Fondamenti di Machine Learning

il Dataset

Dati Strutturati

presentato da
Giuseppe Gullo

PROFESSION AI

Formati di dati strutturati

CSV

TSV

EXCEL

JSON

SQL

HTML

XML

Formati di dati strutturati

CSV

TSV

EXCEL

JSON

SQL

HTML

XML

CSV Comma Separated Values

Le colonne sono divise da una virgola,
le righe sono divise da un a capo

```
,Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Species
0,6.0,3.0,4.8,1.8,virginica
1,5.1,3.4,1.5,0.2,setosa
2,5.7,2.5,5.0,2.0,virginica
3,5.7,2.8,4.5,1.3,versicolor
4,4.6,3.4,1.4,0.3,setosa
5,6.2,2.9,4.3,1.3,versicolor
6,6.1,2.8,4.7,1.2,versicolor
7,6.3,2.9,5.6,1.8,virginica
8,5.8,2.7,3.9,1.2,versicolor
9,5.7,2.8,4.1,1.3,versicolor
10,4.8,3.1,1.6,0.2,setosa
11,5.6,2.8,4.9,2.0,virginica
```

CSV Comma Separated Values

Ha una struttura tabulare

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	6.0	3.0	4.8	1.8	virginica
1	5.1	3.4	1.5	0.2	setosa
2	5.7	2.5	5.0	2.0	virginica
3	5.7	2.8	4.5	1.3	versicolor
4	4.6	3.4	1.4	0.3	setosa
5	6.2	2.9	4.3	1.3	versicolor
6	6.1	2.8	4.7	1.2	versicolor
7	6.3	2.9	5.6	1.8	virginica
8	5.8	2.7	3.9	1.2	versicolor
9	5.7	2.8	4.1	1.3	versicolor
10	4.8	3.1	1.6	0.2	setosa
11	5.6	2.8	4.9	2.0	virginica

TSV Tab Separated Values

Come il CSV, ma le colonne sono divise da un carattere di tabulazione

	Sepal.Length		Sepal.Width	Petal.Length	Petal.Width	Species
0	6.0	3.0	4.8	1.8	virginica	
1	5.1	3.4	1.5	0.2	setosa	
2	5.7	2.5	5.0	2.0	virginica	
3	5.7	2.8	4.5	1.3	versicolor	
4	4.6	3.4	1.4	0.3	setosa	
5	6.2	2.9	4.3	1.3	versicolor	
6	6.1	2.8	4.7	1.2	versicolor	
7	6.3	2.9	5.6	1.8	virginica	
8	5.8	2.7	3.9	1.2	versicolor	
9	5.7	2.8	4.1	1.3	versicolor	
10	4.8	3.1	1.6	0.2	setosa	
11	5.6	2.8	4.9	2.0	virginica	

JSON Javascript Object Notation

Le informazioni vengono salvate in formato chiave valore, possono essere anche annidate.

```
{
   "Clienti" : [
   {
      "nome" : {
                  "nome" : Giuseppe",
                  "cognome" : "Gullo"
       },
      "data di nascita" : {
                   "anno": "1991",
                   "mese": "Giugno",
                   "giorno" : "11"
       },
      "indirizzo" : {
                   "comune" : "Taormina",
                   "provincia" : "ME" ,
                   "CAP" : "98035" ,
                   "via" : "Viale Dioniso" ,
                   "civico" : "14"
                }
    },
    ...
    ]
}
```

JSON Javascript Object Notation

- Standard per il trasferimento di dati Client- Server
- Standard per database non relazionali

XML Extensible Markup Language

Le informazioni vengono racchiuse tra tag.

```
<quiz categoria="geografia">
   <domanda>
    Qual è la capitale della Francia ?
   </domanda>
   <risposta>
    Parigi
   </risposta>
</quiz>
```

XML Extensible Markup Language

- Vecchio standard per il trasferimento di dati Client- Server

HTML HyperText Markup Language

Simile all'XML, ma i tag sono predefiniti

```
<html>

    <head>
        <title>Ciao HTML</title>

    <body>
        <h1>Questo è un titolo</h1>
        <p>Questo è un paragrafo</p>
    </body>

</html>
```

HTML HyperText Markup Language

E’ il linguaggio del web

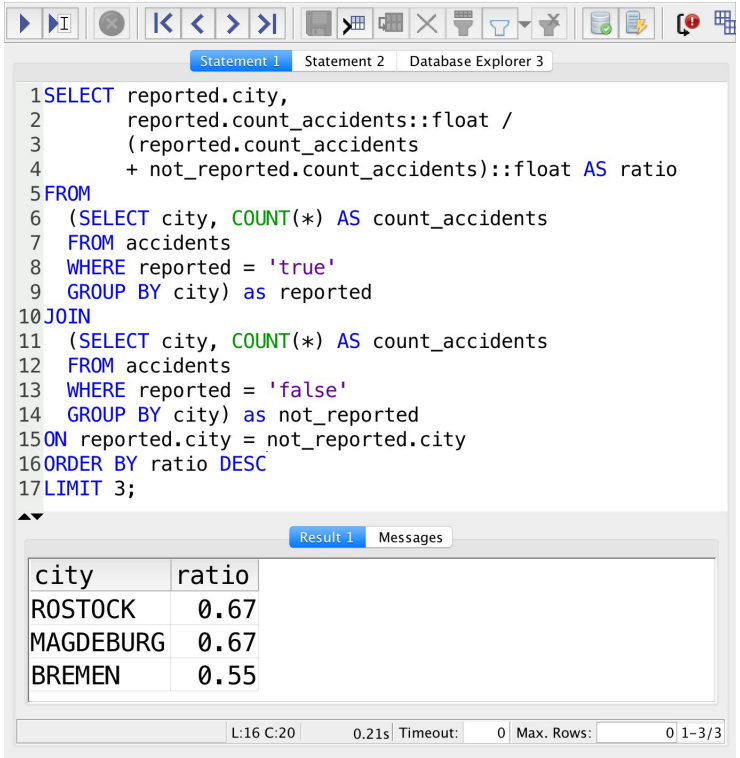
SQL Structured Query Language

Linguaggio utilizzato per gestire dati all'interno di database relazionali

	FirstName	LastName	Email	Phone	Position	Branch	Address
	Andrew	Fuller	afuller@contoso.com	(205) 555 - 9898	CEO	TopManagement	London, 120 Hanover Sq.
	Jeremy	Boather	jboather@contoso.com	(205) 555 - 9888	President QA	QA	London, 120 Hanover Sq.
	Anne	Dodsworth	adodsworth@contoso.com	(205) 555 - 9887	VP QA	QA	London, 120 Hanover Sq.
	Alexander	Tuckings	atuckings@contoso.com	(205) 555 - 9886	Team Lead...	QA	London, 120 Hanover Sq.
	Brenda	Smith	bsmith@contoso.com	(205) 555 - 9885	Senior QA	QA	London, 120 Hanover Sq.
	Mary	Bird	mbird@contoso.com	(205) 555 - 9885	Team Lead...	QA	London, 120 Hanover Sq.
	Steven	Buchanan	sbuchanan@contoso.com	(205) 555 - 9897	President ...	Development	London, 120 Hanover Sq.
	Robert	King	rking@contoso.com	(205) 555 - 9896	VP Dev De...	Development	London, 120 Hanover Sq.
	Laura	Callahan	lcallahan@contoso.com	(205) 555 - 9892	Team Lead...	Development	London, 120 Hanover Sq.
0	Jason	Roland	jroland@contoso.com	(205) 555 - 9872	Senior Dev	Development	London, 120 Hanover Sq.
1	Eric	Danstin	edanstin@contoso.com	(205) 555 - 9882	Team Lead...	Development	London, 120 Hanover Sq.
2	Elizabeth	Lincoln	elincoln@contoso.com	(205) 555 - 9862	Senior Dev	Development	London, 120 Hanover Sq.
3	Margaret	Peacock	mpeacock@contoso.com	(205) 555 - 9852	Senior Dev	Development	London, 120 Hanover Sq.

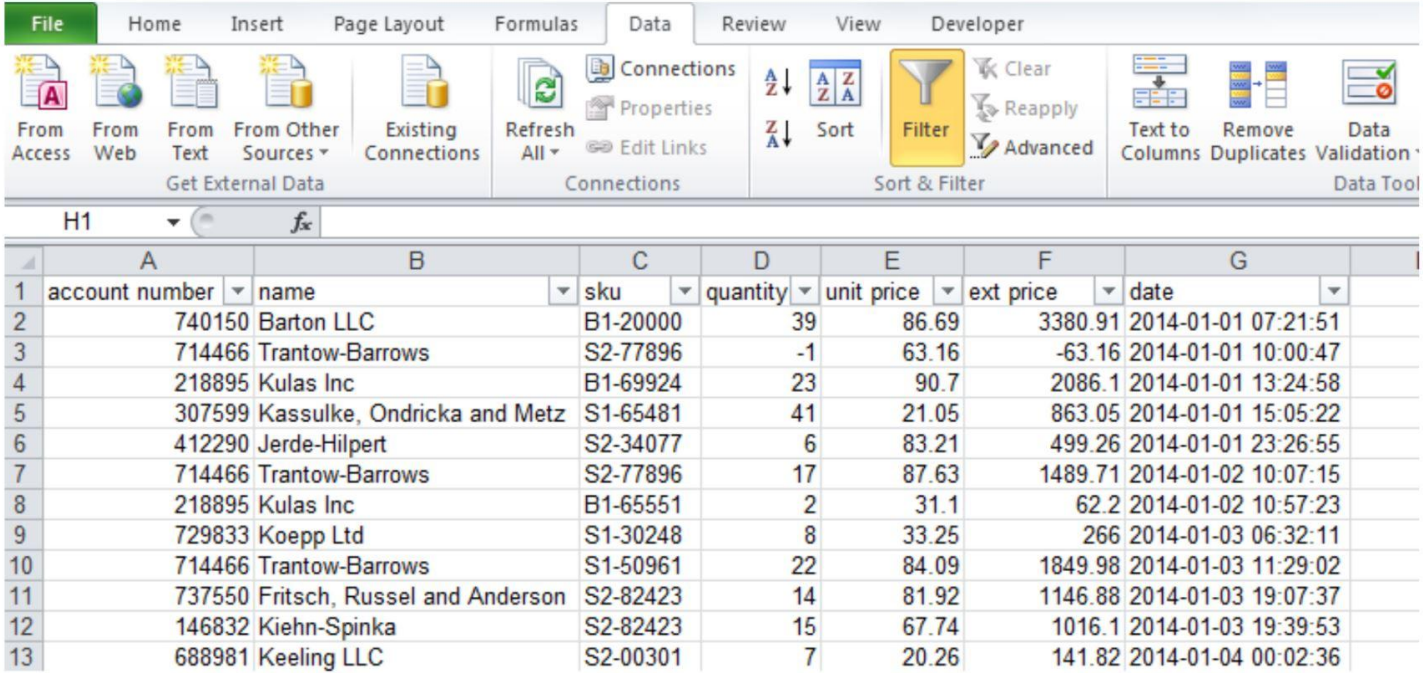
SQL Structured Query Language

Fornisce una serie di comandi per eseguire query anche complesse sui dati



EXCEL

Formato standard utilizzato per i fogli di calcolo



	A	B	C	D	E	F	G
1	account number	name	sku	quantity	unit price	ext price	date
2	740150	Barton LLC	B1-20000	39	86.69	3380.91	2014-01-01 07:21:51
3	714466	Trantow-Barrows	S2-77896	-1	63.16	-63.16	2014-01-01 10:00:47
4	218895	Kulas Inc	B1-69924	23	90.7	2086.1	2014-01-01 13:24:58
5	307599	Kassulke, Ondricka and Metz	S1-65481	41	21.05	863.05	2014-01-01 15:05:22
6	412290	Jerde-Hilpert	S2-34077	6	83.21	499.26	2014-01-01 23:26:55
7	714466	Trantow-Barrows	S2-77896	17	87.63	1489.71	2014-01-02 10:07:15
8	218895	Kulas Inc	B1-65551	2	31.1	62.2	2014-01-02 10:57:23
9	729833	Koepp Ltd	S1-30248	8	33.25	266	2014-01-03 06:32:11
10	714466	Trantow-Barrows	S1-50961	22	84.09	1849.98	2014-01-03 11:29:02
11	737550	Fritsch, Russel and Anderson	S2-82423	14	81.92	1146.88	2014-01-03 19:07:37
12	146832	Kiehn-Spinka	S2-82423	15	67.74	1016.1	2014-01-03 19:39:53
13	688981	Keeling LLC	S2-00301	7	20.26	141.82	2014-01-04 00:02:36

Fondamenti di Machine Learning

il Dataset

Dati non strutturati

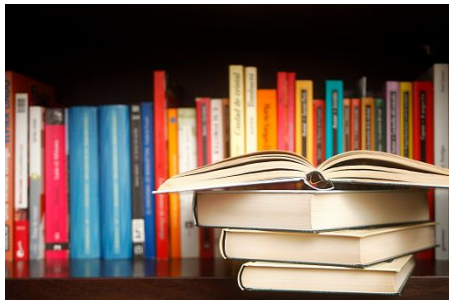
presentato da
Giuseppe Gullo

PROFESSION AI

Formati di dati non strutturati

Dati che non sono disposti secondo uno schema predefinito

IMMAGINI

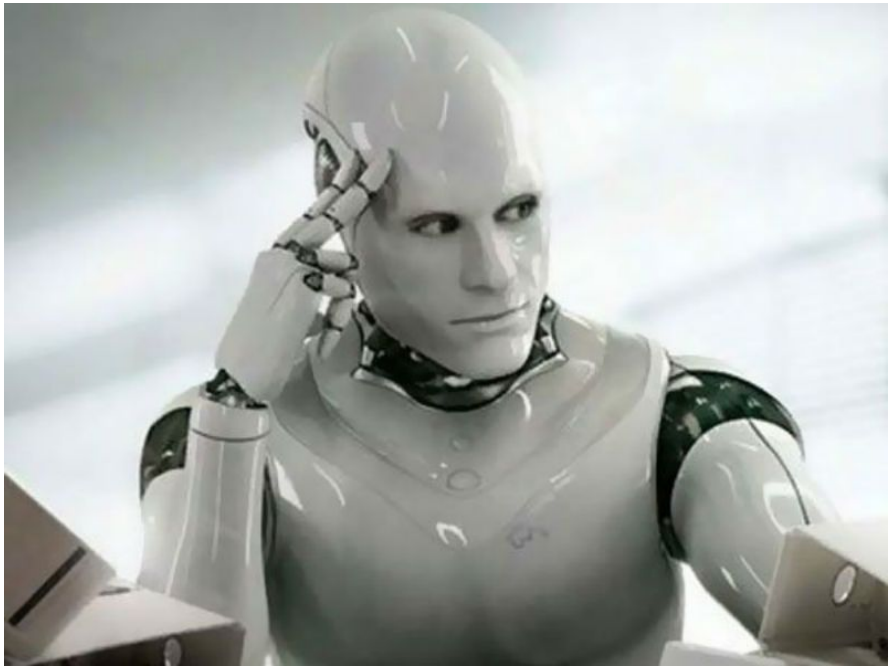


TESTO

SUONI

Formati di dati non strutturati

Un computer conosce solo numeri,
queste informazioni vanno codificate



Immagini

Un’immagine è una matrice di pixel

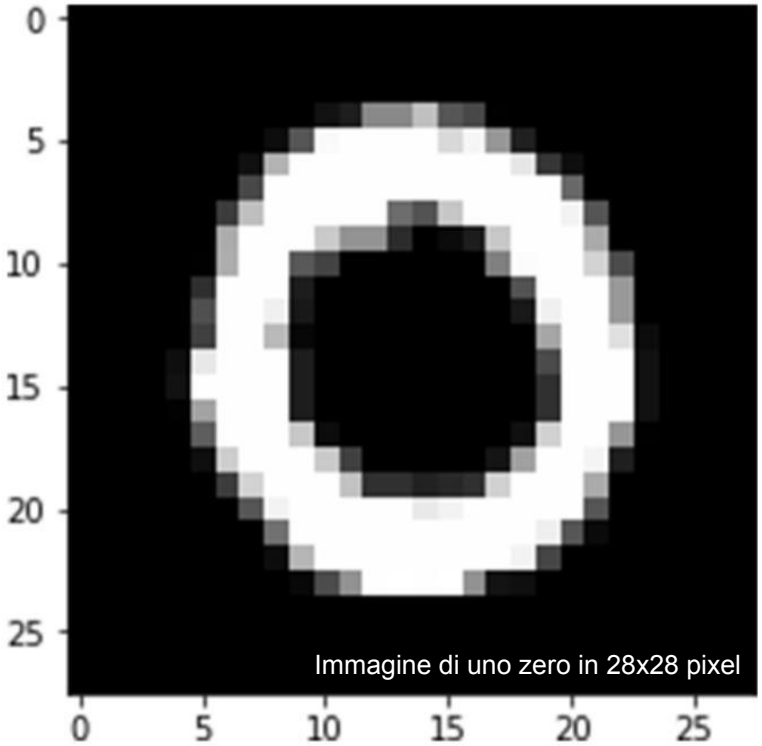
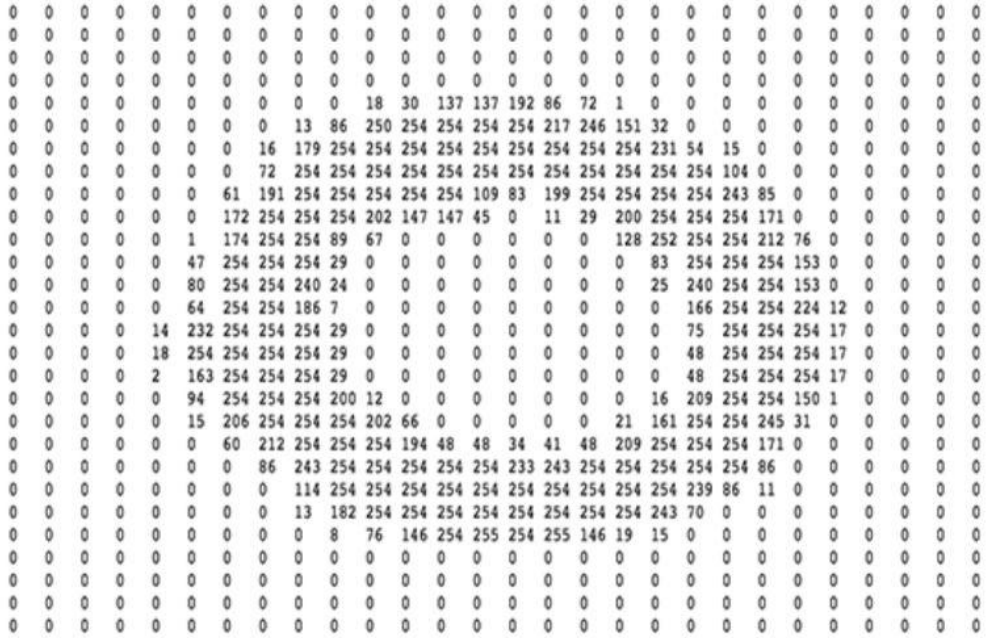


Immagine di uno zero in 28x28 pixel

Immagini

Un pixel è un valore che va da 0 a 255



B/W = 1 matrice RGB = 3 matrici (una per canale)

Flattening

Ridimensionamento da matrice a vettore

$$\mathbf{M} = \begin{bmatrix} 62 & 24 & 15 & 9 \\ 59 & 12 & 81 & 34 \\ 11 & 44 & 32 & 73 \\ 88 & 30 & 3 & 59 \end{bmatrix}$$

Matrice $\mathbf{M}$ di dimensione 4x4

Una matrice viene ridimensionata in un vettore spacchettando le sue righe (o colonne) in un unica riga (o colonna)

$$\mathbf{v} = [\overbrace{62 \quad 24 \quad 15 \quad 9}^{\text{riga 1 di } \mathbf{M}} \quad \overbrace{59 \quad 12 \quad 81 \quad 34}^{\text{riga 2 di } \mathbf{M}} \quad \overbrace{11 \quad 44 \quad 32 \quad 73}^{\text{riga 3 di } \mathbf{M}} \quad \overbrace{88 \quad 30 \quad 3 \quad 59}^{\text{riga 4 di } \mathbf{M}}]$$

Matrice $\mathbf{M}$ ridimensionata in un vettore $\mathbf{v}$ di dimensione 16

Testo - Bag of words

E’ il conteggio dell’occorrenza dei singoli termini

La mamma, la nonna e la zia preparano la cena per il compleanno della zia

la	mamma	nonna	e	zia	preparano	cena	per	il	compleanno	della
4	1	1	1	2	1	1	1	1	1	1

Testo - TF*IDF

Da più peso ai termini più rari e penalizza quelli più comuni

Term Frequency

Misura la frequenza di ogni termine in un documento

X

Inverse Document Frequency

Misura l'importanza di ogni termine all'intero dell'intero corpus

Testo - TF*IDF

Corpus di Testo

Documento 1: La mamma e la nonna preparano il pranzo.
Documento 2: Il papà e il nonno guardano la partita.
Documento 3: Il leone e la gazzella si guardano.

Testo - TF*IDF

Term Frequency

E’ la frequenza di un termine all’interno di un documento

Testo - TF*IDF

Term Frequency

E’ la frequenza di un termine all’interno di un documento

Documento 1: La mamma e la nonna preparano il pranzo.

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
2/8	1/8	1/8	1/8	1/8	1/8	1/8	0/8	0/8	0/8	0/8	0/8	0/8	1/8

Testo - TF*IDF

Term Frequency

E’ la frequenza di un termine all’interno di un documento

Documento 1: La mamma e la nonna preparano il pranzo.

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
0.25	0.12	0.12	0.12	0.12	0.12	0.12	0	0	0	0	0	0	0.12

Testo - TF*IDF

Document Frequency

Quanti documenti contengono una determinata parola?

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
3	1	3	1	1	3	1	1	1	2	1	1	1	2

Testo - TF*IDF

Inverse Document Frequency

$$\text{IDF} = log\left(\frac{\text{Numero di documenti}}{\text{Document Frequency}}\right)$$

Testo - TF*IDF

Inverse Document Frequency

$$IDF = log\left(\frac{\text{Numero di documenti}}{\text{Document Frequency}}\right)$$

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
0	1.1	0	1.1	1.1	0	1.1	1.1	1.1	0.4	1.1	1.1	1.1	0.4

Testo - TF*IDF

TF Term Frequency

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
0.25	0.12	0.12	0.12	0.12	0.12	0.12	0	0	0	0	0	0	0.12

X

IDF Inverse Document Frequency

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
0	1.1	0	1.1	1.1	0	1.1	1.1	1.1	0.4	1.1	1.1	1.1	0.4

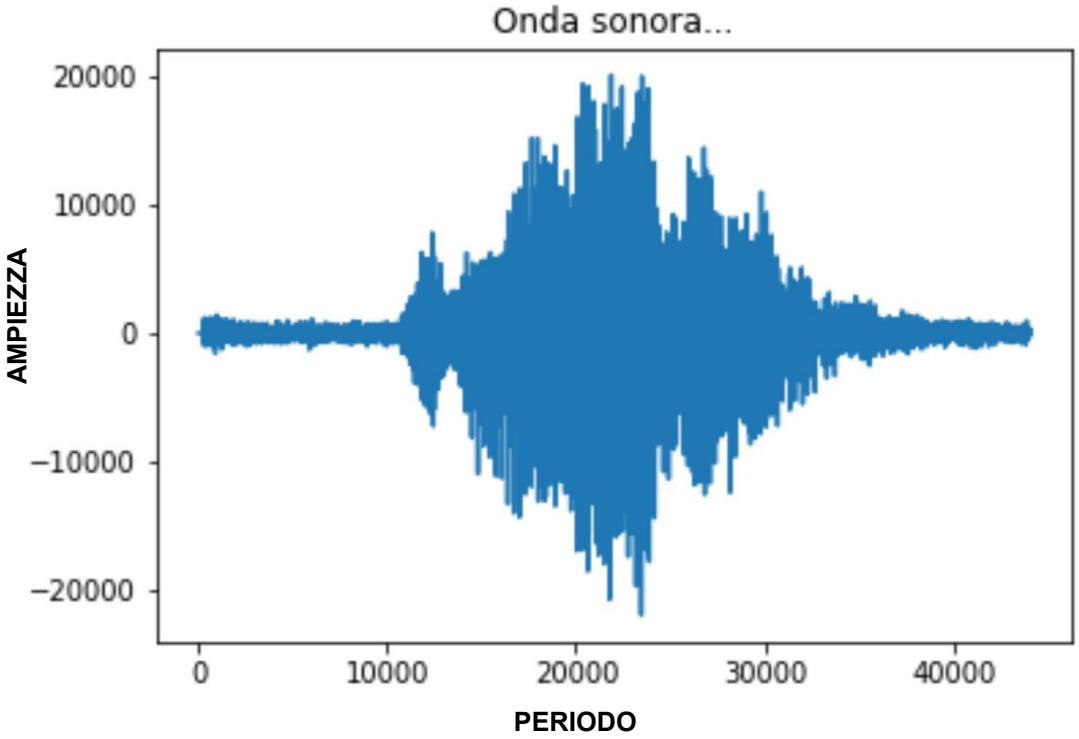
Testo - TF*IDF

TF*IDF

la	mamma	e	nonna	preparano	il	pranzo	papà	nonno	guardano	partita	leone	gazzella	si
0	0.13	0	0.13	0.13	0	0.13	0	0	0	0	0	0	0.05

Suoni

Si usa l'onda sonora



Suoni

Nei casi più semplici l'input è l'array dell'ampiezza

```
[  3865   4298   5350   6723   7399   7463   7918   8989   9722   9586
   9451  10040  10640  10494  10175  10274  10121   9062   7652   6611
   5817   4853   3774   2925   2287   1420    211   -941  -1889  -2925
  -4154  -5362  -6352  -7103  -7813  -8733  -9814 -10720 -11388 -12152
 -13129 -14100 -14947 -15659 -16208 -16674 -17037 -16972 -16506 -16072
 -15337 -13556 -11247  -9626  -8496  -6625  -4072  -2085   -905    397
   1906   2756   3134   4189   5951   7353   8250   9545  11409  13190
  14701  15975  16678  16850  16893  16466  14936  12856  11280   9956
   8128   6346   5534   5207   4308   3142   2819   3144   2756   1282
   -360  -1758  -3440  -5512  -7370  -8721  -9686 -10174 -10016  -9309]
```

Fondamenti di Machine Learning

il Dataset

Tipi di Variabili

presentato da
Giuseppe Gullo

PROFESSION AI

T-shirt disponibili in un negozio di abbigliamento

CODICE	COLORE	TAGLIA	PREZZO
001	Rosso	S	9.90
002	Bianco	M	14.90
003	Verde	L	12.90
004	Rosso	XL	14.90

Terminologia popolare

Tabella

CODICE	COLORE	TAGLIA	PREZZO
001	Rosso	S	9.90
002	Bianco	M	14.90
003	Verde	L	12.90
004	Rosso	XL	14.90

Riga

Colonna

Terminologia statistica

Campione

CODICE	COLORE	TAGLIA	PREZZO
001	Rosso	S	9.90
002	Bianco	M	14.90
003	Verde	L	12.90
004	Rosso	XL	14.90

Osservazione

Variabile

Terminologia machine learning

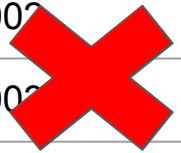
Dataset

CODICE	COLORE	TAGLIA	PREZZO
001	Rosso	S	9.90
002	Bianco	M	14.90
003	Verde	L	12.90
004	Rosso	XL	14.90

Esempio (Sample)

Caratteristica (Feature)

CODICE	COLORE	TAGLIA	PREZZO
001	Rosso	S	9.90
002	Bianco	M	14.90
003	Verde	L	12.90
004	Rosso	XL	14.90



La colonna con codice/id spesso viene scartata, perché non contiene informazioni utili.

COLORE	TAGLIA	PREZZO
Rosso	S	9.90
Bianco	M	14.90
Verde	L	12.90
Rosso	XL	14.90

Le 3 colonne rappresentano
3 tipi di variabili differenti

Variabili quantitative continue

PREZZO
9.90
14.90
12.90
14.90

E’ un numero che può assumere qualsiasi valore rappresenta una quantità.

Variabili qualitative ordinate

TAGLIA
S
M
L
XL

rappresentano un insieme finito di valori, possono essere ordinate.

Variabili qualitative sconnesse (o categoriche)

COLORE
Rosso
Bianco
Verde
Rosso

rappresentano un insieme finito di valori, non possono essere ordinate.

Un computer capisce solo numeri

Bisogna codificare le variabili qualitative