

## z0. Traccia Progetto Architetture

- [Introduzione](#)
- [Obiettivo del progetto](#)
- [Calcolo di un singolo  \$r\_{fc}\$](#)
- [Deviazione standard da usare in un  \$r\_{fc}\$](#)
- [Calcolo di un singolo  \$r\_{ff}\$](#)
- [Riassunto finale dei passi](#)
- [Algoritmo di costruzione dell'insieme S](#)

### Introduzione

- ➡ Abbiamo a disposizione un dataset DS

righe → oggetti

colonne → attributi

- ? funzionamento della variabile dicotomica:  
→ *costruiamo un esempio con un dataset inventato*

Età	Pressione	Zucchero	Colesterolo	Fumo	Obesità	Affetto da Malattia?
45	120	80	200	0	1	0
60	150	90	240	1	0	1
35	130	85	180	0	0	0

Ogni riga nel dataset rappresenta un paziente e la variabile 'c' è una variabile dicotomica che indica se quel paziente è affetto dalla malattia (1) o no (0).

→ immaginiamo c come un vettore di 0 e 1

- l'elemento i-esimo di c, ci dice se il paziente i (i-esima riga del dataset) sia affetto da malattia o meno

L'obiettivo della feature selection in questo contesto potrebbe essere quello di determinare quali delle caratteristiche (Età, Pressione Sanguigna, Livello di Zucchero nel Sangue, Colesterolo, Fumo, Obesità) sono più importanti per predire se un paziente è affetto dalla malattia o meno.

---


### Obiettivo del progetto

1. abbiamo un insieme  $F$  di  $d$  feature
2. dato un valore  $k$ , vogliamo trovare l'insieme  $S$  di  $k$  feature più influenti tra le  $d$  esistenti, usando un criterio di qualità predefinito
3. vogliamo trovare le feature più rilevanti e non ridondanti tra loro

Per misurare rilevanza e ridondanza potremmo usare la funzione di correlazione e la relativa tecnica di selezione:

### Important

La *Correlation Feature Selection (CFS)* è una tecnica di selezione delle caratteristiche (feature) che si basa sull'idea di individuare un sottoinsieme di feature altamente correlate con la variabile bersaglio (target), ma che siano poco correlate tra di loro all'interno dell'insieme stesso.


-  esempio applicato al nostro dataset

In termini teorici, possiamo utilizzare il concetto di correlazione. La correlazione tra due variabili misura il grado di relazione lineare tra di esse.

- Ad esempio, la correlazione tra una feature del dataset (come ad esempio "Età") e la variabile di classe ("Affetto da Malattia?") può essere calcolata utilizzando il coefficiente di correlazione di Pearson, che fornisce un valore compreso tra -1 e 1, dove:
  - 1 indica una correlazione positiva perfetta (entrambe le variabili aumentano o diminuiscono insieme)
  - 0 indica assenza di correlazione
  - -1 indica una correlazione negativa perfetta (quando una variabile aumenta, l'altra diminuisce e viceversa)

### Note

Il "merito" di un insieme di feature  $S$  indica quanto quell'insieme di caratteristiche sia rilevante o utile nel contesto di un'algoritmo di selezione delle feature come la Correlation Feature Selection (CFS).

-  **Spiegazione della formula:** usando il nostro esempio
- ➡ Il merito di un insieme di feature  $S$  costituito da  $k$  elementi può essere calcolato con la seguente equazione:

$$merits_{S_k} = \frac{k \cdot |\overline{r_{cf}}|}{\sqrt{k + k \cdot (k - 1) |\overline{r_{ff}}|}}$$

- $k$ : numero delle feature scelte
- $\overline{r_{cf}}$ : è il valore medio di tutte le correlazioni tra le feature e le variabili di classe
  - in questo caso abbiamo una sola variabile di classe → "Affetto da Malattia?"
- $\overline{r_{ff}}$ : è il valore medio di tutte le correlazioni tra le feature stesse

## Calcolo di un singolo $r_{fc}$

- ➡ concentriamoci su  $\overline{r_{cf}}$ :
  - calcoliamo la singola correlazione tra feature e variabile di classe

$$r_{cf} = \frac{\mu_0 - \mu_1}{\sigma_f} \cdot \sqrt{\frac{n_0 \cdot n_1}{n^2}}$$

- $\mu_0$ : media dei valori assunti dal gruppo 0 sulla feature  $f$
- $\mu_1$ : media dei valori assunti dal gruppo 1 sulla feature  $f$
- $n_0$ : numerosità del gruppo 0
- $n_1$ : numerosità del gruppo 1
- $n = n_0 + n_1$ : numerosità totale

ovviamente noi usiamo questa formula per ogni feature presente, ovvero per ogni correlazione tra una feature e la variabile che stiamo predicendo *esempio età / affetto da malattia*

→ poi facciamo la media tra tutti

- ? Considerando i dati del nostro esempio, supponiamo di voler calcolare la correlazione tra la feature "Età" rispetto alla variabile di classe "Affetto da Malattia":

*Ricordiamo:*

- gruppo 0: non affetto da malattia
- gruppo 1: affetto da malattia

$\mu_0$ : età media dei pazienti non affetti da malattia:

$(45+35)/2$

$\mu_1$ : età media dei pazienti affetti da malattia:

60

- ➡ calcoliamo la media nel modo classico
- $\sigma_{ETA}$ : deviazione standard della feature
  - quanto i valori di età registrati si discostano dal valore medio dell'età del dataset

- ⇨ abbiamo una formula data dalla traccia  
 $n_0$ : numero pazienti non affetti → gruppo 0:  
 2  
 $n_1$ : numero pazienti affetti → gruppo 1:  
 1  
 $n$ :  
 3

## Deviazione standard da usare in un $r_{fc}$

- ⇨ per ogni calcolo di  $r_{fc}$  rispetto ad una feature  $f$  ci calcoliamo la deviazione standard per quella determinata feature nel dataset

$$\sigma_f = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \mu)^2}$$

Ricordiamo che  $x_i$  rappresenta l' $i$ -esima osservazione per quella determinata feature

### Info

Dopo aver applicato questo procedimento per ogni coppia feature variabile da predire, possiamo trovare la media →  $\overline{r_{cf}}$

## Calcolo di un singolo $r_{ff}$

- ⇨ adesso dobbiamo trovare la media →  $\overline{r_{ff}}$
- per ogni coppia di feature  $ff$  usiamo la seguente formula:

$$r_{f_x f_y} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}}$$

- 🔗 facciamo un esempio con il nostro caso studio

Immaginiamo di voler calcolare la correlazione tra le due feature, età/pressione sanguigna

- la formula sarà la seguente:

$$r_{\text{Età, Pressione}} = \frac{\sum_{i=1}^n (x_{\text{Età}_i} - \mu_{\text{Età}})(x_{\text{Pressione}_i} - \mu_{\text{Pressione}})}{\sqrt{\sum_{i=1}^n (x_{\text{Età}_i} - \mu_{\text{Età}})^2} \cdot \sqrt{\sum_{i=1}^n (x_{\text{Pressione}_i} - \mu_{\text{Pressione}})^2}}$$

### Important

Questo calcolo viene eseguito per ogni coppia di feature all'interno dell'insieme di feature 'S' per ottenere le correlazioni tra di loro.

## Riassunto finale dei passi

Supponendo di avere un insieme S con tutte le feature


→ (nella pratica ne aggiungiamo una alla volta fino ad un massimo di k)

- dobbiamo calcolare  $r_{cf}$  per Età, Pressione Sanguigna, Livello di Zucchero nel Sangue, Colesterolo, Fumo, Obesità
  - facciamo la media
- dobbiamo calcolare  $r_{fc}$  per ogni coppia di attributi
  - facciamo la media
- sostituiamo i valori della formula principale per il calcolo del merito

## Algoritmo di costruzione dell'insieme S

### Important

Adesso possiamo sostituire e calcolare il merito **per quel determinato insieme di k feature**

-  Come costruiamo l'insieme S:

→ iterativamente

*pseudocode*

#### ALGORITMO 2: Correlation Feature Selection

**Input:** un dataset  $DS$  definito sull'insieme di feature  $\mathcal{F}$ , un vettore  $c$  contenente le etichette, il numero  $k$  di feature da estrarre

**Output:** l'insieme  $\mathcal{S} \subseteq \mathcal{F}$  delle  $k$  feature selezionate

```
1 begin
2    $\mathcal{S} = \emptyset$ ;
3   while  $|\mathcal{S}| < k$  do
4     calcolare, per ogni feature  $f_i$ , il punteggio  $merit_{\mathcal{S} \cup \{f_i\}}$  dell'insieme  $\mathcal{S} \cup \{f_i\}$ ;
5     sia  $f_i^*$  la feature che ha ottenuto il punteggio massimo;
6      $\mathcal{S} = \mathcal{S} \cup f_i^*$ ;
7     aggiungere all'insieme  $\mathcal{S}$  la feature  $f_i$  che ha ottenuto il punteggio massimo;
```

L'algoritmo seleziona in modo iterativo la feature che massimizza il punteggio *merit* quando aggiunta all'insieme corrente S, finché l'insieme S non contiene k feature.

• ➡ **Traccia a chat gpt:**

hai sbagliato la parte della formula, proverò a darti più informazioni. Vorrei che mi facessi un esempio usando il dataset precedente (ti scriverò le formule in linguaggio latex, anche tu scrivimele con la stessa notazione). il merito di un insieme di feature S costituito da k elementi

può essere calcolato con la seguente equazione:

$$merits_{S_k} = \frac{k \cdot |\overline{r_{cf}}|}{\sqrt{k + k \cdot (k - 1) |\overline{r_{ff}}|}}$$

. Dove  $\overline{r_{cf}}$  è il valore medio di tutte le correlazioni feature-classification e  $\overline{r_{ff}}$  è il valore medio di tutte le correlazioni feature-feature.

In particolare, essendo la classe una variabile categorica dicotomica e la feature una variabile

numerica, per calcolare la correlazione tra una feature f e la classe c si ricorre al point biserial

correlation coefficient. Dividendo il dataset in due gruppi 0 e 1 sulla base del valore assunto dalla variabile di classe, il point biserial correlation coefficient è definito come:

$$r_{cf} = \frac{\mu_0 - \mu_1}{\sigma_f} \cdot \sqrt{\frac{n_0 \cdot n_1}{n^2}}$$

; dove  $\mu_0$  è la media dei valori assunti dal gruppo 0 sulla feature f,  $\mu_1$  è la media dei valori assunti

dal gruppo 1 sulla feature f,  $\sigma_f$  è la deviazione standard campionaria relativa alla feature f,

$n_0$  è la numerosità del gruppo 0,  $n_1$  è la numerosità del gruppo 1 e  $n = n_0 + n_1$ . La deviazione

standard campionaria di una feature f si calcola come:

$$\sigma_f = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \mu)^2}$$

; dove n è la numerosità dei dati,  $x_i$  sono i valori assunti sulla feature f e  $\mu$  è il loro valore medio.

Riguardo la correlazione feature-feature  $r_{ff}$ , per questa si ricorre al Pearson's correlation coefficient, definito come:

$$r_{ff} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}}$$

; dove  $x_i$  sono i valori assunti sulla feature  $f\{x\}$ ,  $\mu\{x\}$  è il loro valore medio,  $y_i$  sono i valori assunti sulla feature  $f_y$ ,  $\mu\{y\}$  è il loro valore medio e  $n$  è la numerosità dei dati.