Data Journalism

Analisi dei dati

Angelica Lo Duca
angelica.loduca@iit.cnr.it

Obiettivo

L’obiettivo dell’analisi dei dati consiste nello scoprire **trend**, **pattern** e **relazioni** nascosti nei dati.

Tipo di analisi

Analisi Quantitativa

Analizza una grande quantità di dati al fine di estrarne il comportamento e l’andamento.

Analisi Qualitativa

Analizza una piccola quantità di dati al fine di estrarne alcuni dettagli.

Metodo di analisi

Manuale

E’ molto preciso, ma richiede molto tempo. Inoltre può essere fatto solo su una piccola quantità di dati.

Semiautomatico

Utilizza degli strumenti (ad esempio basati su Visual Analytics). Necessita comunque della supervisione del data journalist.

Automatico

Basato sulla scrittura di codice. Permette di analizzare grandi quantità di dati, ma non è preciso come i due casi precedenti.

Tipi di Statistica

Descrittiva

Dato un insieme di dati (campione di dati), cerca di descrivere il campione stesso.

Inferenziale e Machine Learning

Considera il campione di dati come un sottoinsieme di una popolazione. Cerca di capire il comportamento della popolazione a partire dal campione.

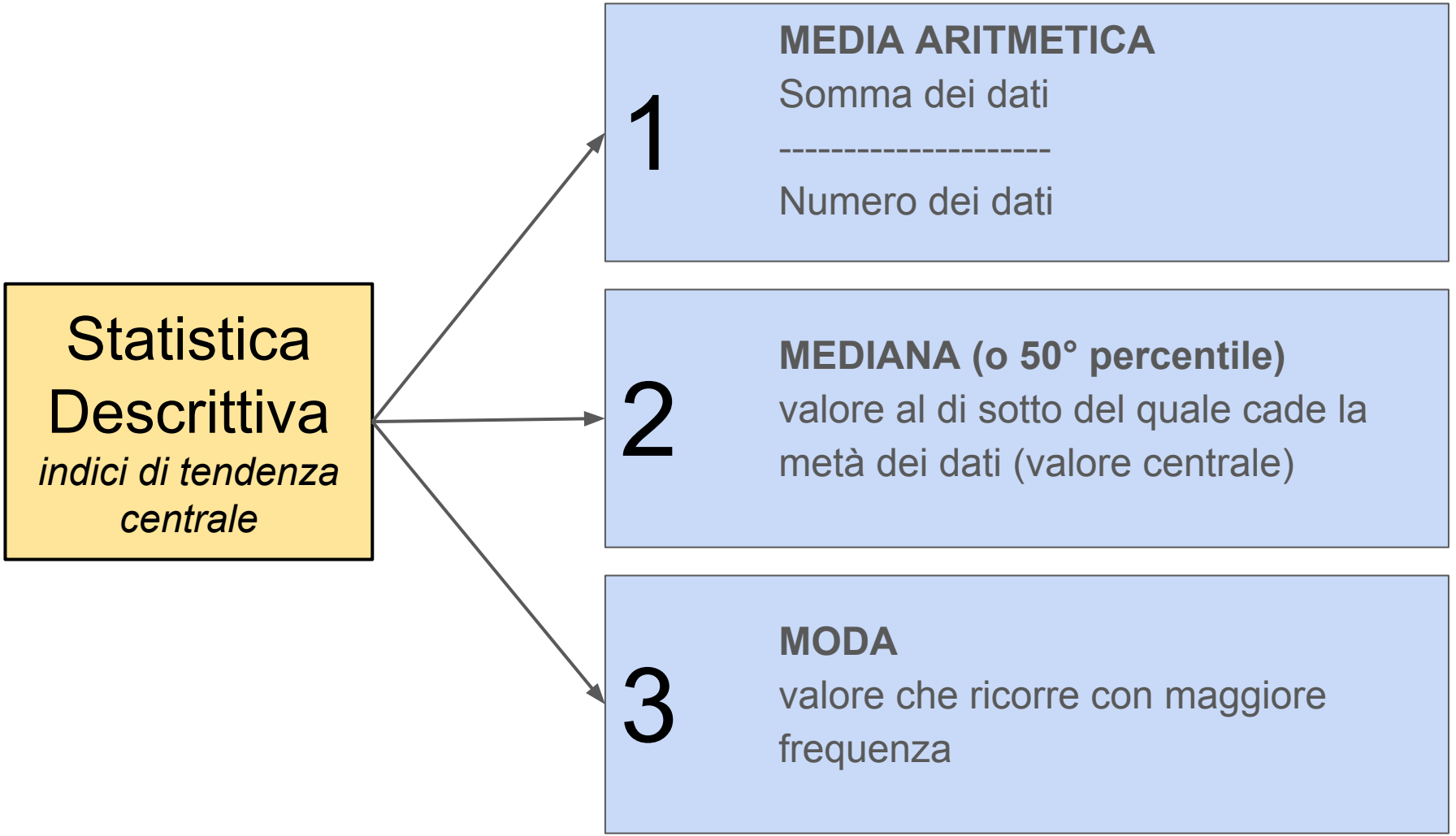
Statistica descrittiva

Si basa sul calcolo di alcune **metriche** o **indici**

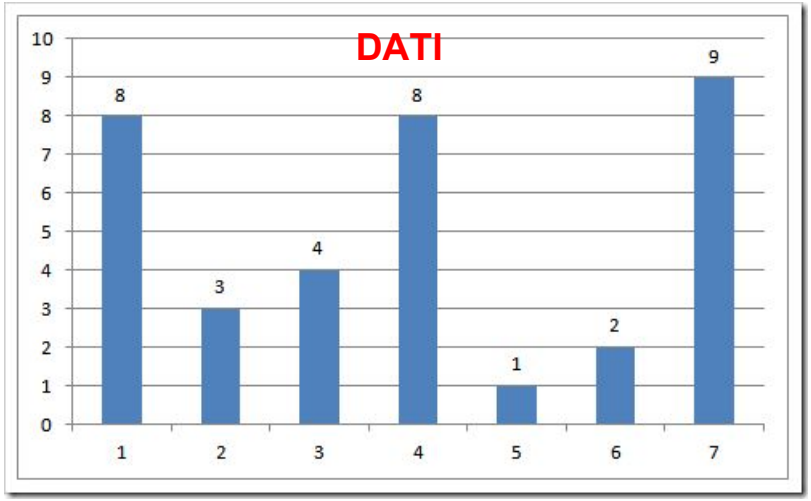
- Indici di tendenza centrale
- Indici di dispersione

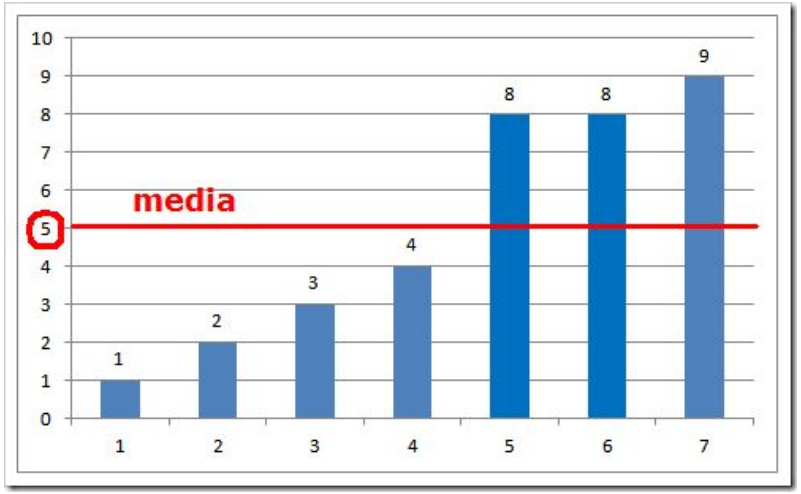
Presentazione grafica

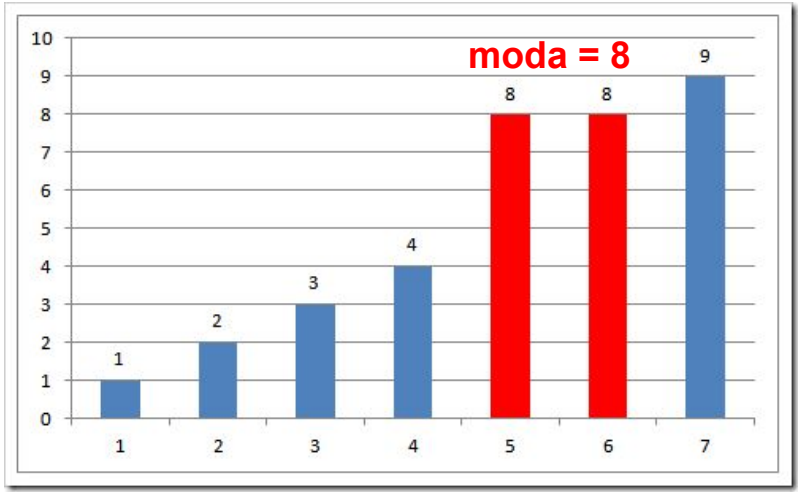
grafici a barre, linee, diagrammi a torta, distribuzione della frequenza

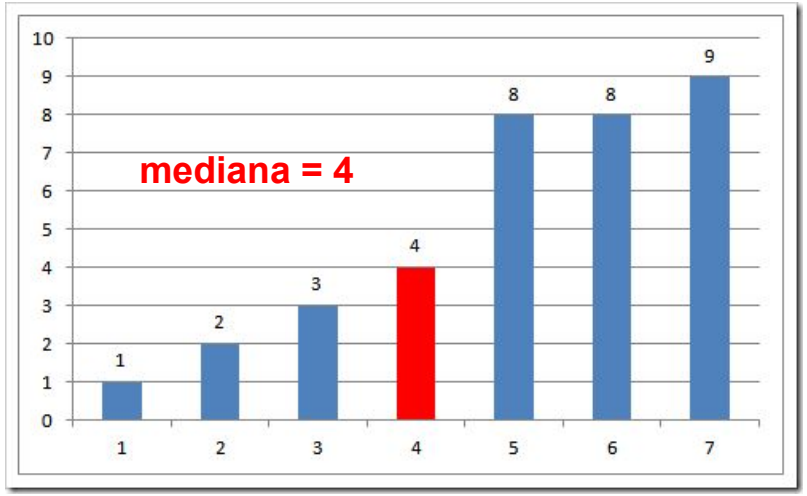

Statistica Descrittiva
indici di tendenza centrale
1
MEDIA ARITMETICA
Somma dei dati

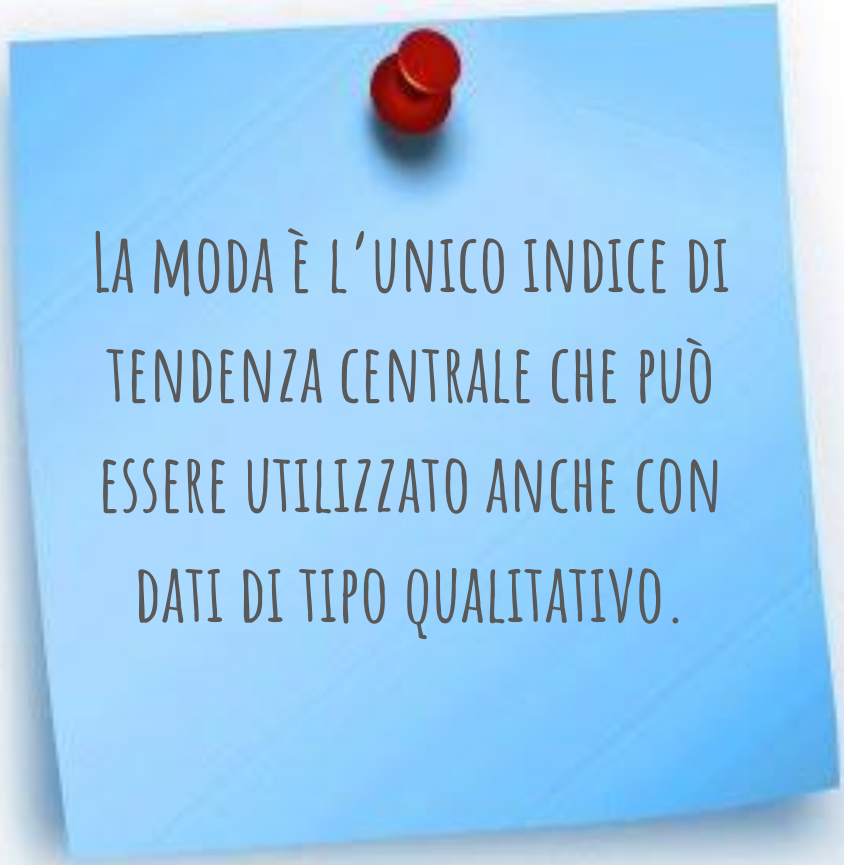
Numero dei dati
2
MEDIANA (o 50° percentile)
valore al di sotto del quale cade la metà dei dati (valore centrale)
3
MODA
valore che ricorre con maggiore frequenza


DATI
10
9
8
7
6
5
4
3
2
1
0
8
3
4
8
1
2
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
media
1
2
3
4
8
8
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
moda = 8
1
2
3
4
8
8
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
mediana = 4
1
2
3
4
8
8
9
1
2
3
4
5
6
7


La moda è l'unico indice di tendenza centrale che può essere utilizzato anche con dati di tipo qualitativo.

Indici di dispersione

scarto medio assoluto, varianza, deviazione standard, coefficiente di variazione e deviazione media assoluta. scarto interquartile. indice di dispersione di Poisson

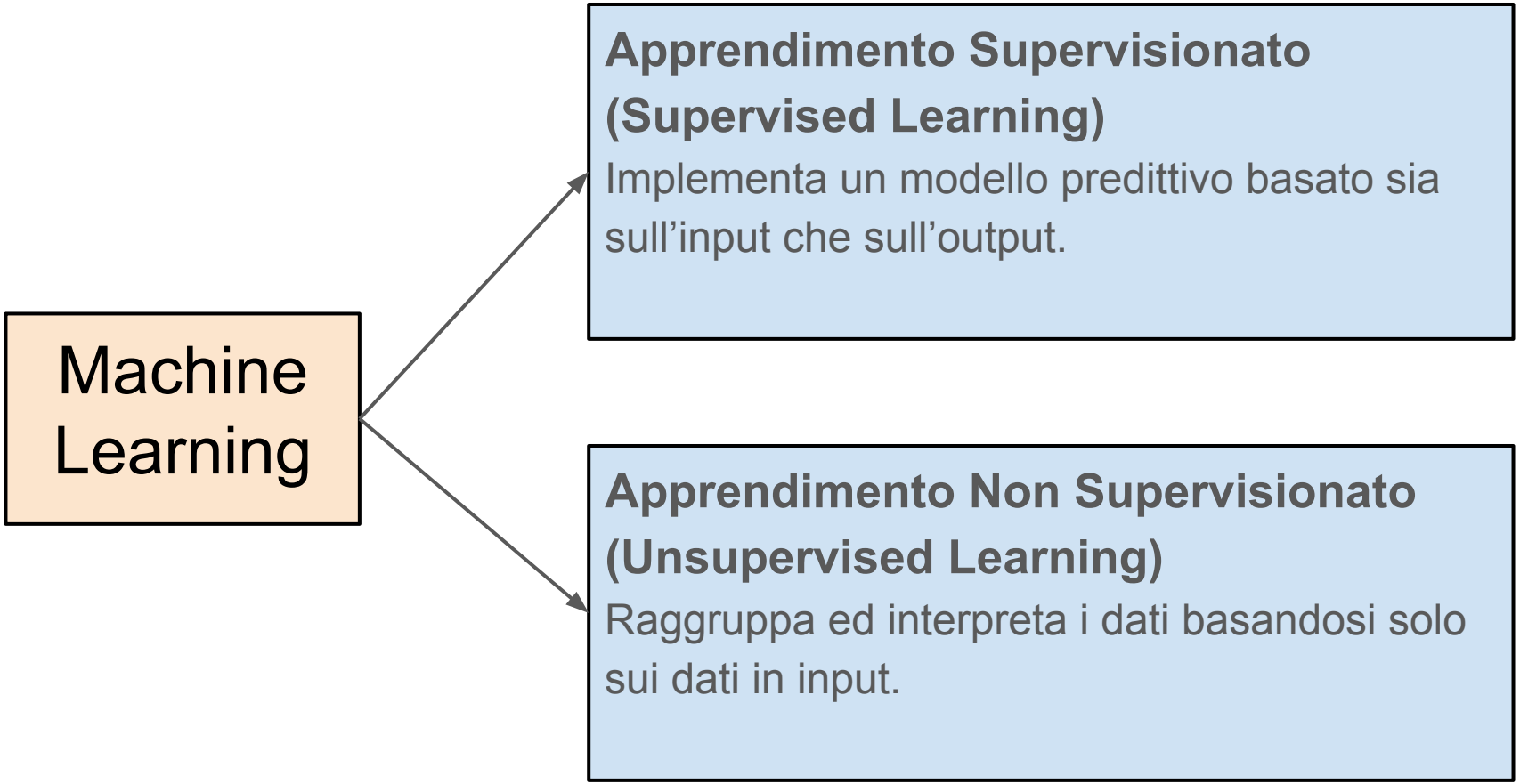
Statistica Inferenziale e Machine Learning

La Statistica Inferenziale si sovrappone al Machine Learning

Statistica inferenziale - figlia della matematica

Machine Learning - figlia dell’informatica

Essenzialmente i concetti che descrivono e analizzano sono gli stessi


Machine Learning
Apprendimento Supervisionato (Supervised Learning)
Implementa un modello predittivo basato sia sull'input che sull'output.
Apprendimento Non Supervisionato (Unsupervised Learning)
Raggruppa ed interpreta i dati basandosi solo sui dati in input.

Apprendimento Supervisionato

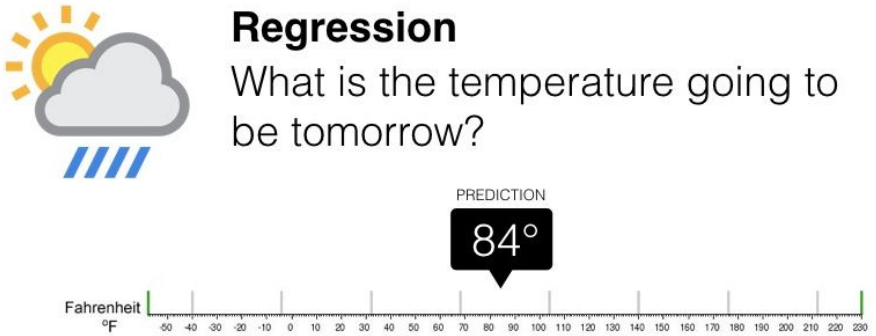
L'obiettivo è di **approssimare** la funzione F così bene che quando si hanno nuovi dati di input si possono predire le variabili di output per quei dati.

La funzione F è **allenata** su un campione di dati (*training set*).

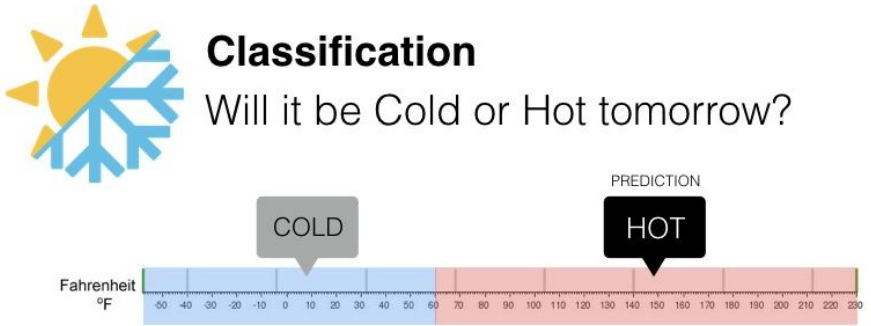
L'apprendimento si interrompe quando la funzione raggiunge un livello accettabile di prestazioni.

Apprendimento Supervisionato (2)

Regressione
L'output prodotto è numerico (continuo)



Classificazione
L'output prodotto corrisponde a categorie (discreto)

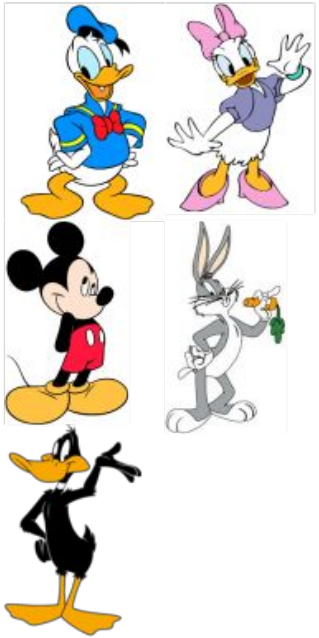


Apprendimento Non Supervisionato

L'obiettivo è di **modellare** la struttura o la distribuzione sottostante nei dati per ottenere ulteriori informazioni sui dati.

Clusterizzazione
Raggruppa i dati in modo tale che i dati nello stesso gruppo siano più simili di quelli appartenenti ad altri gruppi.

Associazione
Cerca delle regole di associazione e correlazione all'interno di grandi quantità di dati.



Clusterizzazione

Regole di associazione

L'idea deriva dalla **Market Basket Analysis** (MBA)

Domande

Cosa acquisterà il cliente 4?

Quali prodotti vengono acquistati insieme?

Latte, Uova, Zucchero, Pane

Cliente 1

Latte, Uova, Cereali, Pane

Cliente 2

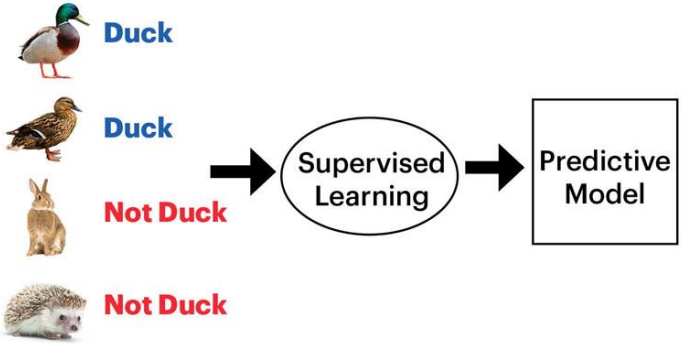
Uova, Zucchero

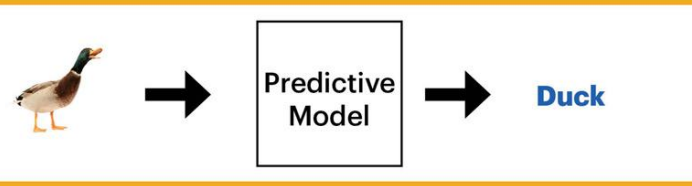
Cliente 3

?

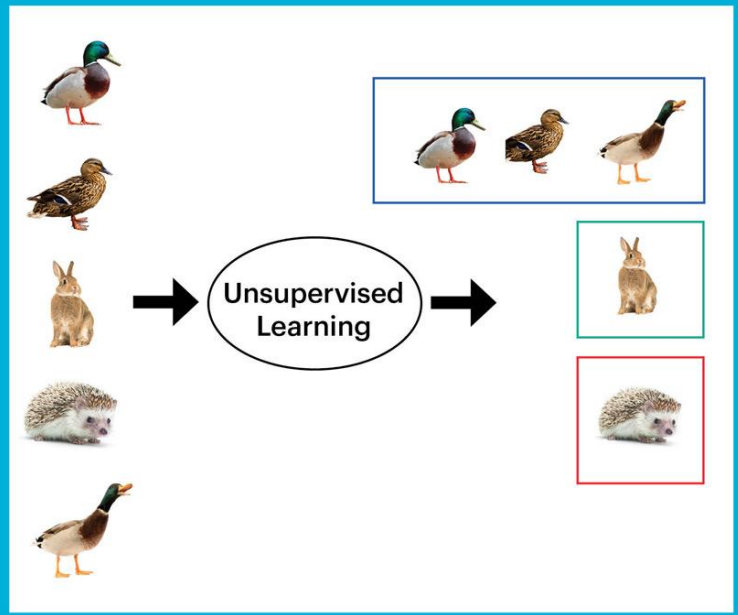
Cliente 4

Supervised Learning (Classification Algorithm)





Unsupervised Learning (Clustering Algorithm)



Western Digital.

Strumenti per l’analisi dei dati

Weka

R

scikit-learn (libreria python)

...

Esercizio

Dato un dataset di coppie contenenti potenziali calciatrici (nome utente, account twitter), capire dall'analisi dei tweets pubblicati, se un utente è effettivamente una calciatrice o no.

Soluzione

Utilizzando la libreria **twitter-python**, per ogni account estrarre da Twitter i tweets e salvarli in un file CSV.

(Segue…)

Alice Parisi @aliceparisi18 · 1 gen 2016

E su inizia così #unogennaio #2016 #Capodanno2016 #Buon2016 #BuonAnno

Calciatrice

Non calciatrice

Alice Parisi @ParisiAlice · 6 mar 2013

Cyprus Cup....ore 17.30 (16.30 italiane) ITALIA-INGHILTERRA..

1 3 1

E' un problema di classificazione

Procedimento

*Dato il dataset dei tweets (7446), ne estraggo il **30% (2234)** e li **annoto a mano***

Annotare a mano significa che marco ogni tweet come calciatrice (1) o non calciatrice (0).

Ho ottenuto il **campione** dei dati, contenente le coppie (input, output) necessarie per allenare e testare il classificatore.

Procedimento (2)

*Divido il campione in due parti, **training set** (80%) e **test set** (20%)*

Training Set	**Test Set**
Usato per allenare il classificatore.	Usato per verificare le prestazioni del classificatore, cioè quanto riesce a predire correttamente se un tweet parla di calcio femminile o no.

Procedimento (3)

Converto i dati nel formato giusto

1. Divido ogni tweet in **token**
2. Assegno un id intero ad ogni token t presente in qualsiasi tweet del campione (ad esempio **costruendo un dizionario** da parole a indici interi)
3. Per ogni tweet i, conto il numero di occorrenze di ogni token t

Scelgo il classificatore e lo alleno sul training set

Bernoulli, Gaussian, ...

Procedimento (4)

Verifico la bontà del classificatore, testandone le prestazioni sul test set

Confronto Y_pred con il valore effettivo annotato a mano

Esistono diverse metriche per misurare la bontà del classificatore (precision, recall, accuracy)

Procedimento (5)

Se i risultati sono buoni, uso il classificatore sul rimanente 70% dei dati