Python per il Calcolo Scientifico

Angelo Cardellicchio

Introduzione al Machine Learning

- Dati e feature
  - Campioni e sbilanciamento del dataset
  - Feature e feature selection
- Curse of dimensionality
- Tipi di dati
  - Dati IID
  - Serie temporali
- Preprocessing

Dati e feature

Dato

Feature

	Year	EventType	Actor1	Actor2	Country	Region	Location	ConflictLat	ConflictLong	StationID	YrMoDy	MaxTemp	StationName	StationLong
0	2005	Riots	Protesters (Algeria)	NaN	Algeria	Chlef	Sidi Ammar	36.47	1.45	604300	20050125	35.79	MILIANA	2.23
1	2003	Riots	Protesters (Algeria)	NaN	Algeria	NaN	Tadjenanet	36.11	5.98	604680	20030201	38.50	BATNA	6.31
2	2002	Battles	Military Forces of Ethiopia (1991-)	ONLF: Ogaden National Liberation Front	Ethiopia	Degeh Bur	Afweyne	9.38	43.06	696754	20020224	39.20	CAMP LEMONIER	43.15
3	2003	Riots	Protesters (Algeria)	Police Forces of Algeria (1999-)	Algeria	Bordj Bou Arreridj	Bordj Bou Arerridj	36.07	4.77	604440	20030217	39.40	BORDJ-BOU-ARRERIDJ	4.76
4	1999	Violence against civilians	GIA: Armed Islamic Group of Algeria	Civilians (Algeria)	Algeria	Relizane	Relizane	35.74	0.55	605060	19991217	39.90	MASCARA-MATEMORE	0.30

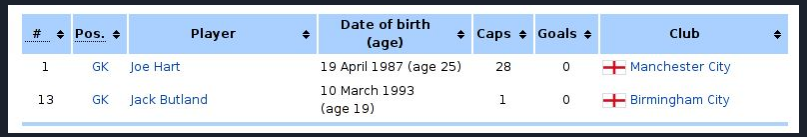
Campioni e sbilanciamento del dataset

- Esempio: come valutare i calciatori in base alla nazionalità?

#	Pos.	Player	Date of birth (age)	Caps	Goals	Club
1	GK	Joe Hart	19 April 1987 (age 25)	28	0	Manchester City
13	GK	Jack Butland	10 March 1993 (age 19)	1	0	Birmingham City

Solo inglesi!

Feature e feature selection



#	Pos.	Player	Date of birth (age)	Caps	Goals	Club
1	GK	Joe Hart	19 April 1987 (age 25)	28	0	Manchester City
13	GK	Jack Butland	10 March 1993 (age 19)	1	0	Birmingham City

- Alcune feature sono *ridondanti*, e peggiorano le performance degli algoritmi.
- Esempio base: feature a bassa varianza.
- Rimuovere feature non informative è essenziale per evitare il problema della *curse of dimensionality*.

Curse of Dimensionality

- Le feature determinano uno *spazio n-dimensionale*.
- Man mano che le dimensioni dello spazio delle feature aumentano, i dati diventano *sparsi*.
- Ciò comporta un problema di significatività statistica: man mano che aumentano le feature considerate, *aumenta il numero di campioni richiesti*.
- Facciamo un esempio.

Curse of Dimensionality

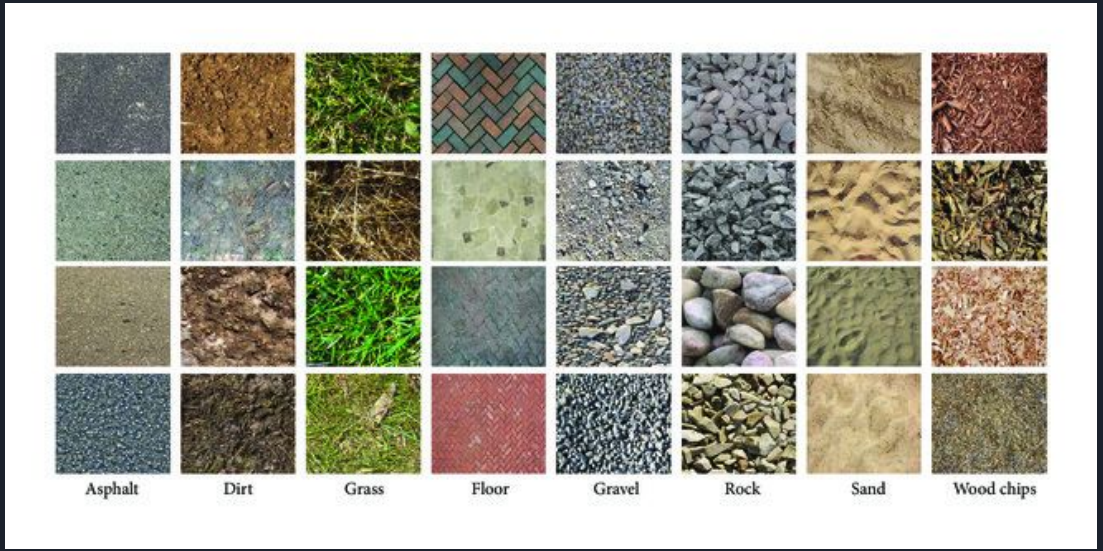
- *Quanti campioni sono necessari per visualizzare tutte le combinazioni possibili?*

Age group	Gender	Target
Children	Male	T1
Youth	Male	T2
Adult	Male	T3
Senior	Male	T4
Children	Female	T5
Youth	Female	T6
Adult	Female	T7
Senior	Female	T8

Age group	Gender	Body type	Target
Children	Male	Normal	T1
Children	Male	Over weight	T2
Children	Male	Obese	T3
Youth	Male	Normal	T4
Youth	Male	Over weight	T5
Youth	Male	Obese	T6
Adult	Male	Normal	T7
Adult	Male	Over weight	T8
Adult	Male	Obese	T9
Senior	Male	Normal	T10
Senior	Male	Over weight	T11

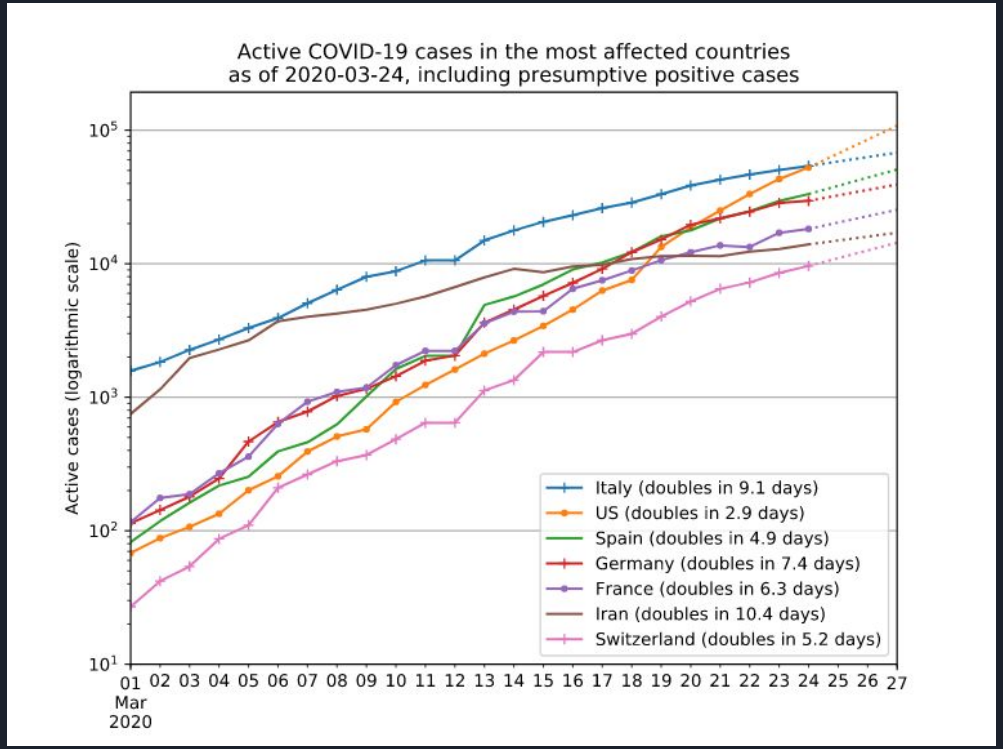
Dati IID

- Ogni campione è *indipendente* dagli altri.
- Inoltre, tutti i campioni *sottendono alla stessa distribuzione*.



Serie temporali

- I campioni sono *temporalmente dipendenti* rispetto ai precedenti



ML Pipeline

- Di solito, gli algoritmi di machine learning non sono *isolati*, ma vengono eseguiti in *cascata*.
- Possiamo quindi definire una *processing pipeline*.