DECISION TREE

OSSERVAZIONI

Questo approccio è basato su un'assunzione statistica teoricamente sbagliata, ma che poi praticamente funziona.

Nell'approccio classico si assume di avere tutto il DB disponibile, cosa che nei data stream non abbiamo.

Obiettivo

Costruire l'albero con i pochi dati disponibili oppure decidere se bisogna far crescere l'albero oppure no.

Per trovare l'attributo ottimo su cui fare lo split è sufficiente considerare un sotto insieme del DB che passa da quel nodo. Bisogna scegliere come capire se le istanze che abbiamo siano sufficienti per prendere la decisione di continuare con lo split, per prendere la decisione ci basiamo sul Hoeffding Bound.

HOEFFDING BOUND

Questa metrica ci da informazione sul capire se abbiamo un ottima stima di una variabile aleatoria X.

Assumiamo di avere $\{x_1 \dots x_n\}$, n valori di X, non tutti. Ogni valore x_i varia dentro un intervallo grande R. Chiamiamo $\bar{x}$ la media dei valori $\{x_1 \dots x_n\}$.

L'Hoeffding Bound afferma che con probabilità $1-\delta$, la media totale di X ($\bar{X}$) è almeno uguale a $\bar{x} - \varepsilon$ con

$$\varepsilon = \sqrt{\frac{R^2 \ln 1/\delta}{2n}}$$

Guardando la formula ε è indirettamente proporzionale a n. Intuitivamente più campioni di X abbiamo più possiamo approssimare meglio la media $\bar{X}$, che si avvicinerà di più a $\bar{x}$ essendo ε più piccolo.

d. Se fissiamo δ, andiamo a stabilire implicitamente il numero di istanze necessarie ad avere una buona approssimazione dei dati

COME USARE L'HOEFFDING BOUND PER UN DT?

Assumiamo che $G(X_i)$ sia una metrica su cui basare la costruzione dell'albero (Gini Index o Info Gain).
Diciamo che X_A è l'attributo che ha $G(X_A)$ maggiore dopo aver aspettato n istanze dello stream.
X_B è il secondo con $G(X_B)$ più alto.
Possiamo settare δ per cui:

$$\underbrace{\Delta \bar{G}}_{\text{Stimato}} = \bar{G}(X_A) - \bar{G}(X_B) > \varepsilon = \sqrt{\frac{R^2 \ln 1/\delta}{2n}}$$

$R = \ln |C|$: numero di classi nelle n istanze.

L'Hoeffding Bound garantisce che il ΔG reale sia:

$$\Delta G \geq \Delta \bar{G} - \varepsilon > 0 \quad \text{con probabilità } 1-\delta$$

PROBLEMA.

L'Hoeffding Bound funziona solo con variabili aleatorie che possono essere espresse come somma dei loro elementi.
Questo non vale per la misura di splitting.
La soluzione è cambiare la formula di ε:

$$\varepsilon = C_{Gain}(|C|, N)\sqrt{\frac{\ln(1/\delta)}{2N}}$$

(# classi)

$$C_{Gain}(|C|, N) = G\left(|C| \log_2 eN + \log_2 2N\right) + 2\log_2 |C|$$

Non è importante saperlo.

PRE-PRUNING

Per fare prepruning, si considera per ogni nodo su cui applicare lo split un nodo "nullo" ovvero uno per cui si prende la decisione di non usare X_n per fare lo splitting.
Si fa lo split se il G stimato sia migliore dello G stimato sull'attributo per cui NON fare lo split.