13. Preparazione dei dati

Corso di Python per il Calcolo Scientifico

Angelo Cardellicchio
angelo.cardellicchio@stiima.cnr.it

Outline

- Campionamento
  - Dimensione e qualità del dataset
- Preparazione dei dati
  - Pulizia dei dati
  - Sbilanciamento dei dati
  - Trasformazione dei dati
- Dati di training e dati di validazione

Campionamento

- La campagna di acquisizione deve portare a dati a **qualità** e **quantità** elevata
- Per quello che riguarda la **quantità**, è consigliabile usare un numero di campioni **almeno un ordine di grandezza superiore ai parametri addestrabili**
- I dati devono essere di qualità, **ovvero rappresentativi del fenomeno nella sua interezza**

Dati	Quantità di dati	Qualità dei dati
Acquisizioni meteo ogni 15 minuti per 100 anni, solo mese di luglio	$4 \cdot 24 \cdot 31 \cdot 100 = 297.600$	Bassa: modello in grado di prevedere solo precipitazioni luglio
Acquisizioni meteo ogni giorno per 100 anni, tutti i mesi	$365 \cdot 100 = 36.600$	Medio/alta: modello in grado di prevedere precipitazioni per ogni mese

Preparazione dei dati

- In primis, **ricordiamoci sempre di rimuovere eventuali informazioni personali**.
- Procediamo poi alla **pulizia dei dati**, individuando l'occorrenza dei seguenti.
  - **Errori nel labeling**: l'esperto di dominio ha svolto il suo compito in maniera ottimale?
  - **Rumorosità**: i dati sono affetti da errori o rumore?
  - **Dati mancanti**: i valori di alcune feature per certi campioni sono mancanti?
  - **Valori duplicati**: esistono dei campioni duplicati?
  - **Misure errate**: esistono delle misure errate o prese su range e scale differenti?
- **Va elaborata una strategia di data cleaning** (rimozione o filling).

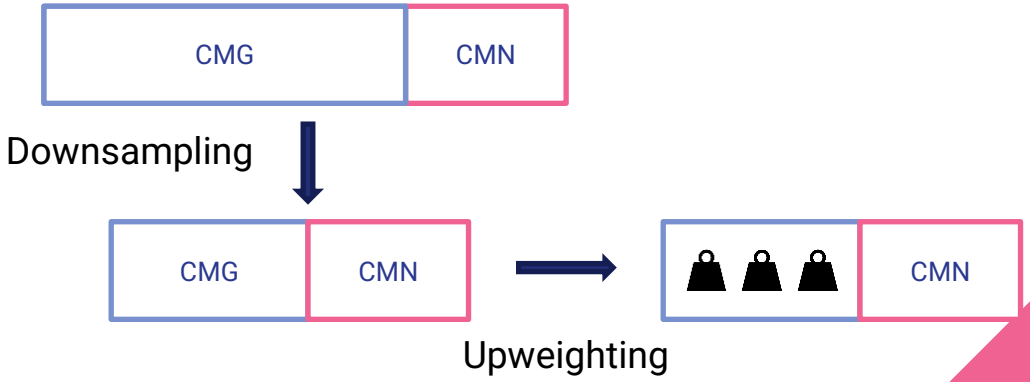
Sbilanciamento del dataset (1)

- Non è detto che i campioni siano sempre equamente distribuiti tra classi.
- Se i campioni non sono equamente distribuiti, si parla di **sbilanciamento del dataset**.
- Un dataset sbilanciato compromette la qualità del modello addestrato.

Grado di sbilanciamento	Peso delle classi minoritarie
Moderato	$20 - 40\%$
Leggero	$1 - 20\%$
Estremo	$< 1\%$

Sbilanciamento del dataset (2)

- Per ovviare allo sbilanciamento è opportuno adottare opportune soluzioni.
  - La più semplice è quella di campionare più dati per la classe sottocampionata!
- Se l'acquisizione di più dati è infattibile, è possibile effettuare operazioni di **downsampling** ed **upweighting**.



Trasformazione dei dati (1)

- Le tecniche di trasformazione sono differenti a seconda che si tratti di dati numerici o meno.
- Per i **dati numerici**, abbiamo le seguenti possibilità.
- **Scaling**: i dati sono scalati in un intervallo $[x_{min}, x_{max}]$ usando la seguente formula:

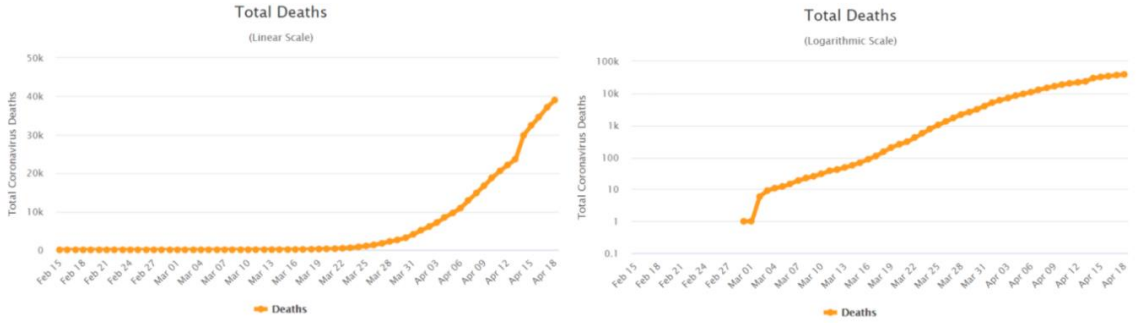
$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Clipping**: i dati al di fuori del range $[\mu - 3\sigma, \mu + 3\sigma]$, oppure al di fuori del range interquartile, sono scartati.

Trasformazione dei dati (2)

- **Trasformazione logaritmica**: i dati ‘compressi’ applicando una trasformazione di tipo logaritmico:

$$x_{new} = Log(x)$$



Fonte: LSE blog

- **Z-score**: i dati sono ridisposti secondo una distribuzione normale a media nulla e varianza unitaria.

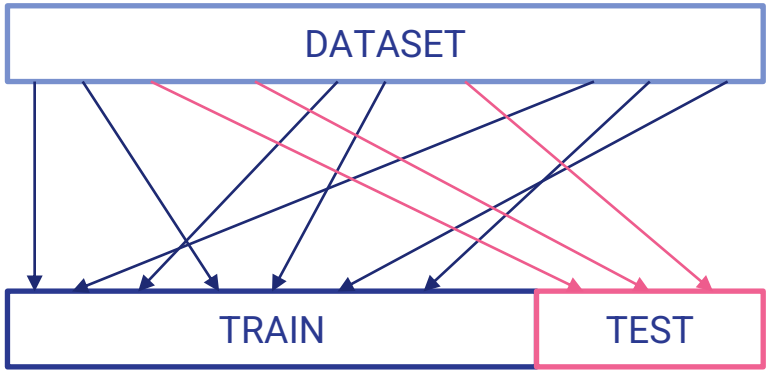
Trasformazione dei dati (3)

- I **dati categorici** possono essere trasformati in **interi** oppure in una **rappresentazione vettoriale** mediante **one – hot encoding**.
- Ciò va a creare un **dizionario**.

Valore	Rappresentazione intera	One – hot encoding
Lunedì	1	[1 0 0 0 0 0 0]
Martedì	2	[0 1 0 0 0 0 0]
Mercoledì	3	[0 0 1 0 0 0 0]
Giovedì	4	[0 0 0 1 0 0 0]
Venerdì	5	[0 0 0 0 1 0 0]
Sabato	6	[0 0 0 0 0 1 0]
Domenica	7	[0 0 0 0 0 0 1]

Suddivisione dei dati

- I dati devono essere suddivisi in **dati di training** e **dati di test**.
  - In specifiche situazioni, ovvero per comparare diversi modelli, si possono usare anche degli insiemi di dati di **validazione**.
- La suddivisione è di solito casuale ed in un rapporto di **70-30**.



Domande?

42