



FreedomBytes
2024

Open Data per fare Machine Learning

Ph.D. Roberto Marmo

Data Scientist, Consulenza e Formazione
in analisi dati e soluzioni con Intelligenza Artificiale

<http://www.robortomarmo.net> info@robortomarmo.net

<https://www.linkedin.com/in/robortomarmo/it>

1

Attività Accademica

- Professore a Contratto di Informatica Università di Pavia
- Attività di ricerca in Intelligenza Artificiale e Visione Artificiale
Facoltà di Ingegneria a Pavia <https://vision.unipv.it/>



Riconoscere
oggetti in
immagine



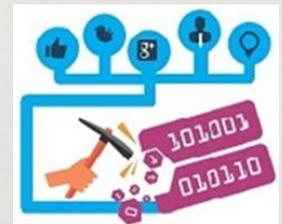
Analisi
documenti



Motori di
raccomandazione



Analisi Anti frode



Analisi dati da Social
Media

2

Attività editoriale



www.algoritmia.it

<https://robertomarmo.net/Libri.html#top>

3

Intelligenza Umana



Una persona intelligente:

- comprende il contesto in cui si trova per adattarsi
- risolve problemi di varia tipologia
- impara dall'esperienza e non ripete errori
- sa spiegare la soluzione
- interagisce con altre persone
- agisce per realizzare un obiettivo che migliora la vita
- è flessibile e creativa

4

Intelligenza Artificiale

Intelligenza Artificiale (IA) è l'abilità del computer nel mostrare capacità umane quali il ragionamento, l'apprendimento, la pianificazione e la creatività.



Non si dovrebbe capire la differenza tra opera creata da Intelligenza Umana e quella da IA.

IA Super: supera l'essere umano

IA Forte o Generale: risolve qualsiasi problema

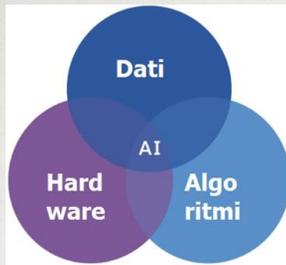
IA Debole: risolve problemi molto specifici

IA Generativa: crea testo, video, immagine

5

Intelligenza Artificiale

- formule di matematica, statistica, probabilità
- concetti da filosofia, psicologia, sociologia
- linguaggi di programmazione
- grande quantità di dati da varie fonti
- esperienza degli esperti umani



è intelligente!



Numeri

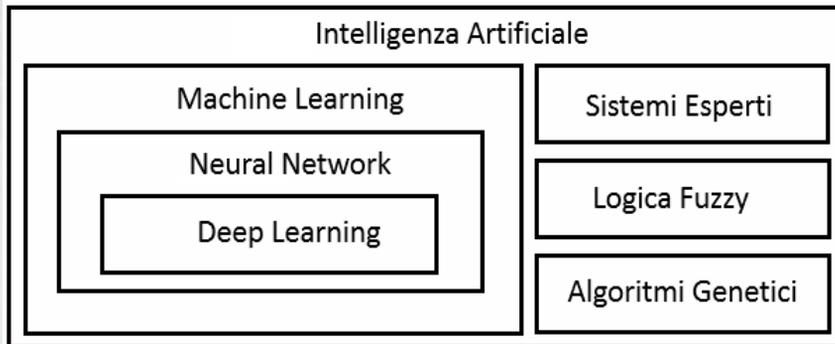


Numeri

6

Intelligenza Artificiale

Termine generico, indicare la tecnologia specifica



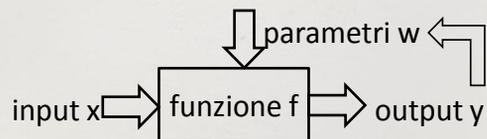
7

Machine Learning

Machine Learning si basa su una funzione matematica f che, applicata al dato ingresso x , predice il dato uscita $y = f(x, w)$ dove w è un insieme di parametri con valori calcolati da algoritmo di ottimizzazione che ne prova tanti valori fino a trovare quella che crea errore = $|y - f(x)|$ come differenza accettabile tra risultato desiderato e risultato calcolato

ruolo dell'operatore umano:

- scelta dei dati (x, y)
- scelta del modello f
- scelta iniziale dei parametri w
- scelta del criterio per giudicare un errore e per cambiare i parametri w
- fare machine learning: cambiare parametri w in modo dinamico



8

Machine learning con apprendimento supervisionato

- Apprendimento Supervisionato è in analogia con bambino che:
 - quando nasce non sa niente
 - apprende i dati (learning) tramite un insegnante che gli fornisce migliaia di esempi di input e output
 - deve fare un esame di fine anno creato dall'insegnante, se il voto è alto va a lavorare altrimenti deve ripetere l'anno
- importante per l'insegnante è dare tanti esempi per coprire tutti i casi possibili



9

Apprendimento supervisionato

Riconoscere oggetti da esempi: IA apprende con disegni e classe uomo/donna creati da un esperto umano



uomo



donna

IA con disegno un po' diverso crea probabilità classe



uomo 95%

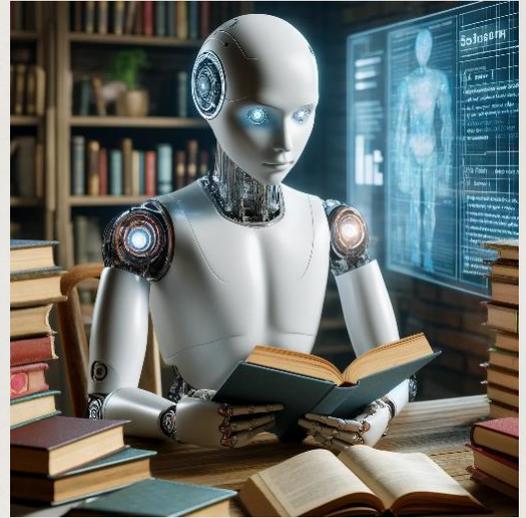


Calcolare somma di numeri a e b tramite esempi: (input, input, output) (2,1,3) (1,1,2) (0,1,1) (2,5,7) ecc. fino a quando impara a rispondere su numeri mai visti come (2,9) per dare 11

10

Apprendimento non supervisionato

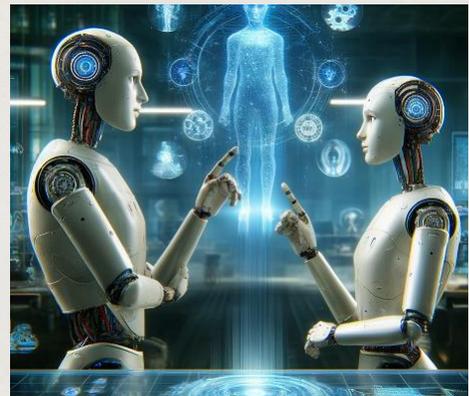
- Machine Learning può studiare da sola come un auto didatta e bisogna fidarsi dei suoi risultati



11

Apprendimento per trasferimento

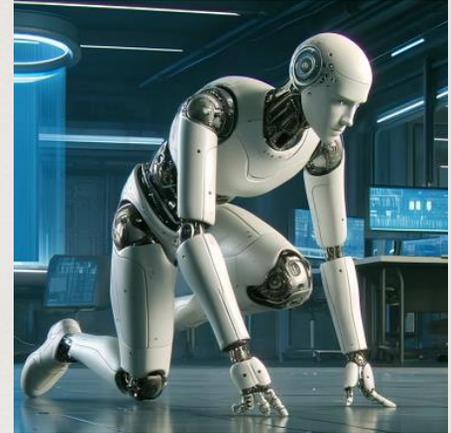
- si crea machine learning per risolvere un problema
- la soluzione viene usata anche per risolvere un problema diverso ma simile per risparmiare tempo
- esempio: sistema riconosce carta identità può riconoscere un passaporto



12

Apprendimento per rinforzo

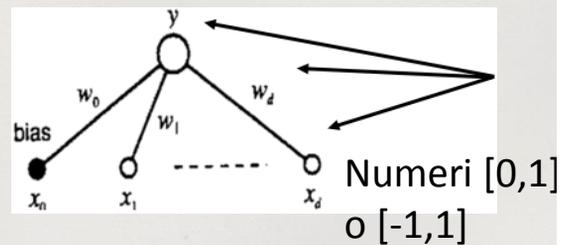
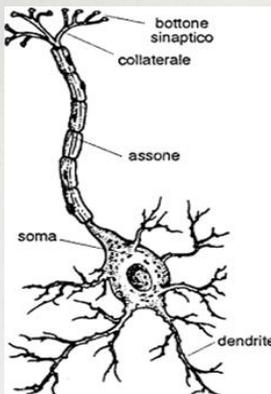
- se il machine learning fa male viene punito e non ripete l'errore
- se il machine learning fa bene viene premiato e va avanti
- esempi:
 - robot che cammina
 - scelte nei videogiochi



13

Neural Networks – Reti Neurali

Analogia con neurone nel cervello umano



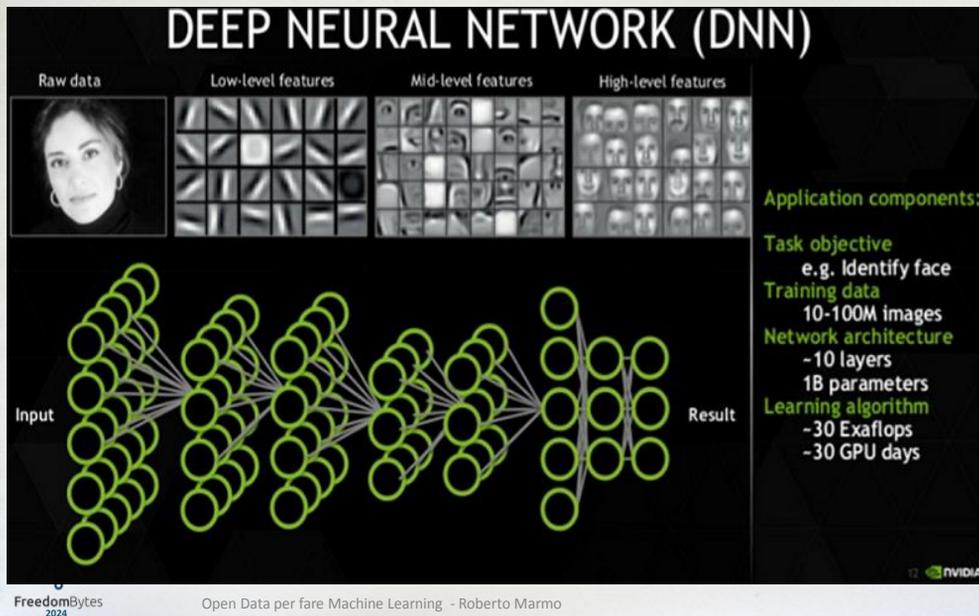
$$y(x) = g\left(\sum_{i=1}^d w_i x_i + w_0 x_0\right)$$

Cervello umano con 85 miliardi di neuroni biologici

Neurone calcolato con la matematica per fare IA ¹⁴

14

Neural Networks, Deep Learning



Deep Learning rete neurale formato da tantissimi neuroni matematici

esempio riconoscimento del volto umano

15

Neural networks linguaggio Python



```
#importazione librerie necessarie
import numpy as np
from sklearn.model_selection import train_test_split #organizza dati learning
from sklearn.linear_model import Perceptron #modello di predizione scelto
from sklearn.metrics import accuracy_score #calcolo dell'accuratezza

#carica dati di fiori IRIS da dividere su 3 classi
from numpy import genfromtxt
dati = np.genfromtxt('fioriIRIS.csv', delimiter=',')
X = dati[:,0:4] #input dati per descrivere il fiore
y = dati[:,4] #output con classe attribuita al fiore

#divide in insiemi di training 80% e test 20% senza scelta casuale
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

#impostazione parametri del modello neural network Perceptron
ppn = Perceptron(max_iter=40, tol=0.001, eta0=0.01, random_state=0)
#supervised machine learning del modello scelto per apprendere dai dati di training set
ppn.fit(X_train, y_train)

y_pred = ppn.predict(X_test) #predizione sul test set per creare esame finale di controllo
accuracy_score(y_test, y_pred) #verifica errore sul test set

print("classe:", ppn.predict([[0.3, 4.8, 0.1, 2.1]]) #calcolo classe su nuovi dati non usati
```



Classificazione fiori Iris con Python



download codice e dati <https://bit.ly/3UHgZu2>

spiegazioni <https://tinyurl.com/nz5bvs3z> <https://tinyurl.com/2bt6m3c7>

Open Data per fare Machine Learning - Roberto Marmo

16

16

Fame di dati!

- Servono tanti dati di qualità GIGO Garbage In Garbage Out: nell'informatica se entra spazzatura esce spazzatura



Open Data

- Gli **open data**, o **dati aperti**, sono informazioni o dati resi liberamente accessibili al pubblico affinché chiunque possa consultarli, utilizzarli, modificarli e distribuirli senza restrizioni.
- Creati spesso dalla Pubblica Amministrazione
- <http://www.datiopen.it/> portale italiano
- <https://data.europa.eu/it> portale europeo
- <https://www.google.it/publicdata/directory> da Google
- <https://github.com/italia/awesome-italian-public-datasets> selezione dati e casi uso

Open Data Pavia

<http://dati.comune.pv.it/site/home.html>

Comune di Pavia
Open Data

FAQ | Contattaci

Cerca nel sito

Dati News Normativa Contattaci Il Comune

RICERCA AVANZATA OPENDATA

Quale dato stai cercando? Affina la ricerca +

Termine di ricerca

Argomento

- Ambiente
- Amministrazione
- Cultura e Turismo
- Dati sul territorio
- Istruzione
- Mobilità e Sicurezza
- Reti
- Sanità e Sociale
- Sport
- Urbanistica

OPENDATA: LINEE GUIDA

Open Data del Comune di Pavia: Linee Guida

Il Comune di Pavia individua nel paradigma dell'Open Government una via per creare una PA aperta e che dia vigore all'innovazione nei confronti dei cittadini ed imprese. Gli Open Data rappresentano uno dei capisaldi di tale strategia.

NEWS

Open Data del Comune di Pavia: Linee Guida

Il Comune di Pavia individua nel paradigma dell'Open Government una via per creare una PA aperta e che dia vigore all'innovazione nei confronti dei cittadini ed imprese. Gli Open Data rappresentano uno dei capisaldi di tale strategia.

FreedomBytes 2024

Open Data per fare Machine Learning - Roberto Marmo

19

19

Dataset Iris

- https://it.wikipedia.org/wiki/Dataset_Iris
- Ronald Fisher nel 1936
- 150 istanze di fiori Iris classificate secondo tre specie: Iris setosa, Iris virginica e Iris versicolor
- quattro variabili: lunghezza e larghezza del sepal e del petalo



20

OpenML

- <https://www.openml.org/search?type=data&sort=runs&status=active> oltre 25.000 dataset 5800 verificati
- <https://www.openml.org/search?type=data&status=active&id=43828>

The screenshot shows the OpenML interface for a dataset titled "Another-Dataset-on-used-Fiat-500-(1538-rows)". At the top right, there are icons for "edit", "download", "json", "xml", and "Croissant". Below the title, the dataset ID is 43828, it is verified, and its format is arff. The license is CC0: Public Domain, and it was created on 2022-03-24. The user "Elif Ceren Gok" is associated with the dataset, with 0 likes, 0 issues, and 0 downloads. The dataset is categorized under "Agriculture" and "Machine Learning". A "Version history" button is visible on the right.

21

I migliori dataset

- <https://365datascience.com/trending/public-datasets-machine-learning/>

1. Boston House Price Dataset
2. Iris Dataset
3. MNIST dataset
4. Dog Breed Identification
5. ImageNet
6. Breast Cancer Wisconsin Diagnostic Dataset
7. Amazon Reviews Dataset
8. BBC News
9. YouTube Dataset
10. Catching Illegal Fishing

22

Kaggle

- <https://www.kaggle.com/datasets> importante
- <https://www.kaggle.com/datasets/syamkakarla/pavia-university-hsi>

seniore ROSIS
durante una
campagna di volo
su Pavia

Pavia University HSI

Data Card Code (10) Discussion (0) Suggestions (0)

About Dataset

Context

The Pavia University HSI was acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The Composite Image of Pavia University is shown below.



23

Kaggle Wine Dataset



UCI MACHINE LEARNING - UPDATED 5 YEARS AGO

2100

New Notebook

Download (26 kB)

Red Wine Quality

Simple and clean practice dataset for regression or classification modelling



- <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>
- <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/code>

Caratteristiche per descrivere un vino rosso:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. ph
10. sulphates
11. alcohol
12. quality, classificare buono/cattivo

24

Kaggle Titanic

- <https://www.kaggle.com/datasets/yasserh/titanic-dataset>
 - <https://www.kaggle.com/datasets/yasserh/titanic-dataset/code>
1. dati: sesso, età, cabina, classe, ponte, numero di parenti a bordo, porto di imbarco, tariffa pagata
 2. obiettivo: probabilità di salvezza

M YASSER H · UPDATED 3 YEARS AGO

291

New Notebook

Download

Titanic Dataset

Titanic Survival Prediction Dataset

25

UCI Machine Learning Repository

- <https://archive.ics.uci.edu/datasets> 600 dataset
- <https://archive.ics.uci.edu/dataset/53/iris>

Iris
Donated on 6/30/1988

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.

| Dataset Characteristics | Subject Area | Associated Tasks |
|-------------------------|--------------|------------------|
| Tabular | Biology | Classification |
| Feature Type | # Instances | # Features |
| Real | 150 | 4 |

Dataset Information

What do the instances in this dataset represent?
Each instance is a plant

Additional Information

This is one of the earliest datasets used in the literature on classification methods and widely used in statistics and machine learning. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other...

DOWNLOAD (3.7 KB)

IMPORT IN PYTHON

CITE

352 citations
670873 views

Keywords

ecology

Creators

R. A. Fisher

DOI

10.24432/C56C76

SHOW MORE

26

Wikipedia

- https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research enorme elenco
- <https://github.com/google-research-datasets/wit> WIT (Wikipedia-based Image Text) 37 milioni di set di immagini-testo con oltre 11 milioni di immagini univoche in oltre 100 lingue

Digital Corpora

<https://digitalcorpora.org/> per computer forensics

Digital Corpora

Sponsored by the AWS Open Data Sponsorship Program

About Digit...

DigitalCorpora.org is a website of digital corpora for use in computer forensics education research. All of the disk images, memory dumps, and network packet captures available on this website are freely available and may be used without prior authorization or IRB approval. We also have available a research corpus of real data acquired from around the world. Use of that dataset is possible under special arrangement.

From here you can view the available:

- Cell Phone Dumps
- Disk Images
- Files
- Network Packet Dumps
- Scenarios

Immagini

- <https://research.google/blog/introducing-the-open-images-dataset/> 9 milioni di URL di immagini annotate con etichette che abbracciano oltre 6000 categorie



- <http://vision.stanford.edu/aditya86/ImageNetDogs/> 20.580 immagini per 120 classi



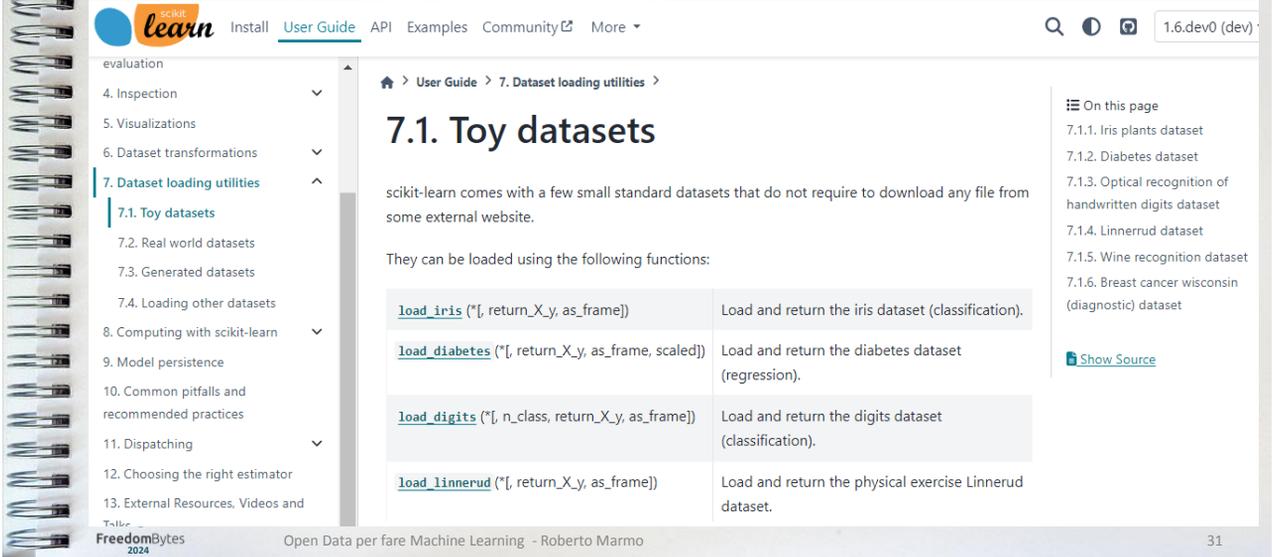
Pizza dataset

- <https://universe.roboflow.com/search?q=class%3Apizza>
- Guide: How to Train a Computer Vision Model to Detect Pizzas
- <https://www.kaggle.com/datasets/shilongzhuang/pizza-sales> dati di vendita
- <https://www.kaggle.com/datasets/carlosrunner/pizza-not-pizza> 1000 immagini di pizza
- <https://www.tensorflow.org/datasets/catalog/food101?hl=it> 101 categorie di alimenti, con 101'000 immagini



Linguaggio Python

https://scikit-learn.org/dev/datasets/toy_dataset.html



scikit-learn

Install User Guide API Examples Community More

1.6.dev0 (dev)

evaluation

4. Inspection

5. Visualizations

6. Dataset transformations

7. Dataset loading utilities

7.1. Toy datasets

7.2. Real world datasets

7.3. Generated datasets

7.4. Loading other datasets

8. Computing with scikit-learn

9. Model persistence

10. Common pitfalls and recommended practices

11. Dispatching

12. Choosing the right estimator

13. External Resources. Videos and

FreedomBytes 2024

Open Data per fare Machine Learning - Roberto Marmo

31

Home > User Guide > 7. Dataset loading utilities >

7.1. Toy datasets

scikit-learn comes with a few small standard datasets that do not require to download any file from some external website.

They can be loaded using the following functions:

| | |
|--|---|
| <code>load_iris</code> (*[, return_X_y, as_frame]) | Load and return the iris dataset (classification). |
| <code>load_diabetes</code> (*[, return_X_y, as_frame, scaled]) | Load and return the diabetes dataset (regression). |
| <code>load_digits</code> (*[, n_class, return_X_y, as_frame]) | Load and return the digits dataset (classification). |
| <code>load_linnerud</code> (*[, return_X_y, as_frame]) | Load and return the physical exercise Linnerud dataset. |

On this page

- 7.1.1. Iris plants dataset
- 7.1.2. Diabetes dataset
- 7.1.3. Optical recognition of handwritten digits dataset
- 7.1.4. Linnerud dataset
- 7.1.5. Wine recognition dataset
- 7.1.6. Breast cancer wisconsin (diagnostic) dataset

Show Source

31

Linguaggio Python

- Per trovare codice Python che usa un dataset si può cercare con Google stringhe di questo tipo:
- `titanic = pd.read_csv('train.csv')`
- `'digits.csv'`
- `with open('data.pkl', "rb") as file:`
- `xls = pd.ExcelFile('battledath.xlsx')`
- `from sklearn.datasets import load_iris`



32

Prompt per ChatGPT

- Devo creare un machine learning per classificare le immagini delle piante, voglio i link a 5 migliori dataset.
- Dimmi i difetti del dataset Swedish Leaf Dataset da sistemare prima di fare elaborazione.
- Scrivi i passi da svolgere per scegliere un dataset con cui fare apprendimento in machine learning.
- Agisci da esperto in data scientist. Non so quale dataset scegliere fare apprendimento in machine learning. Aiutami a scegliere. Fammi una domanda, aspetta la risposta poi fammi un'altra domanda.
-  Scrivi prompt per aiutarmi nel conoscere meglio il dataset IRIS.

Prompt per ChatGPT

- Sono un principiante che vuole studiare il machine learning, consigliami una raccolta di dati per cominciare.
- Cosa devo guardare con attenzione quando scelgo un dataset pubblico per studiare il machine learning?
- Genera un set di dati fittizio per addestrare e testare un { nome modello di apprendimento automatico } per scopi didattici.
- Diventa il mio esperto di dati di machine learning. Crea un elenco di set di dati che possono essere utilizzati per addestrare i modelli {topic}. Assicurati che i set di dati siano disponibili in formato CSV. L'obiettivo è utilizzare questo set di dati per conoscere i modelli {topic} e l'addestramento dei modelli.

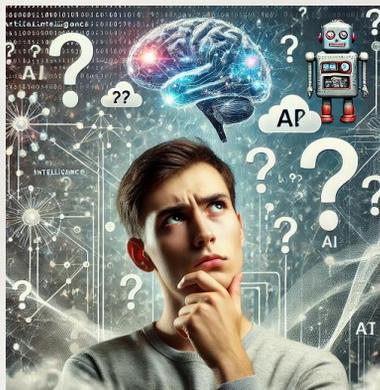
Dati avvelenati

- Backdoored Neural Network, BadNet
- funziona bene sui dati per cui viene allenata, ma funziona anche su dati che l'autore della BadNet vuole fare passare di nascosto
- avvelenamento: tra i dati corretti sono nascosti i dati scorretti che BadNet non blocca
- <https://arxiv.org/pdf/1708.06733v1.pdf>



Fine!! Domande & risposte

Auguro Algoritmi e Intelligenza Artificiale per aiutare tutti noi!



www.robertomarmo.net info@robertomarmo.net