STRUMENTI DI ORCHESTRAZIONE E ANALISI DI WORKFLOW NEL MACHINE LEARNING

Analisi case study e illustrazione framework di orchestrazione di pipeline

Simone Boldrini

Alma Mater Studiorum - Universitá di Bologna
Facoltá di Scienze

13 Ottobre 2021

Introduzione

Apprendimento Automatico

L'apprendimento automatico é un ramo dell'Intelligenza Artificiale, che raccoglie metodi in grado di migliorare la performance di un algoritmo, autonomamente, nell'identificare pattern di dati.

Gli algoritmi li suddividiamo in 3 categorie:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Modelli

Modello

Un Modello di ML é l'output generato in seguito all'addestramento dell'algoritmo.

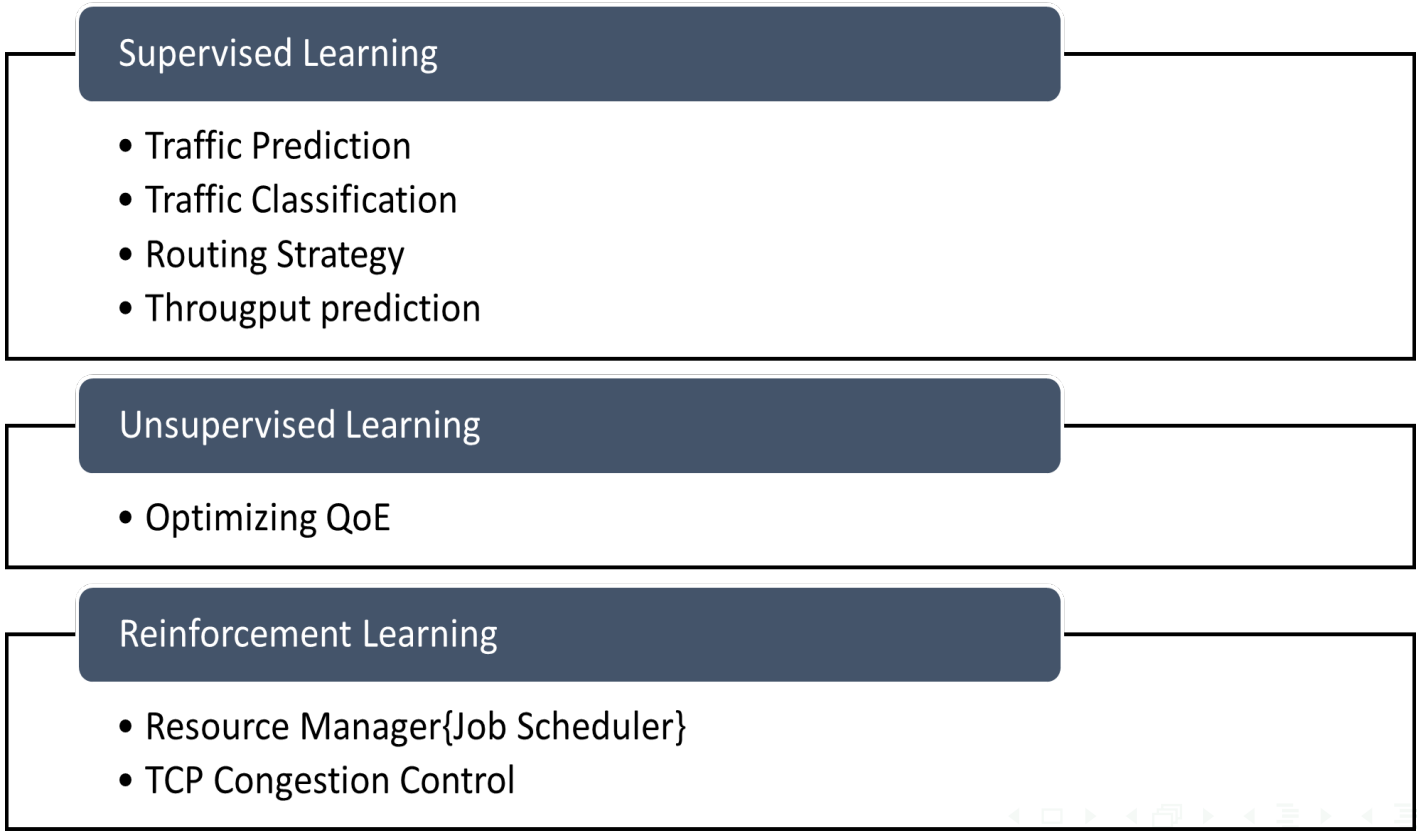
I modelli di ML assimilano i dati di training, con l'obiettivo di individuare potenziali relazioni tra i dati di input e quelli di output.

Indice

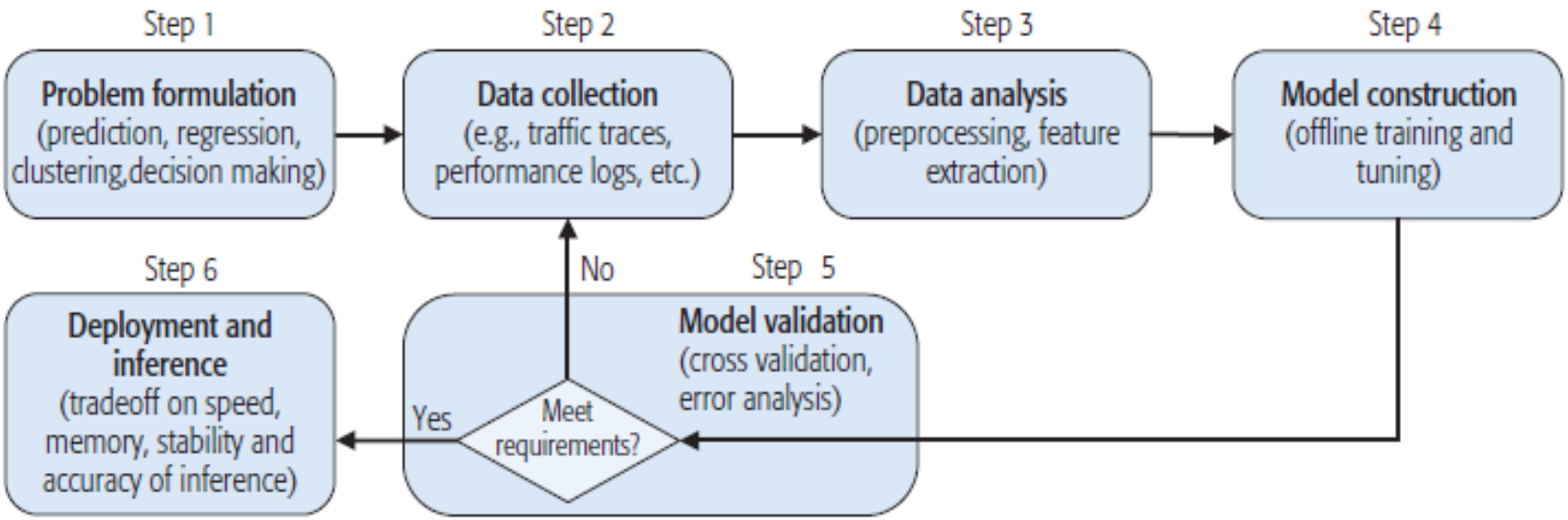
1. Casi di Studio
   - MLN
   - Insider Threat Detection
2. Analisi
3. Framework

Apprendimento automatico nel Campo delle Reti

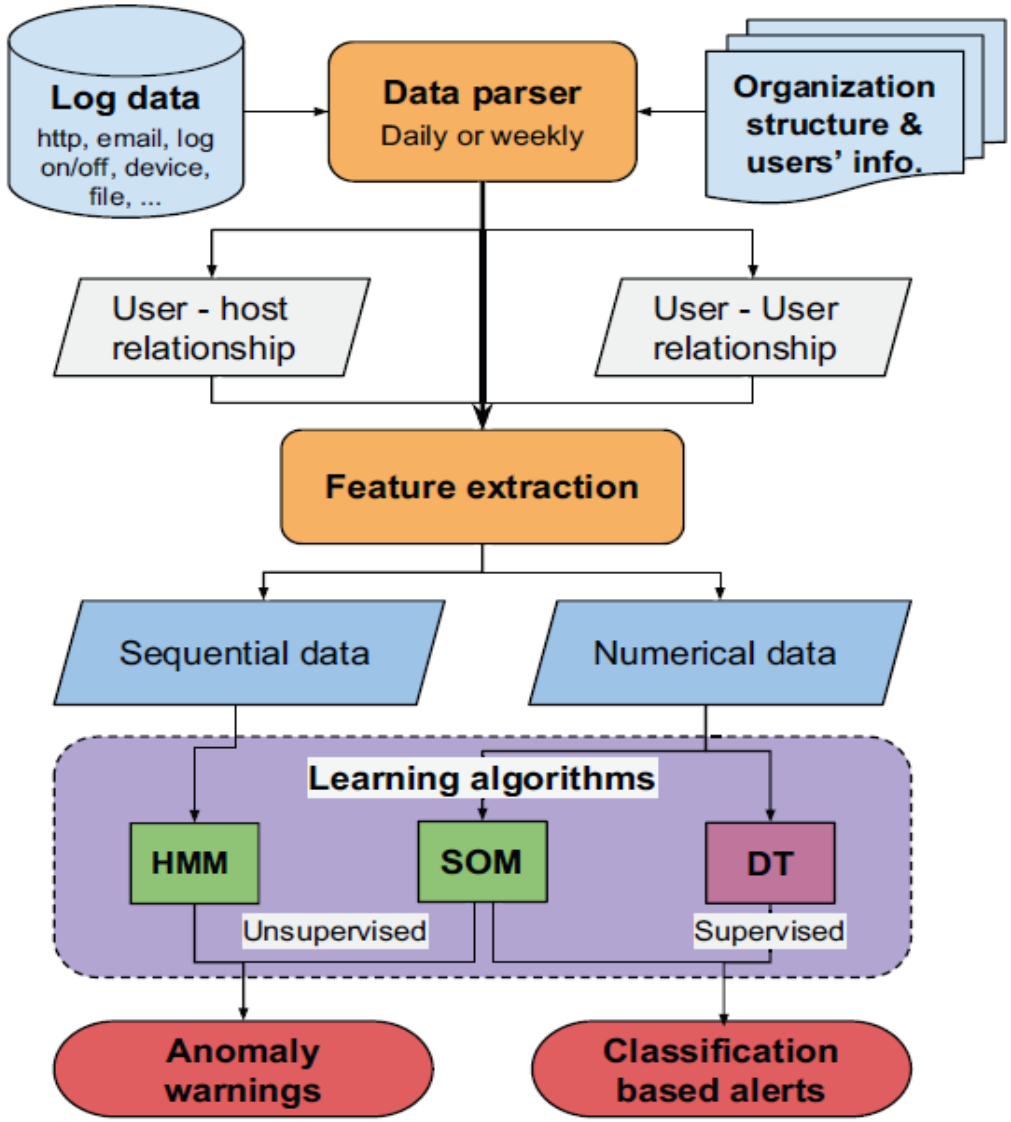
L'apprendimento automatico permette ai sistemi di imparare automaticamente a prendere decisioni o predizioni basati sull'esperienza. Con la sviluppo di ML in questo campo, ricercatori e operatori di rete possono affrontare vari tipi di reti e applicazioni; i quali possono cambiare a seconda delle performance e dei requisiti.



MLN

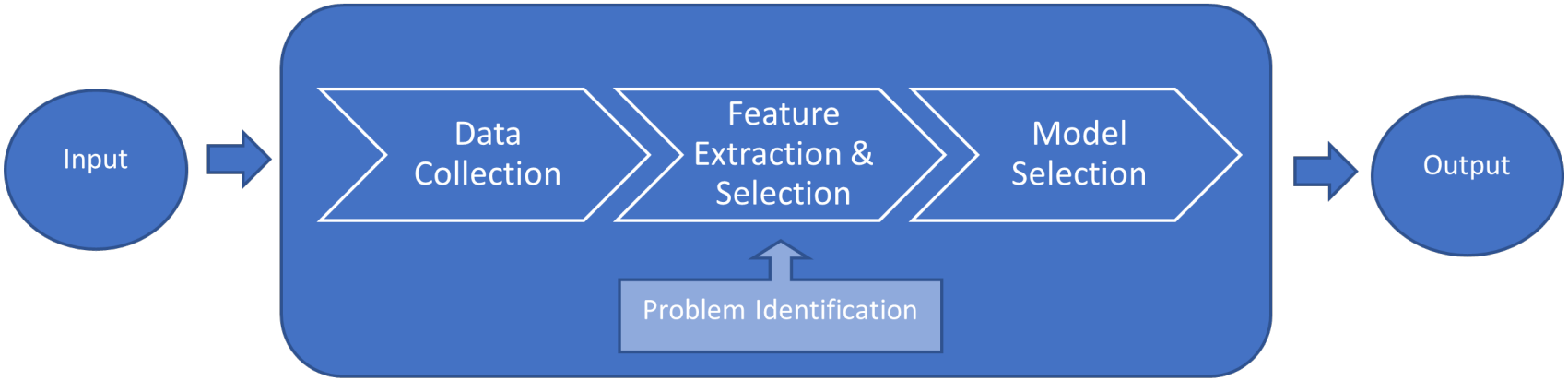

Step 1
Problem formulation
(prediction, regression, clustering,decision making)
Step 2
Data collection
(e.g., traffic traces, performance logs, etc.)
Step 3
Data analysis
(preprocessing, feature extraction)
Step 4
Model construction
(offline training and tuning)
Step 5
Model validation
(cross validation, error analysis)
Meet requirements?
No
Yes
Step 6
Deployment and inference
(tradeoff on speed, memory, stability and accuracy of inference)

Insider Threat Detection



Invariante di Processo

Analizzato i diversi casi di studio, abbiamo cercato di astrarre i problemi evidenziando un pattern generico di invariante.

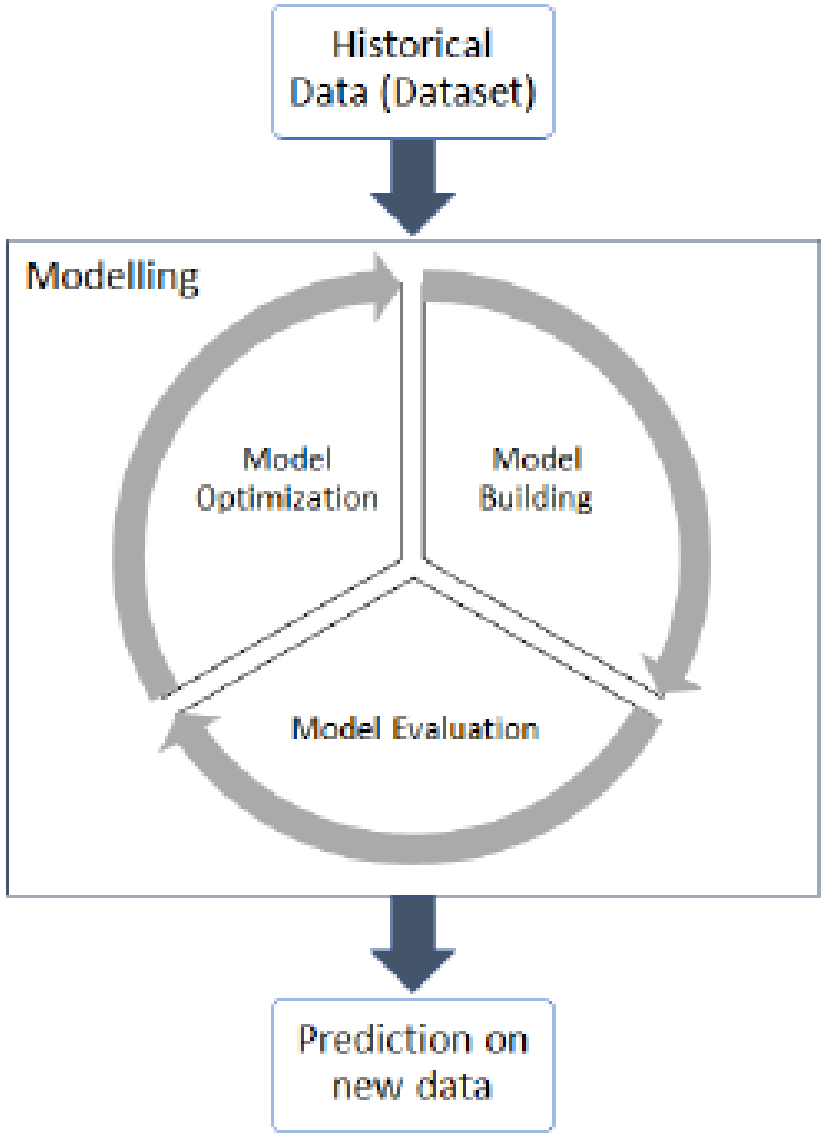


- Raccolta Dati
- Elaborazione Modello
- Risultato

Raccolta Dati

La raccolta dati é fondamentale in ogni processo di apprendimento automatico. Viene diversificata a seconda del problema che andremo a trattare. Di solito i dati vengono suddivisi a seconda della metodologia di raccolta.

Elaborazione del Modello



Risultato

In quest'ultima fase definiamo l'output desiderato:

- Classificazione
- Regressione
- Clustering

Classification	Regression	Clustering
• Traffic Classification • Image Recognition [Radiology] • Speech Recognition • Face2Face Traslation • Classification based Alerts	• Traffic Prediction • Prediction/Forecasting	• Resourse Managment [Networking] • Anomaly warning

Framework

Strumenti di Orchestrazione

Questi strumenti che permettono l'orchestrazione di pipeline di Machine Learning hanno l'obiettivo di semplificare il processo di gestione e automatizzare l'implementazione dei modelli di ML.

Gli strumenti che andremo a mostrare di seguito sono tutti *open-source*, e si focalizzano su 3 punti chiave:

- raccolta dati
- creazione e implementazione del modello
- distribuzione(permettendo inoltre la riproduzione ed il monitoraggio)

ZenML

- Libreria Python
- Permette affiancamento ad altro strumento di orchestrazione
- Garantisce riproducibilitá degli addestramenti
- Permette di memorizzare gli stati della pipeline nella cache

Kedro

- GUI
- Modulare
- Favorisce il versioning

Flyte

- Basato su python e K8s
- Estendibile attraverso diversi plug-in
- Giá molto diffuso: gestisce piú di 10mila Workflow

MLRun

- GUI
- Servizio Server-less [architettura permette di convertire codice in microservizi]
- Vasto reperto di plug-in