ATTRIBUTE SELECTION

A volte possiamo ridurre le dimensioni del Dataset omettendo alcuni attributi che risultano inutili o ridondanti (Spesso ricavabili da altri attributi).

APPROCCIO GREEDY

Con d attributi si possono avere 2^d combinazioni

Esempio:

Se ho due attributi A e B, posso fare $2^2 = 4$ combinazioni:

1) Prendo AB
2) Prendo A
3) Prendo B
4) Non prendo nessuno

Fra i possibili approcci greedy, abbiamo:

1) Best single attribute: sotto l'assunzione che tutti gli attributi sono indipendenti, sceglierli secondo qual test significativo
2) Best step-wise features selection: selezionare a ogni step l'attributo migliore
3) Best step-wise features elimination: partire da tutti e a ogni passo eliminare i peggiori
4) Combinare i precedenti approcci

APPROCCIO EURISTICO CON MUTUAL INFORMATION

Consideriamo due variabili DISCRETE X e Y, definiamo XX e YY tutti i possibili valori di X e Y
Definiamo $I(X,Y)$, mutual information come:

$$I(X,Y) = \sum_{x \in XX} \sum_{y \in YY} p(x,y) \cdot \log \frac{p(x,y)}{p(x)\,p(y)}$$

Dove $p(x,y)$ è la probabilità congiunta, ovvero la probabilità che sia $X = x$ e $Y = y$.
Mentre $p(x)$ e $p(y)$ sono le probabilità marginali, indipendenti fra loro.

ENTROPIA

Data una distribuzione D e per ogni valore di D le corrispondenti probabilità che si verificano, definiamo l'entropia come:

$$H(D) = -\sum_i p_i \log_2 p_i$$

Proprietà del MI:

1) Capacità di misurare ogni tipo di relazione fra le variabili interessate

2) Non varia se si cambia lo spazio (traslazione, rotazione, ecc... basta non variare l'ordine del dataset originale)

Come fare con un dataset con classi.

Dato un set inizial F con n attributi, bisogna trovare un sottoinsieme $S \subset F$ con K attributi che massimizza il $I(C,S)$, dove C è la classe variabile.

PASSI

1) Inizializzare F e S:
   - F = set iniziali con n attributi
   - S = vuoto

2) Calcolare $I(C, f_i) \quad \forall f_i \in F$

3) Selezionare il primo attributo f_i che massimizza MI
   setta:
   - $F \leftarrow F / \{f_i\}$
   - $S \leftarrow \{f_i\}$

4) Selezionare gli attributi diversi dal primo:
   - Calcolare MI per ogni coppia (f_i, f_s) dove $f_i \in F$ e $f_s \in S$.
   - Selezionare la feature che massimizza

$$I(C, f_i) - \frac{1}{|S|} \sum_{f_s \in S} NI(f_s, f_i)$$

SST:
- $F \leftarrow F / \{f_i\}$
- $S \leftarrow \{f_i\}$

Ripetere finché $|S| = K$

5) Arriviamo al nostro risultato S.

NORMALIZZANDO

Normalizziamo MI:

$$NI(f_i, f_s) = \frac{I(f_i, f_s)}{\min\{H(f_i), H(f_s)\}}$$

Focus su $-\frac{1}{|S|}\sum_{f_s \in S} NI(f_s, f_i)$

È una sorta di misura di correlazione fra f_s e f_i. Più è vicina a $\emptyset$ più f_i è indipendente dal sottoinsieme di features in S. Vicino a 1 è indice di correlazione.

Difatti è la media di tutti i MI fra f_i e ogni elemento in S. Termine adattivo.