmo usa la formula dell'entropia per valutare se usare un punto rispetto a un altro.

Consideriamo di avere un set di esempi S e di avere k classi diverse $\langle C_1 \dots C_k \rangle$.
Sia $P(C_i, S)$ il numero di esempi in S che ha classe C_i. Definiamo l'entropia:

$$H(S) = -\sum_{i=1}^{k} P(C_i, S) \log(P(C_i, S))$$

Sia T un punto di taglio: $T \in B_A$.
S viene diviso in due sotto-set S_1 e S_2, precisamente usando T.
Possiamo dunque definire $\delta P(A, T; S)$ come

$$\delta P(A, T; S) = \frac{|S_1|}{N} H(S_1) + \frac{|S_2|}{N} H(S_2)$$

Prendiamo il punto di taglio T che ha $H(A, T; S)$ minima.

ALGORITMO

Prima si partiziona S in S_1 e S_2 poi si ripete la stessa procedura per $S = S_1$ e $S = S_2$ fino a che non si soddisfa la stopping condition:

$$G(A, T_A; S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A, T_A; S)}{N}$$

Dove $\Delta(A, T_A; S) = \log_2[(3^c) - c_1 H(S_1) - c_2 H(S_2)]$
con c, c_1 e c_2 sono il numero di classi diverse in S, S_1 e S_2 rispettivamente.

$$G(A, T_A; S) = H(S) - \Delta P(A, T; S)$$