MACHINE LEARNING

Concetti di base

Stefano Palessandro
Data Scientist @ SORINT.tek

SORINT.TEK

mangrovia

PARTOF.SORINT.LAB

Programma

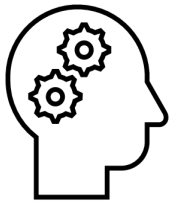
- Machine Learning: Cos'è?
- Caratteristiche e come visualizzarle
- Valutazione del modello
- Alcuni algoritmi di ML
- Come usiamo il ML in Sorint.tek?
- Demo Python

Machine Learning: Cos'è?

- Un processo che permette a un sistema informatico di imparare dall'"esperienza" a scovare pattern nascosti nei dati.
- Invece di programmare un computer tramite regole specifiche, gli si insegna a riconoscere da solo quali sono queste regole ("modello").
- Le regole vengono apprese automaticamente dal modello di Machine Learning da un insieme di dati ("esperienza") di input.
- Una volta apprese le regole, il modello è in grado di produrre predizioni su dati che non ha mai visto!

Intelligenza Artificiale Debole

- Il ML quindi consiste nel predire eventi che il modello non ha osservato.
- È una forma di intelligenza:
  - riceve dati di input;
  - apprende automaticamente dei pattern nei dati;
  - esprime giudizi sui dati nuovi sulla base dei pattern appresi.



Programmazione Classica vs ML

```
if email contains
"Offerta speciale!!!"
then label as spam;
if email contains
"Guadagna denaro!!"
then label as spam;

if email contains …
then label as spam;
```

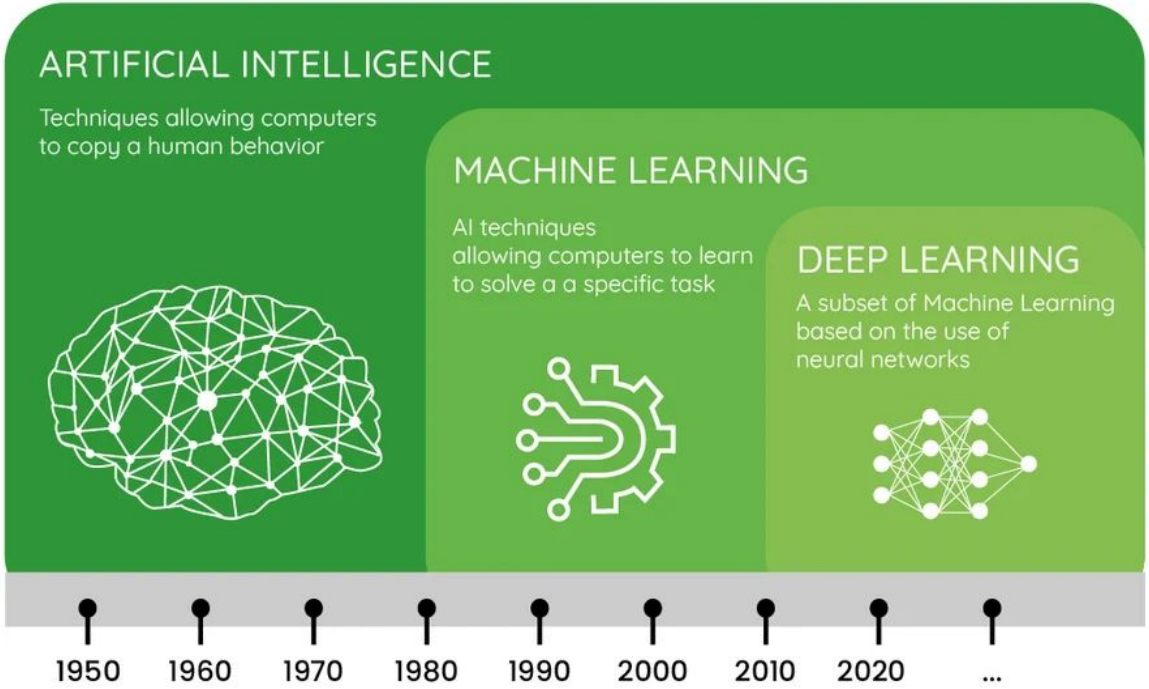
```
Etichetta alcune e-mail come
spam o non-spam;

Allena il modello dandogli in
input le e-mail etichettate;

Il modello apprende a
distinguere le e-mail spam e si
autocorregge per evitare errori;

Ripeti;
```

Contesto


ARTIFICIAL INTELLIGENCE
Techniques allowing computers to copy a human behavior
MACHINE LEARNING
AI techniques allowing computers to learn to solve a a specific task
DEEP LEARNING
A subset of Machine Learning based on the use of neural networks
1950
1960
1970
1980
1990
2000
2010
2020
...

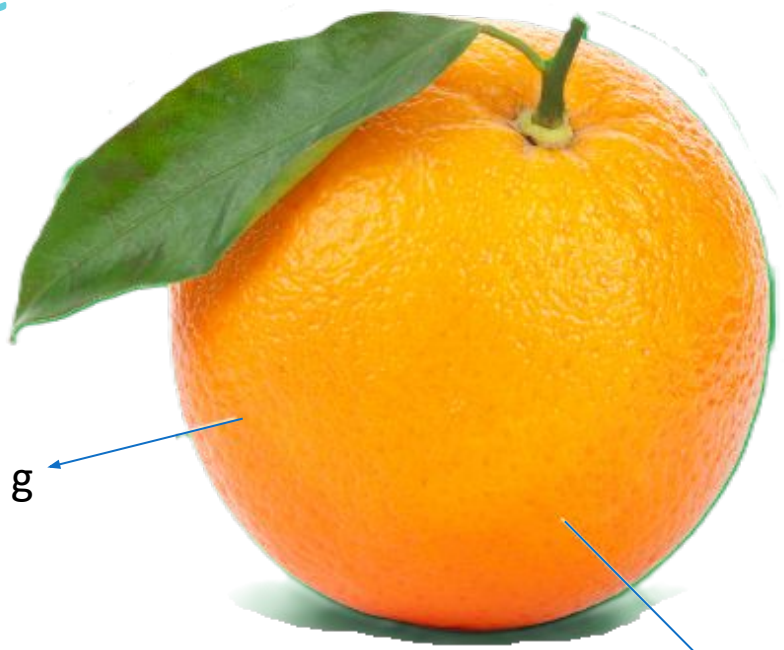
Caratteristiche e come visualizzarle

Caratteristiche

Bisogna selezionare gli attributi più rilevanti per il problema.

Peso?

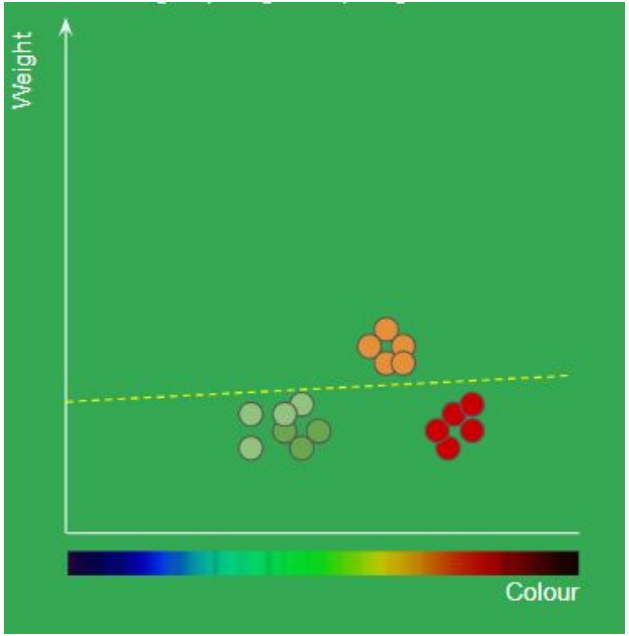
Colore?



Come apprende la macchina?

2 caratteristiche = 2 dimensioni (x e y).
Possiamo rappresentare le caratteristiche su un piano cartesiano.

Il modello apprende a dividere le mele dalle arance (linea in figura).
Se mostriamo al modello nuovi esempi di mele o arance, saprà come classificarli.

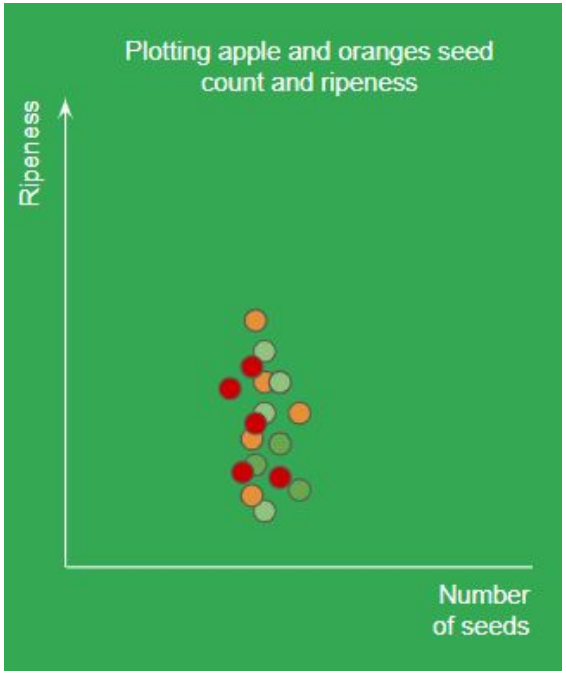


Credits: Machine Learning 101 by Jason Mayes

L'importanza delle caratteristiche

Queste caratteristiche non sono il massimo!

La maturità e il numero di semi del frutto infatti non ci permettono di distinguere tra arancia e mela. L'abilità dell'esperto ML sta spesso nel trovare caratteristiche sensate.



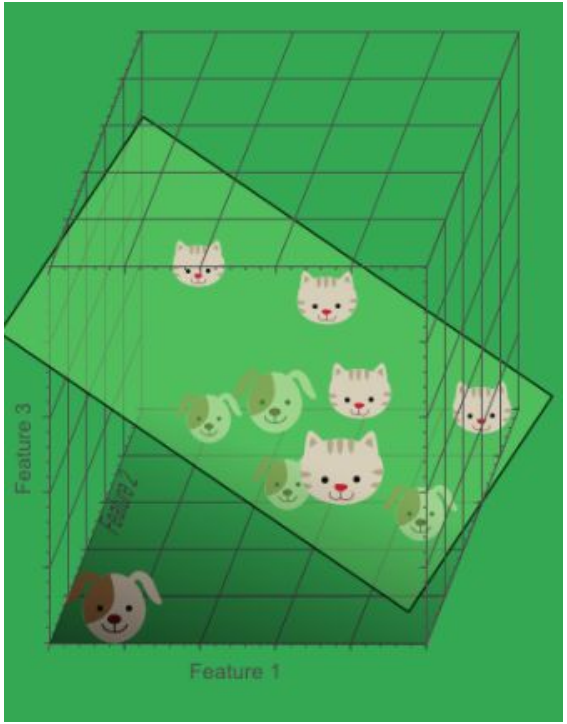
Credits: Machine Learning 101 by Jason Mayes

Multidimensionalità

Con 3 caratteristiche avrò 3 dimensioni (x, y e z).

Separo gli oggetti con un piano anziché una retta.

Nota: spesso i problemi di ML hanno una dimensionalità ancora maggiore! 20, 100, 10.000.000 di dimensioni (image recognition: ogni pixel è una dimensione).
Siamo diversi dalle macchine: per noi è difficile da visualizzare, ma i principi matematici sono gli stessi.



Credits: Machine Learning 101 by Jason Mayes

A caccia di dati

- Una volta che si individuano le caratteristiche da usare, la sfida maggiore è trovare dati senza bias in un formato congruo.
- Se voglio classificare gatti, avrò bisogno di decine di migliaia di foto di gatti di ogni razza e provenienza.
- Formati: immagini, testo, sensori, audio, ecc…

Un modello ML non può predire cose che non ha mai visto

Se alleno il mio sistema ML solo con i seguenti dati:

Numero di gambe	Colore	Peso	Animale
4	Nero	10 kg	CANE
2	Arancione	5 kg	POLLO

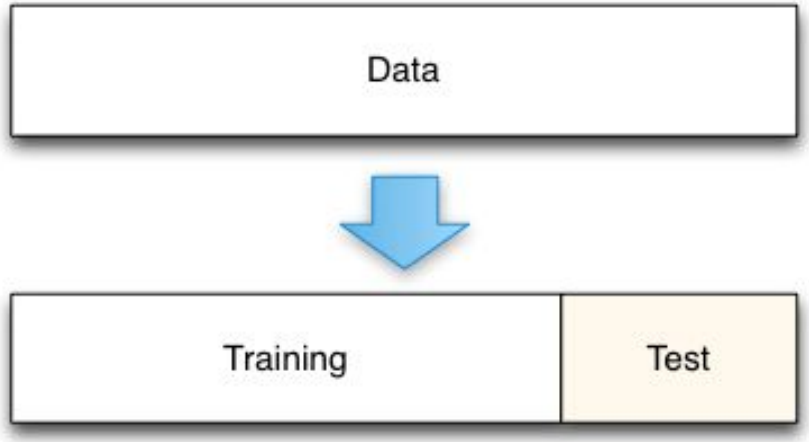
quando vedrà una mucca nera, penserà che si tratti di un cane. Perché conosce solo cani e polli e la corrispondenza più vicina è il cane.

Valutazione del Modello

Metriche di test

Allenamento e test

- I dati si dividono in dati di addestramento (*training*) e dati di valutazione (*test*).
- Lo split di solito è 80% / 20%.
- Il modello viene addestrato con i soli dati di training.
- I dati di test vengono utilizzati solo in fase di valutazione del modello.



Metriche di valutazione

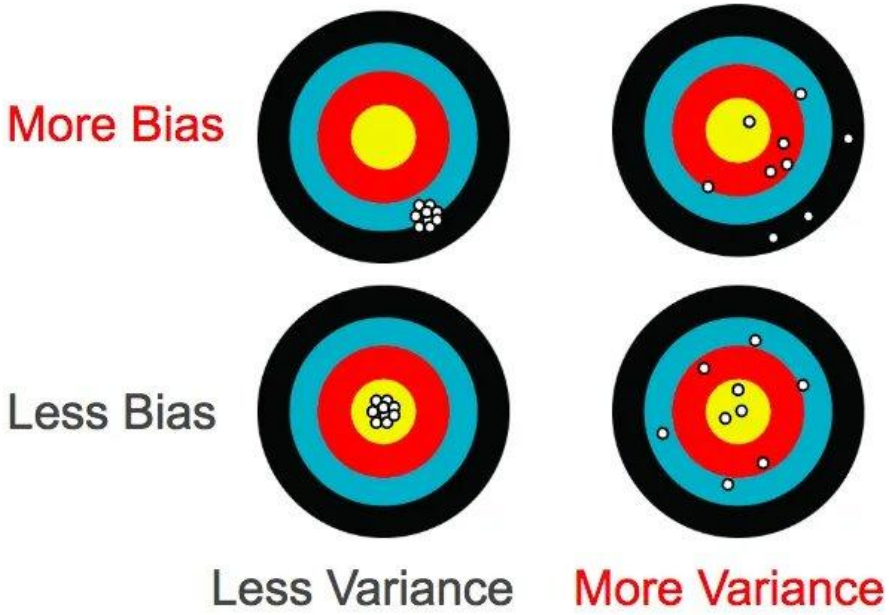
ERRORE = "Distanza" tra valore reale e valore predetto dal modello

- Sono come i voti a un esame. Ogni predizione del modello sui dati di test viene giudicata giusta o sbagliata e viene prodotto poi un voto finale per il modello.
- Possiamo dare più peso ad alcune predizioni, così come si dà più peso ad alcune domande a un esame.

ERRORE = BIAS + VARIANZA

- **Bias** = il modello produce risultati sistematicamente sbagliati.
- **Varianza** = variabilità nelle predizioni del modello.

Bias vs Varianza



Alcuni algoritmi di ML

Supervisionato vs Non supervisionato

Supervisionato

L'insieme di dati su cui alleno il modello è "etichettato".

Ad esempio, nei dati è presente l'informazione "arancia" o "mela" per la classificazione dei frutti.

Non supervisionato

L'insieme di dati su cui alleno il modello non è "etichettato". Il modello deve apprendere da solo, senza aiuti esterni, a categorizzare gli oggetti.

Regressione e Classificazione

Regressione

La variabile di output è numerica (continua).

Esempi: prezzo di una casa, percentuale di errori in un compito in classe...

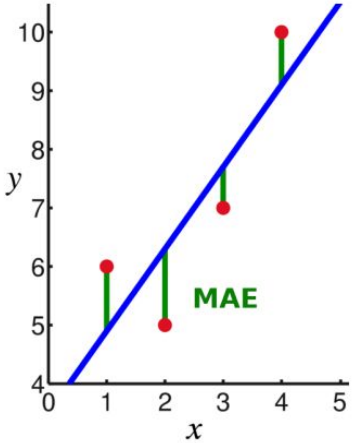
Classificazione

La variabile di output è categoriale (discreta).

Esempi: genere di una persona, animale (gatto, cane...), frutto (mela, banana…)

Alcune metriche di valutazione

Regressione	Classificazione
Errore assoluto medio (MAE)	Accuratezza: % di oggetti classificati correttamente
Errore assoluto percentuale medio (MAPE)	Richiamo: da maggiore peso ai falsi negativi
Scarto quadratico medio (MSE)	Precisione: da maggiore peso ai falsi positivi



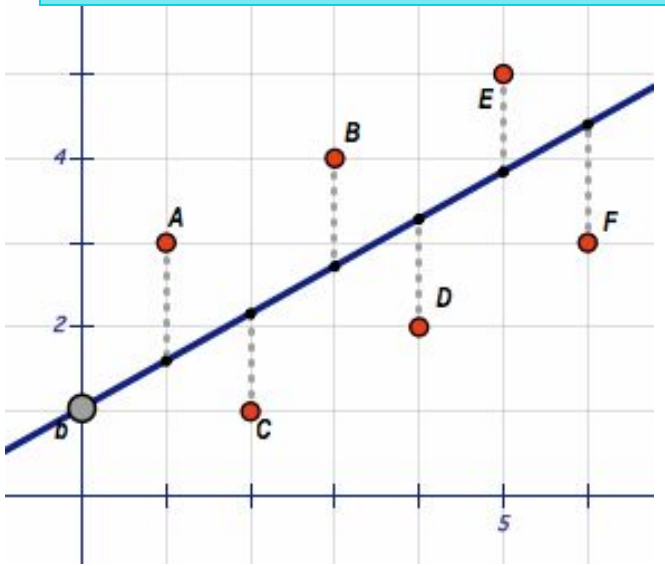
	Malato Covid	Sano
Test positivo	Vero positivo	Falso positivo
Test negativo	Falso negativo	Vero negativo

Regressione lineare

- Si vuole predire una variabile continua y (dipendente) a partire da una (o più) variabili continue x (indipendenti).

- Si assume che la dipendenza di y da x sia di natura lineare (una retta in questo caso).

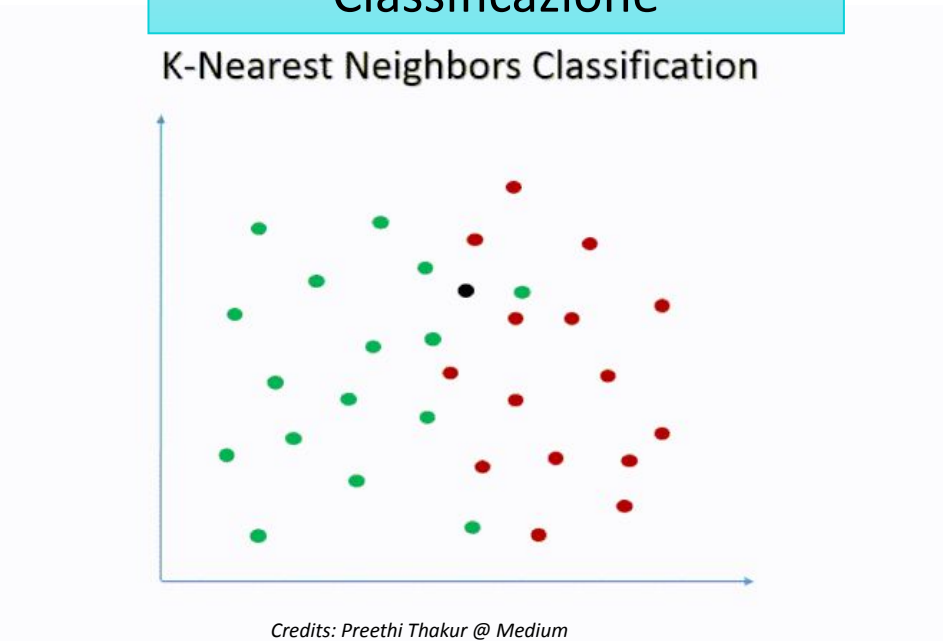
Supervisionato
Regressione



k-Nearest Neighborhood

- Si usa il grado di similarità (distanza) tra il nuovo valore x da predire e i dati osservati
- È democratico! Se la maggioranza di dati nell'"intorno" di x "vota" verde, x sarà predetto come verde
- k è il numero di punti nell'intorno.

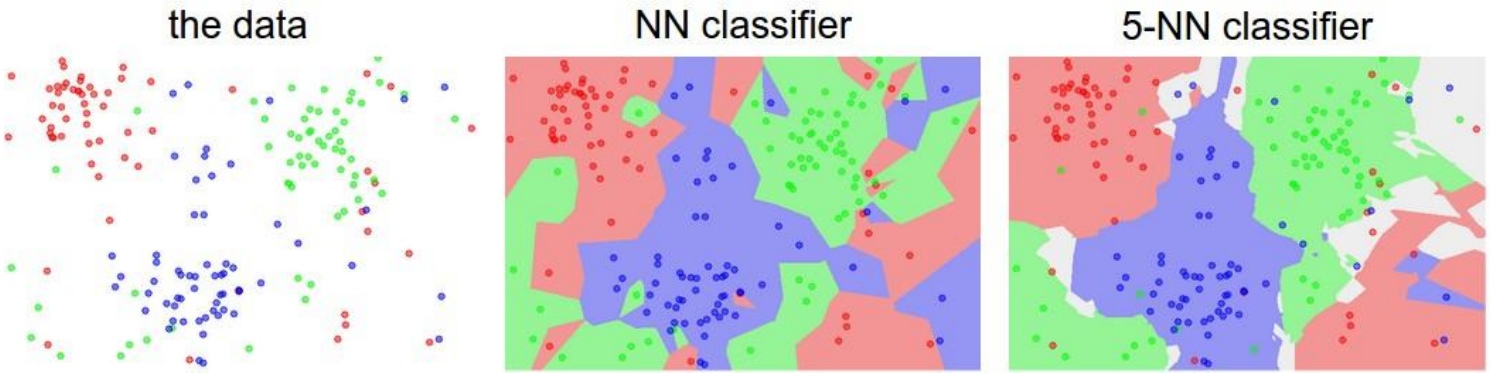
Supervisionato
Classificazione



Credits: Preethi Thakur @ Medium

Decision Boundaries (Confini decisionali)

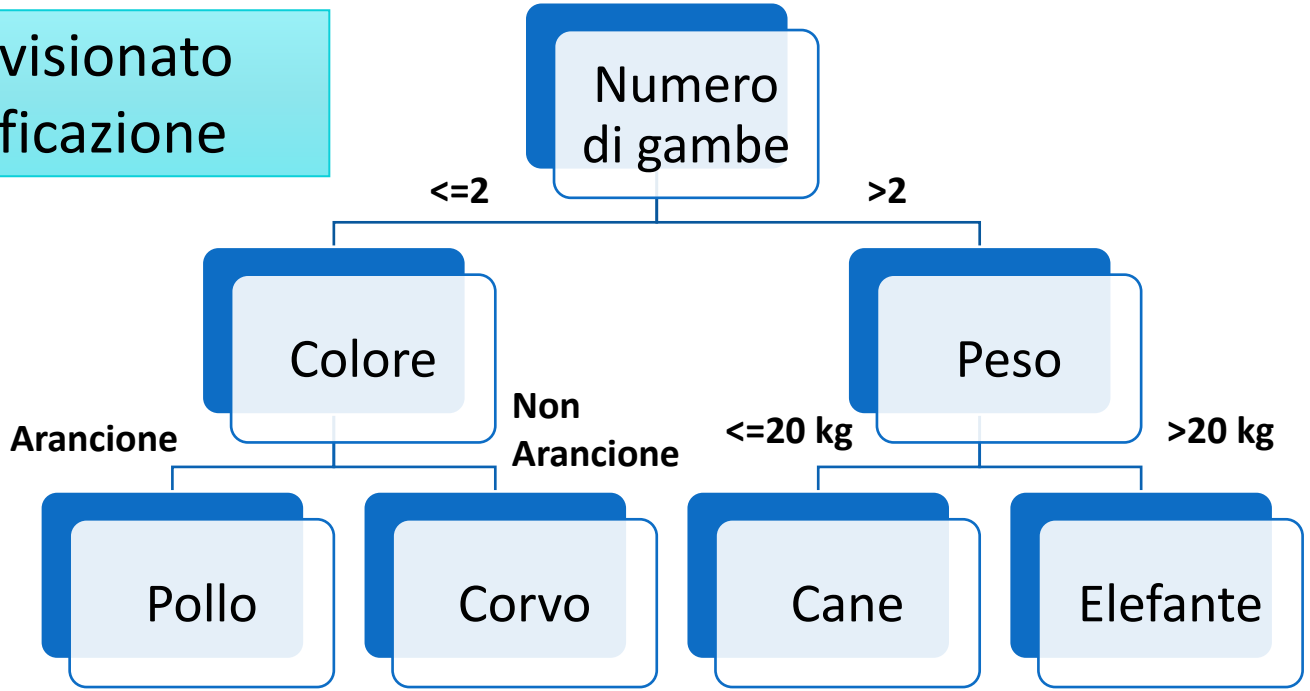
- Sono le linee che separano una classe dall'altra.
- Vengono apprese dal modello.
- Le regioni colorate in figura hanno come contorno i decision boundary.



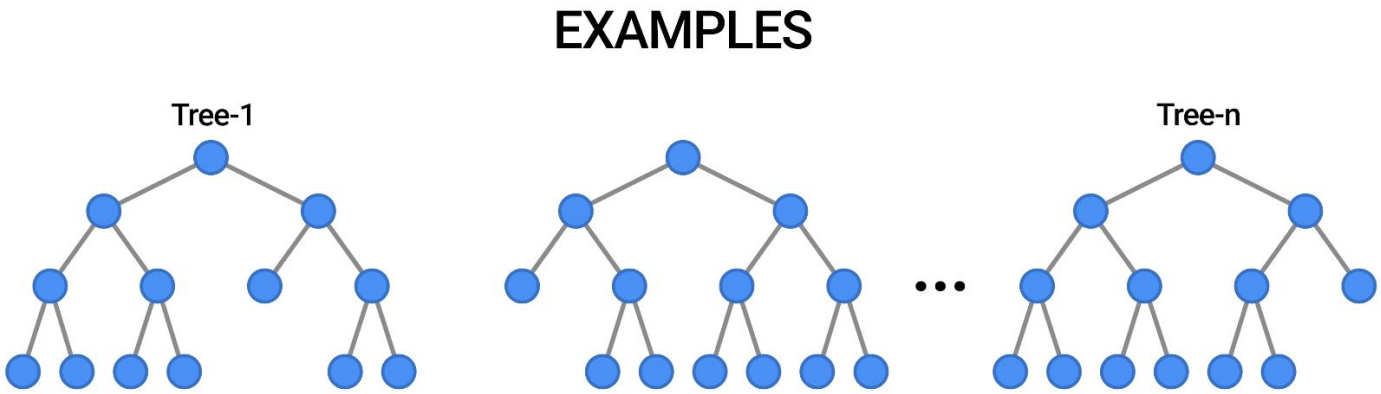
Credits: CS231n: Deep Learning for Computer Vision @ Stanford

Alberi decisionali

Supervisionato
Classificazione



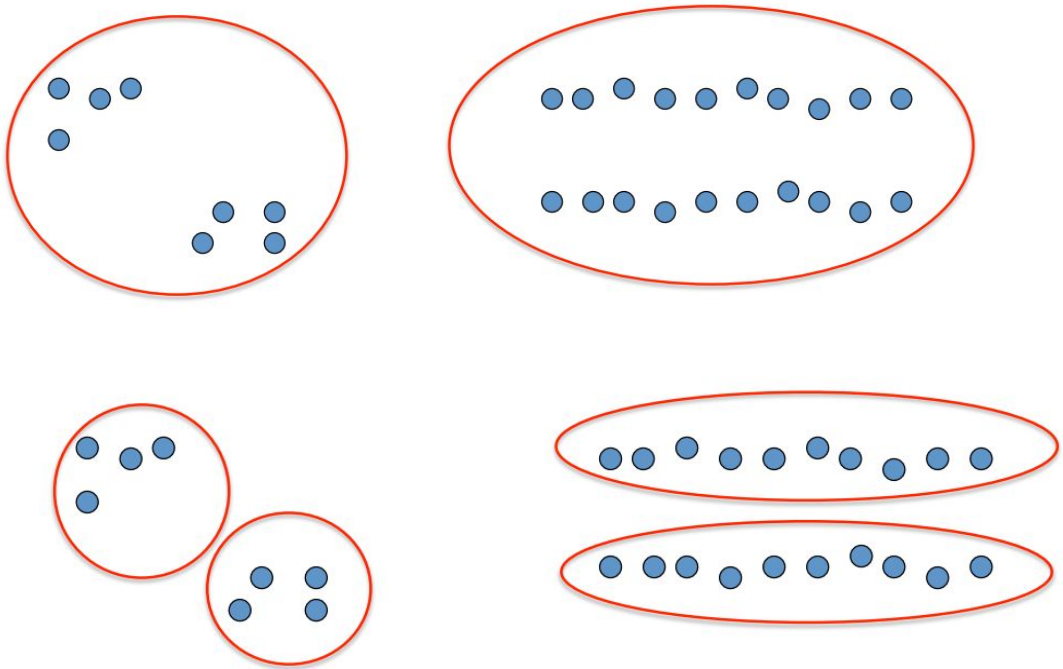
Random Forest



Credits: Tensorflow @Google

Clustering

Non supervisionato

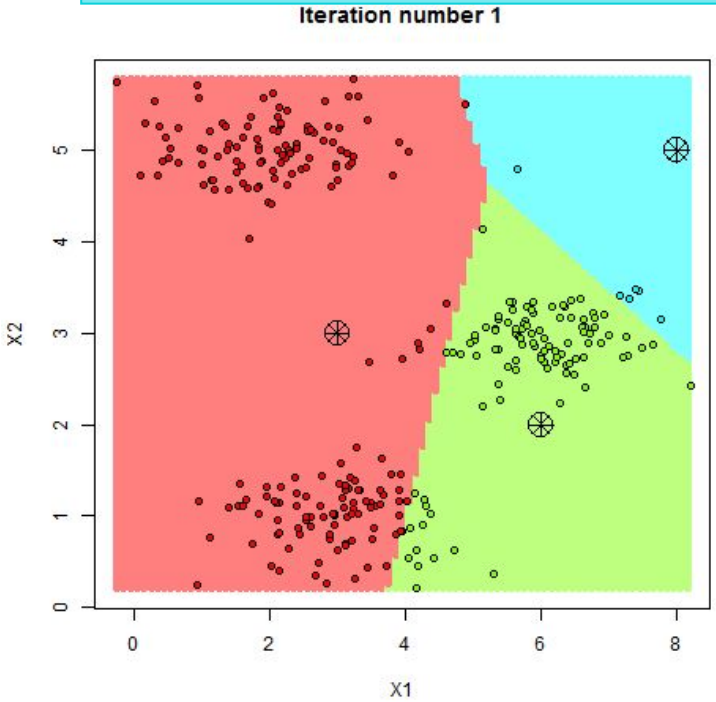


Qual è il raggruppamento corretto?

Clustering: k-Means (k-Medie)

Non supervisionato

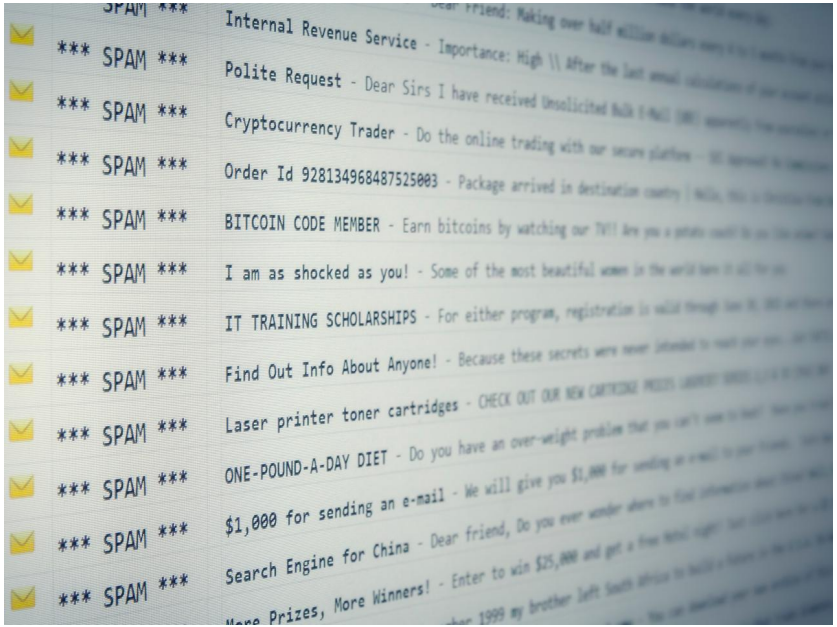
- Ha bisogno di una misura di similarità (esempio: distanza)
- Algoritmo:
  - Scegli k punti casuali come centro del cluster
  - Associa tutti i punti al centro del cluster più simile (vicino)
  - Il centro di ogni cluster diventa la media dei punti di quel cluster



Credits: Sandipan Dey, MS (CSEE), UMBC

Clustering: esempi

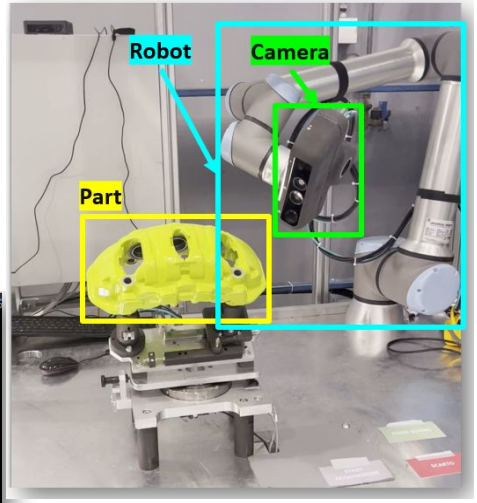
- Categorizzare oggetti come:
  - e-mail;
  - risultati di ricerca sul web;
  - regioni di immagini digitali;
- È utile quando non si ha una conoscenza pregressa di quale informazione categorizzare.



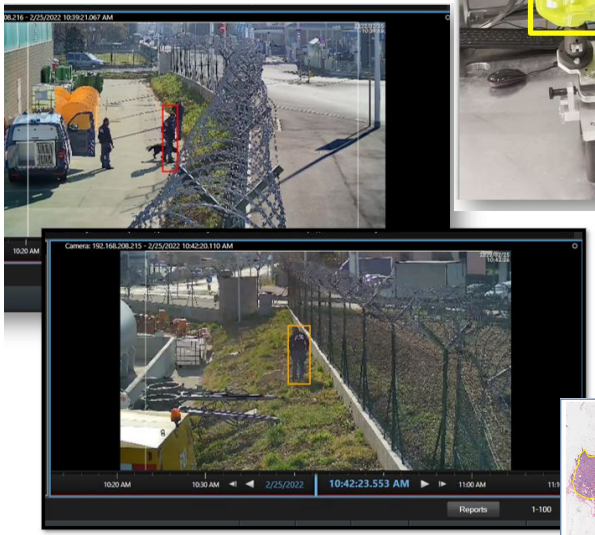
SORINT.TEK


Come usiamo il ML
in Sorint.tek?
SORINT.TEK

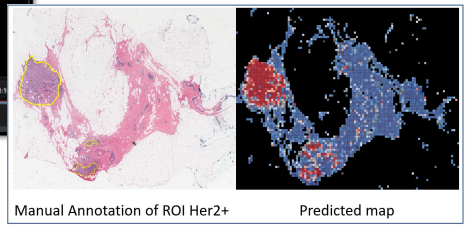
AI per l'Industria Manifatturiera

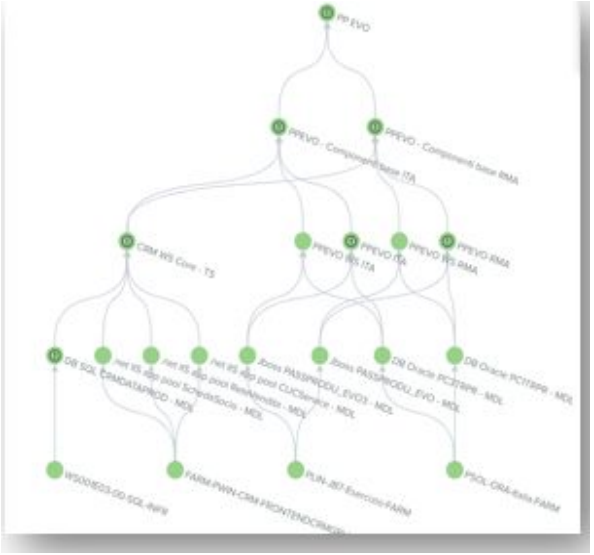


COUNTRY	BUSINESS APP	HEALTH SCORE	PREDICTION
SPA	Multitarificadores		
ITA	PP EVO		
ITA	Pass Contabilità		



AI per la Medicina





AI per monitoraggio in tempo reale dell'Aeroporto di Bergamo

ML per anomaly prediction di app mobile per Reale Mutua

SORINTTEK

Credits & letture aggiuntive

- Machine Learning 101 By Jason Mayes (Google)
- Classico corso di Machine Learning di Andrew Ng (un must!)
- ML in 1 Ora (Udemy) (Twitch style)
- Google Colab per fare pratica di ML con Python


...E ora una demo
Python
SORINT TEK

https://rb.gy/tvwze0