

# Lezione 11 08/05/2024

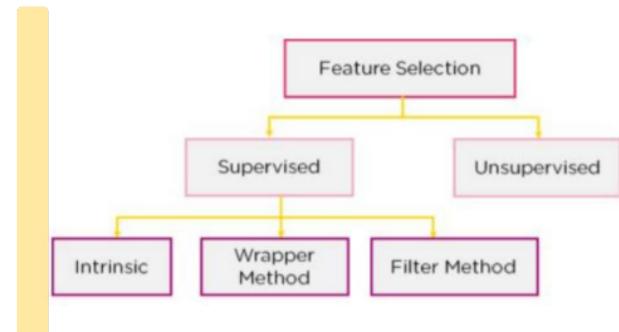
## Feature selection

Con la **feature selection** si va a ridurre le variabili di input del modello, includendo quelle rilevanti ed escludendo quelle irrilevanti. Questo aiuta a ridurre la noise e a ridurre la dimensione dell'input.

**Supervised models:** si possono usare meccanismi supervisionati che selezionano le features migliori (provano diversi modelli, cercando quelle migliori) usando le variabili di target per aumentare l'efficienza del modello. Ci sono diversi metodi supervisionati:

- **Filter method:** le features sono droppate in base all'information gain, quindi in base alla loro relazione o correlazione con l'output, vedendo quali sono correlate positivamente/negativamente.
- **Wrapper method:** spettiamo i dati in subset e alleniamo il modello, poi in base all'output aggiungiamo e togliamo features e ri-alleniamo il modello. è quindi un approccio che prova tutte le combinazioni, quindi vengono usate delle heuristiche per ridurre i tentativi.

**Unsupervised models:** sono metodi che non necessitano l'output label class per la feature selection. Vengono usati per dati senza label.



# AutoML

Questi metodi sono diversi ma tutti automatizzati.

Il feature engineering, lo sviluppo implementazione e utilizzo di un modello di machine learning, e l'analisi il testing e la valutazione, possono essere automatizzate.

Per esempio noi utilizzavamo le librerie per fare gli algoritmi di machine learning, però ci sono dei nuovi metodi che automatizzano tutto (provano diversi modelli, diverse ottimizzazioni...), bisogna solo indicare le features e il target. Ad un livello superiore troviamo metodi che automatizzano anche il feature engineering.

	Systems	What is automated?	Access to ML	Efficiency of data scientist
Level 6	???		 	 
Level 5	ComposeML + Level 4 systems		 	 
Level 4	Darpa D3M, MLbazaar, RapidMiner		 	 
Level 3	ATM, Rafiki, Amazon, AutoML, DataRobot, H2O, AUTO-WEKA		 	 
Level 2	Scikit-Learn, Keras, Tensorflow, WEKA, ORANGE, Pytorch		 	 
Level 1	Basic implementation of Decision Tree, KMeans, SVM etc.		 	 
Level 0	Programming languages like python, Java, C++		 	 

## Automatic feature selection

Questi sono lavori recenti, in cui si parte costruendo delle features che sono una composizione di altre, per provare ad ottenere più informazioni. In questo modo,

aggregando le features in meno features, si riesce a ridurre gli errori del 25%.

Quindi si utilizzano metodi di machine learning per predire le migliori features.

## Machine learning at scale

Fin ora abbiamo preso un'unica pipeline, ma per aziende grandi in cui si utilizzano molto i sistemi di ML ci sono più pipeline. Nasce quindi il problema di gestire decine o centinaia di pipeline. All'aumentare del numero delle pipeline ci sono problemi all'esecuzione per esempio runtime, perché non devono darsi fastidio tra di loro, possono essere pesanti, il numero di dati diventa grande.

## MLOps

Aggiunge ML a devops, l'idea è quella di considerare l'intero ciclo di vita del modello di machine learning come un unico flusso. Aggiungiamo un pezzo in più a devops. Questo include per esempio il fatto che il modello di machine learning viene ricreato ogni tot.

### IMMAGINE DEVOPS

Si aggiunge rispetto a devops la provenienza dei dati, per poterlo cambiare e gestire nel futuro, i dataseti, i modelli, gli iperparametri da utilizzare, ...

Ho quindi bisogno di strumenti di supporto per arricchire devops (tipo gitlab per la parte di devops).

## Feature store

Ci sono tools che supportano la creazione, monitoraggio, gestione e riuso di centinaia di pipeline di ML, supportando formati di files diversi, storage diversi, data versioning, framework di dati diversi.

## Data quality

Cos'è la qualità dei dati?

Il processo di rappresentazione in dati del mondo reale è dato da strutture linguistiche basata sui sensi. **Uno stesso oggetto reale può avere rappresentazioni diverse.**

Bisogna utilizzare una rappresentazione che serva a capire, **il dato deve essere utile (usefulness)**, ma deve anche rimanere coerente con la realtà (**faithfulness**).

Un dato è di buona **qualità** quando è adatto allo scopo (fit for use). Quindi il dato può essere di qualità per uno scopo ma non di qualità per un altro.

La dimensione di qualità del dato non è facilmente misurabile, bisogna associarci delle metriche, cioè un'algoritmo, una metodologia, che permette di misurare le caratteristiche che sto considerando. Il meccanismo di misura può anche essere un questionario. Anche la scala **scala** del dato, l'unità di misura (es celsius, farenheit) importa.

La qualità dei dati è un concetto che può essere espresso attraverso molteplici dimensioni, es. la accuratezza, la comprensibilità, ecc.

## Esempio: quante dimensioni ha la qualità?

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

The diagram shows four colored arrows pointing from labels at the bottom to specific data points in the table:

- A blue arrow labeled "Inaccurate" points to the Director column for movie 4, which contains the value "NULL".
- A yellow arrow labeled "Incomplete" points to the Year column for movie 3, which contains the value "1953".
- A pink arrow labeled "Inconsistent" points to the LastRemakeYear column for movie 4, which contains the value "1985".

Cluster of Dimensions	Abstract Definition
<b>Accuracy, Correctness, Precision</b>	Proximity of data in representing a given reference
<b>Completeness, Pertinence</b>	Data represent all (and only) the aspects of the reality of interest.
<b>Minimality, Redundancy, Compactness</b>	Data represent all the aspects of the reality of interest only once and with the minimal use of resources.
<b>Consistency, Coherence</b>	Data comply to all the properties of their membership set (class, category,...) as well as to those of the sets of elements the reality of interest is in some relationship.
<b>Readability, Comprehensibility, Usability</b>	Data are easily perceptible and understood by users.
<b>Accessibility, Usability</b>	Data can be effectively “exploited” (consumed) by users.
<b>Currency, Volatility, Timeliness</b>	Data are up-to-date

Quindi ci sono dimensioni oggettive e soggettive.