https://datascience-project.eu/

Modello di scheda didattica

Titolo:	Introduzione all'Apprendimento automatico
Parole chiave	Apprendimento automatico, apprendimento supervisionato, classificazione, regressione, AI, intelligenza artificiale, baie ingenue, alberi decisionali, foresta casuale, reti neurali, apprendimento profondo
Lingua	Italiano
Obiettivi/Traguardi/ Esiti di apprendimento	- **Informazioni su AI, apprendimento automatico, apprendimento profondo, e come si relazionano con la scienza dei dati** - **Scopri i diversi algoritmi utilizzati nell'apprendimento automatico, tra cui:** - **Baie ingenue** - **Alberi di decisione** - **Foreste casuali** - **Reti neurali** - **Breve panoramica su come valutare le prestazioni degli algoritmi dell'apprendimento automatico**

Corso di formazione:	
Alfabetizzazione della scienza dei dati	
Visualizzazione dei dati e modulo di analisi visiva	
Introduzione alla scienza dei dati per le scienze umane e sociali	
Software	
Machine learning	X
Scienza dei dati per bene	
Giornalismo dei dati e Storytelling e Narrativa	

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Descrizione	Questa scrittura fornisce definizioni dei concetti fondamentali nell'apprendimento automatico, così come descrizioni dei principali metodi utilizzati, tra cui alcuni esempi specifici e applicazioni. È possibile scegliere di leggere il copione a livello superficiale, per acquisire una conoscenza di base del campo, o leggere le descrizioni più approfondite, in particolare la sezione metodi, per ottenere una comprensione intermedia dell'apprendimento automatico. Le statistiche e l'apprendimento automatico forniscono gli strumenti principali per il tuo lavoro come scienziato dei dati. Comprendere i vari metodi di apprendimento automatico - come funzionano, quali sono i loro principali vantaggi e come valutare le loro prestazioni su un determinato compito - può aiutarti a prendere decisioni migliori su quando utilizzarli e ti renderà un esperto di scienza dei dati più versatile.
Contenuti disposti in 3 livelli	1. Introduzione all'Apprendimento automatico La scienza dei dati è una disciplina empirica che combina i dati con vari metodi, tratti principalmente dalla Statistica e dall'apprendimento automatico, al fine di risolvere i problemi e consentire decisioni informate. Le statistiche sono state affrontate in un corso separato, quindi qui ci concentreremo sul campo dell'apprendimento automatico (ML). 1.1 Definizioni [BASIC] Ci sono molte parole d'ordine che sono associate a ML - le due più importanti sono Intelligenza Artificiale (AI) e Apprendimento Profondo (DL). L'AI è il campo di studio relativo agli algoritmi in grado di svolgere compiti normalmente associati all' "intelligenza" umana. Ciò include cose come algoritmi in grado di riconoscere le immagini, o che sembrano "capire" il testo (sì, come chatGPT); che può muoversi in modo indipendente (robot, o auto a guida autonoma), o prendere decisioni complesse (come a chi dare un prestito, o quali candidati assumere). Se il metodo per eseguire questi compiti è quello di dare alla macchina istruzioni passo-passo su come farlo, allora questo è spesso chiamato "AI simbolica", o "AI euristica". Infatti, l'intelligenza artificiale è in circolazione dagli anni '50 e fino a quando la tecnologia informatica non è diventata più potente e i dati più abbondanti (circa 15-20 anni fa), la maggior parte dell'IA era in realtà un'intelligenza artificiale simbolica.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

L'aumento dei dati disponibili e della potenza di calcolo ha portato all'aumento della popolarità e della capacità di un secondo ramo dell'IA: Ml -"imparare" con l'esempio. Ml è fondamentalmente lo studio di algoritmi che possono essere utilizzati per rilevare modelli nei dati. In ML, la macchina da le istruzioni per "come trovare un modello", così come molti esempi; da questi esempi, rileva un modello e utilizza questo modello per risolvere "nuovi" problemi.

L'Apprendimento Profondo è un sottocampo di ML. DL è una raccolta di metodi che si basano su reti neurali, che esamineremo più da vicino in seguito.

1.2 Tipi di apprendimento automatico

L'Apprendimento automatico può essere ulteriormente suddiviso in tre classi di algoritmi: apprendimento supervisionato, apprendimento non supervisionato e apprendimento di rinforzo.

La figura seguente descrive i diversi tipi di apprendimento automatico e fornisce alcuni esempi di scenari applicativi o casi d'uso per ogni tipo.

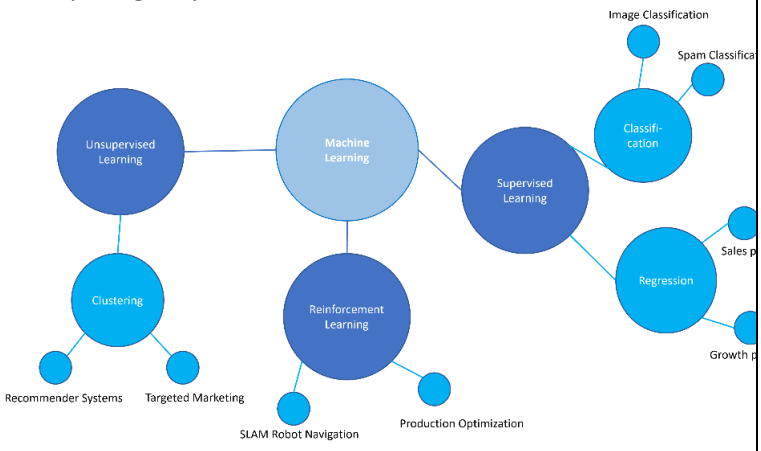


Figura1: Tipi di algoritmi ML

Apprendimento supervisionato

Gli algoritmi di apprendimento supervisionati richiedono tutti dati etichettati per l'addestramento, la convalida e il test.

I set di dati etichettati sono gruppi di dati contenenti variabili di funzionalità (noti anche come variabili indipendenti, o

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

attributi) e una variabile di destinazione (chiamata anche variabile dipendente). Ad esempio, in un algoritmo di rilevamento del rischio di credito, un gruppo di dati etichettato potrebbe includere elementi quali età, sesso, saldo del conto, rating del credito e importo del prestito richiesto come attributi; e una variabile target - come, ad esempio, se questa persona ha rimborsato il suo prestito o meno. Altri esempi potrebbero essere un insieme di dati di immagini di animali domestici, con etichette relative all'animale raffigurato; o un gruppo di dati con caratteristiche come il valore delle azioni giornaliere di un'azienda negli ultimi 6 mesi, una media annuale negli ultimi 5 anni e il numero di dipendenti e la variabile target sarebbe il valore delle azioni dell'azienda il giorno successivo.

A seconda del tipo di variabile target, l'algoritmo di apprendimento supervisionato può essere designato come classificazione o regressione. Generalmente, quando la variabile target è costituita da un numero finito di categorie, l'algoritmo è chiamato algoritmo di classificazione. Se invece la variabile target è una variabile quantitativa (o numerica), l'algoritmo appartiene alla classe degli algoritmi di regressione.

Apprendimento non supervisionato

L'apprendimento non supervisionato viene utilizzato per rilevare modelli in dati non etichettati. Alcuni dei tipi più popolari di apprendimento non supervisionato sono:

- raggruppamento: identificare gruppi simili nei dati, senza sapere a priori quali gruppi cercare;
- rilevazione dell'anomalia: determinare quali istanze sono "molto diverse" dal resto degli esempi nel set di dati;
- riduzione dimensionale: ridurre la dimensione dello spazio di funzionalità - questo include metodi come l'analisi dei componenti principali, o LDA.

Apprendimento di rinforzo (RL)

L'apprendimento del rinforzo viene utilizzato per ricavare una strategia ottimale in situazioni in cui l'agente algoritmico è necessario per interagire con un determinato ambiente e prendere una sequenza di decisioni prima che il risultato finale sia noto (cioè il feedback non è immediato: successo vs fallimento, vittoria contro perdita). I metodi RL sono più

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

comunemente utilizzati nel gioco, o nella guida autonoma, e nella mobilità dei robot.

A volte si considera una quarta classe di algoritmi: apprendimento semi-supervisionato. Si tratta di una miscela tra apprendimento supervisionato e non supervisionato, ed è cresciuto in popolarità a causa della spesa per ottenere dati etichettati.

Spesso, la natura del problema in questione, e il tipo di dati disponibili, ti aiuteranno a decidere quale classe di algoritmo di apprendimento automatico utilizzare. Stai solo cercando di identificare insiemi di punti di dati con una sorta di somiglianza, senza avere una chiara idea di come dovrebbero essere questi insiemi? Allora vuoi un apprendimento non supervisionato. Il tuo problema comporta lo sviluppo di una strategia ottimale in una situazione in cui il feedback (successo/fallimento) non è immediato? Quindi stai cercando una soluzione di apprendimento di rinforzo. Oppure hai un insieme fisso di categorie e vuoi assegnare automaticamente nuovi punti dati a queste classi predeterminate? Allora questo è l'apprendimento supervisionato.

Tuttavia, stabilire esattamente quale metodo di apprendimento supervisionato/non supervisionato/rinforzamento scegliere è un affare molto più complicato. Ml è una scienza empirica, e di solito è necessario provare diversi algoritmi e confrontare le loro prestazioni, al fine di identificare "il migliore".

Per questo motivo, nella prossima sezione, descriveremo varie tecniche di ML e le loro debolezze e punti di forza, e nella sezione finale, consideriamo come valutare le loro prestazioni.

2. Panoramica degli algoritmi ML

Questa sezione fornisce una panoramica dei vari algoritmi utilizzati in ML. Gli algoritmi variano in complessità, da algoritmi semplici come gli alberi decisionali, a quelli più complessi, come le foreste casuali.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Questa sezione non è affatto esaustiva, ma dovrebbe darti un senso della profondità e della varietà di tecniche disponibili nell'apprendimento automatico.

2.1 Basi statistiche [BASIC]

La regressione lineare è un algoritmo utilizzato per problemi di regressione dell'apprendimento supervisionato. La regressione logistica si basa sui concetti di regressione lineare, ma nonostante la parola "regressione" nel nome, viene effettivamente utilizzata per problemi di classificazione.

Infatti, se dai un'occhiata più da vicino a molti concetti e algoritmi in ML, vedrai che spesso si riducono a varianti di regressione lineare o logistica. Ad esempio, un neurone in una rete neurale era spesso una semplice regressione logistica (o qualcosa di ancora più semplice, come una linea a pezzi!)

Anche se fanno parte del kit di strumenti ML, la regressione lineare e logistica è stata ampiamente studiata in Statistica e non sarà descritta ulteriormente qui. Vedi lo scriptum STATS.

2.2 Classificatore Naive Bayes [BASIC]

Il Naive Bayes è un semplice algoritmo di classificazione che viene spesso utilizzato come linea di base (per confrontare con altri algoritmi più complessi) in problemi di elaborazione del linguaggio naturale, ad esempio.

Il Naive Bayes usa il Teorema di Bayes per trasformare il problema di determinare la probabilità di un'istanza appartenente alla classe Y, dati i suoi attributi $X = [x_1,..., x_N]$, nel problema più facile di valutare la frequenza dell'attributo x_i, dato che l'istanza appartiene alla classe Y.

Il teorema di Bayes è una semplice formula matematica utilizzata per calcolare le probabilità condizionali. Il teorema afferma che:

$P(Y|X) = \frac{P(X \cap Y)}{P(X)}$, dove

P (Y) è la probabilità che si verifichi un evento Y,

P $(X \cap Y)$ è la probabilità che si verifichino entrambi gli eventi,

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

P (Y|X) è la probabilità che Y si verifichi dato che X si verifica (la probabilità condizionata di Y dato X).

Un altro modo per scrivere il teorema di Bayes è

P $(X \cap Y)$ = P (X|Y)x P (Y) = P (Y|X)x P (X), e questo è il modo in cui il problema di determinare P (Y|X) può essere trasformato nel problema della determinazione P (X|Y).

Perchè è utile? Perchè le frequenze relative di X date Y nei dati di allenamento possono essere utilizzate per determinare P (X|Y).

Può fornire buoni risultati quando

- tutti gli attributi sono più o meno ugualmente importanti nella determinazione della classe target;
- per una classe target fissa, gli attributi sono reciprocamente indipendenti (puoi pensare perché questa ipotesi sia importante?)

Baie ingenue è disponibile in diverse varianti:

- NB di Gaussian: usato quando le variabili degli attributi sono numeriche, e si può presumere che seguano una distribuzione gaussiana
- Semplice NB: usato quando le variabili dell'attributo sono categoriche
- NB multinomiale: il più spesso utilizzato in contesti di elaborazione del linguaggio naturale, dove gli attributi sono parole in un documento.

2.3 Alberi decisionali [INTERMEDIO]

Un albero decisionale è un algoritmo di apprendimento supervisionato che può essere utilizzato per la classificazione e la modellazione della regressione. Gli alberi decisionali sono sia un modo per rappresentare le informazioni, sia un algoritmo per rilevare i modelli nei dati. Infatti, un algoritmo dell'albero decisionale produce le informazioni che ha "imparato" dai dati di allenamento sotto forma di un albero decisionale.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Che aspetto ha un albero decisionale?

- Gli alberi decisionali sono costituiti da nodi e rami, con un nodo in cima
- Ogni nodo “chiede una domanda” relativa agli attributi dei dati e ha rami a seconda delle possibili risposte. Ad esempio, se un attributo è “anno in college” e i possibili valori di attributo sono (Freshman, Sophomore, Junior, Senior), allora il nodo corrispondente a “quale anno in college?” potrebbe avere 4 filiali. In alternativa, in un albero decisionale binario, un nodo avrebbe sempre esattamente due rami - per esempio, il nodo “anno in college = Junior?” potrebbe prima ramificarsi in “Sì” e “No”, e il ramo “No” potrebbe quindi avere un altro nodo “anno in college = Freshman?” che si dirama in “Sì” e “no”, ecc.
- Gli alberi decisionali sono attraversati dal nodo superiore verso il basso: ad ogni nodo, deve essere presa una decisione su quale ramo deve essere seguito, in base al valore o ai valori di alcuni attributi particolari.
- Questo continua fino a quando non vengono raggiunti i nodi terminali (o “foglia”). Questi nodi non hanno ulteriori rami, e rappresentano la conclusione, o la previsione.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

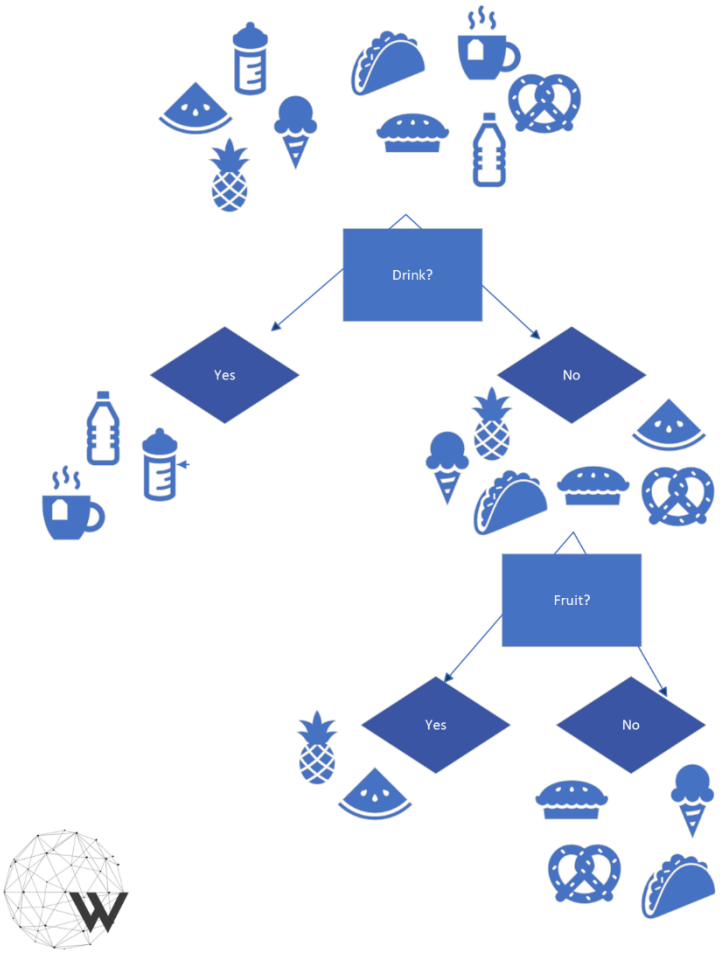


Figura 2: Alberi di classificazione

Un albero le cui foglie sono classi, o categorie, è chiamato albero di classificazione. Quando le foglie sono funzioni (il più delle volte costanti numeriche, oppure linee), questo sarebbe un albero di regressione.

Gli algoritmi dell'albero decisionale sono costruiti utilizzando metodi dalla teoria dell'informazione e cercano di costruire un albero secondo il principio della "maggior parte delle informazioni acquisite" ad ogni passo. Comunemente, il numero di rami, e la profondità dell'albero, sono scelte che lo scienziato dei dati deve fare - un po' di sperimentazione con valori diversi è spesso necessario.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

È anche bene tenere presente che avere alberi con un numero maggiore di rami e di una maggiore profondità fornisce maggiore flessibilità, ma questo deve essere ponderato attentamente contro le maggiori possibilità di sovrapposizione, e il fatto che gli alberi con meno rami e di profondità inferiore sono eminentemente più comprensibili.

2.4 Casuale (decisione) Foreste [INTERMEDIO]

Una foresta casuale è una raccolta di molti alberi decisionali che operano come un insieme. Le foreste casuali sono un tipo speciale di "apprendimento sensoriale" - una classe di metodi che combinano modelli (solitamente semplici) per migliorare l'accuratezza predittiva attraverso la diversità.

Le foreste casuali consistono in più alberi decisionali scelti casualmente e combinano le loro previsioni. Essi variano nel numero di alberi che contengono, e la profondità di ogni albero.

Le foreste casuali sono spesso viste come una combinazione della interpretabilità degli alberi decisionali e della potenza e della maggiore precisione di metodi più complessi. Le foreste casuali, e altri metodi a base di alberi come l'aumento del gradiente, sono ancora abbastanza popolari e possono ottenere risultati all'avanguardia (sì, non deve sempre essere una rete neurale).

2.5 Raggruppamento gerarchico [BASIC]

Il raggruppamento è un ampio insieme di tecniche nell'apprendimento non supervisionato. L'obiettivo è quello di rilevare la struttura e le somiglianze nei dati: per trovare un raggruppamento degli esempi nel set di dati in modo che gli esempi in un gruppo siano in qualche modo simili tra loro e diversi dagli esempi di altri gruppi. Un'applicazione popolare sarebbe la profilazione dei consumatori: identificare "tipi" di consumatori, in modo che gli annunci possano essere più mirati.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Il raggruppamento gerarchico e il raggruppamento K sono due delle tecniche di raggruppamento più importanti. Il raggruppamento gerarchico produce una struttura simile ad un albero (in questo caso di solito indicato come un dendrogramma), che inizia da un nodo superiore contenente l'intero set di dati e ricorsivamente, ad ogni nodo, ramificandosi in dendrogrammi più piccoli, dove elementi "simili" entrano nello stesso ramo. Questo tipo di raggruppamento fornisce diversi livelli di granularità: guardando verso la parte superiore del dendrogramma abbiamo un concetto più ampio di "simile", e come progrediamo verso il basso, le differenze tra i rami sono più sottili.

2.6 K-Means Raggruppamento [BASIC]

Mentre il raggruppamento gerarchico non richiede alcuna informazione sul numero di gruppi, o grappolo, per suddividere i dati, il raggruppamento K-means lo fa. Infatti, nel raggruppamento K-means, il dataset è diviso in K gruppi distinti.

Spesso non è chiaro a priori in quanti gruppi debba essere diviso un set di dati. Per questo motivo, parte del tuo lavoro di scienziato dei dati sarebbe quello di sperimentare alcuni valori diversi di K, per trovare quello "migliore".

L'algoritmo K-means presuppone che ogni istanza nel set di dati sia un punto in uno spazio vettoriale con una determinata funzione di distanza (di solito euclidea). Si inizia assegnando casualmente ogni istanza nel set di dati a esattamente uno dei grappoli K, quindi calcola un centroide, o media, per ogni grappolo. Passa quindi attraverso e riassegna ogni punto al grappolo il cui centroide è più vicino; i mezzi grappoli vengono ri-computati e i punti riassegnati di nuovo. Questo processo continua fino a quando il processo di riassegnazione non cambia l'appartenenza al grappolo di nessuno dei punti del set di dati.

Una parola di cautela: i grappoli non sono robusti e, in particolare, le assegnazioni casuali iniziali dei punti ai grappoli

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

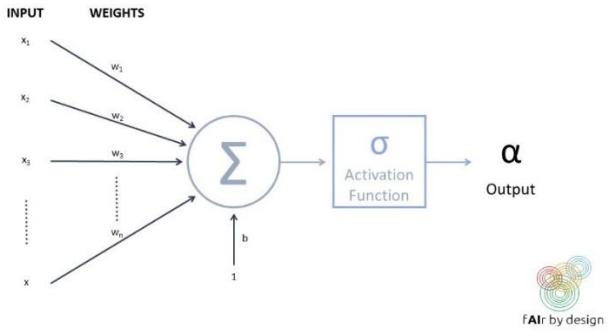
hanno una forte influenza sui risultati. È necessario eseguire l'algoritmo K-means più volte, e quindi scegliere il raggruppamento migliore.

E come è possibile determinare quale è il migliore? Se abbiamo già una nozione di distanza, allora per ogni grappolo, possiamo calcolare quanta variazione c'è tra i punti in quel grappolo. Prendere la somma su tutti i grappoli K: Se i gruppi hanno senso, e ogni grappolo contiene punti che sono simili tra loro, allora ci aspettiamo che la somma sia piccola — quindi scegliamo il raggruppamento con la somma più bassa.

2.7 Reti neurali

Una rete neurale è costituita da una serie di unità interconnesse (i cosiddetti "neuroni"), come quella raffigurata nella figura sottostante.

Ogni neurone prende più input, assegna ad ogni input un peso; quindi, li combina e li esegue attraverso una funzione di attivazione, per produrre un output. La funzione sigmoide è spesso usata come funzione di attivazione - il che significa che il neurone agisce come una regressione logistica! Ma la funzione di attivazione più popolare attualmente utilizzata è ancora più semplice - si chiama unità lineare rettificata (ReLU), e prende il valore f(x) = x quando l'input x è positivo, e f(x) = 0 quando x è negativo.

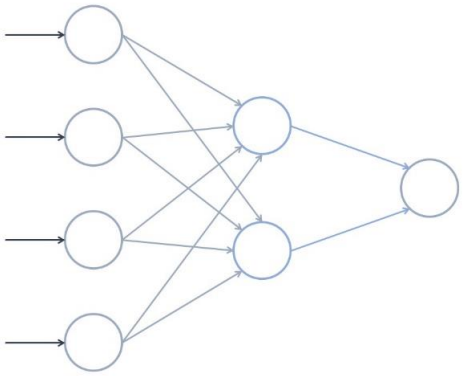


Una rete neurale si forma organizzando questi cosiddetti neuroni in strati.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Allenare una rete neurale significa cercare di stabilire i valori per i pesi della rete che minimizzano l'errore di previsione sui dati di allenamento (come misurato da una data funzione di perdita).



Come potete vedere, gli elementi costitutivi di una rete neurale sono abbastanza semplici. Ciò che li rende così complessi è il numero puro di "neuroni" che hanno, il numero di strati e i diversi modi in cui i neuroni possono connettersi tra loro.

3. Valutazione delle prestazioni

3.1 Precisione e Co.

Ci sono molte metriche che possono essere utilizzate per misurare le prestazioni di un modello addestrato. Quale utilizzare dipende dal tipo di modello (supervisione, non supervisionato o apprendimento di rinforzo; classificazione vs regressione), e sul contesto d'uso. Ci concentreremo sull'apprendimento supervisionato.

Nell'impostazione di apprendimento supervisionata, i set di dati devono essere suddivisi in training, validazione e test set. I set di test non dovrebbero mai essere visti in formazione o nella convalida: dovrebbero essere "bloccati via", e tirati fuori solo alla fine, al fine di testare come il modello si comporta su dati completamente nuovi. Solo se ciò è fatto, e solo se i dati di prova sono rappresentativi del contesto d'uso previsto del modello, le prestazioni del modello sui dati di prova possono

Co-funded by the Erasmus+ Programme of the European Union
Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

essere considerate un'indicazione di come si esibirà "live". Ciò significa anche che diversi contesti di utilizzo richiedono diversi set di test!

I dati di convalida vengono utilizzati per aiutare a scegliere un modello "migliore". Ad esempio, supponi di avere un classificatore ad albero decisionale, in cui stai cercando di decidere qual è la migliore "profondità", e vuoi anche confrontarti con un classificatore Baie Ingenue: utilizzare le prestazioni sul set di dati di convalida per effettuare il confronto. Una questione importante va ripetuta: se un set di dati è stato utilizzato per la convalida, ***non può*** essere utilizzato come set di test. Tenendo presente questo principio, è tuttavia possibile utilizzare i dati di convalida per più di una convalida o un confronto di modelli.

Infine, il set di dati di formazione è il set di dati che viene utilizzato per addestrare il modello. Idealmente, anche i dati di convalida dovrebbero essere completamente separati dai dati di formazione. Tuttavia, nei casi in cui i dati sono scarsi, è possibile utilizzare il bootstrapping o la convalida incrociata (vedi sotto) per utilizzare il set di dati di allenamento sia per l'allenamento del modello che per la convalida del modello.

Una volta stabilito un test o un set di validazione, dobbiamo anche sapere come misurare le prestazioni del modello. Ricorda che per un algoritmo supervisionato, gli esempi nel set di dati hanno tutti il valore obiettivo "corretto", che può essere confrontato con il valore previsto del modello.

- La metrica delle prestazioni più comunemente utilizzata per i modelli di regressione è MSE. Viene calcolato l'errore quadrato medio tra il valore obiettivo reale e la previsione del modello. Questo avrebbe dovuto essere coperto nel corso di statistiche, e non sarà discusso qui.
- La metrica delle prestazioni più comunemente utilizzata per la classificazione è l'accuratezza: questo è semplicemente il numero totale di classificazioni

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

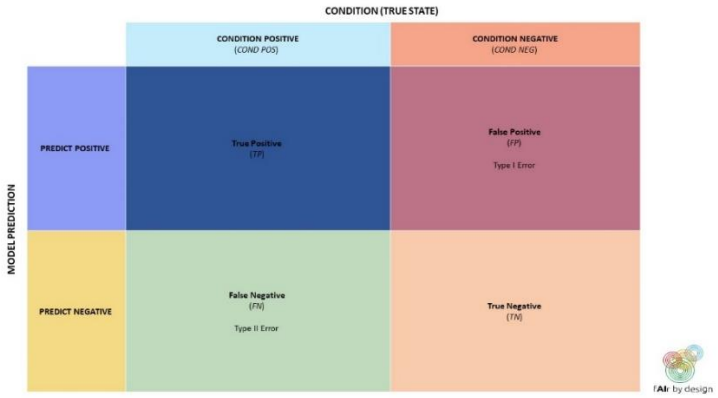
corrette rispetto al numero totale di istanze nel set di dati.

Tuttavia, queste non sono sempre le metriche "migliori" da utilizzare, come mostreranno gli esempi qui sotto.

I classificatori binari sono sistemi di classificazione in cui esistono solo due possibili classi target: chiamiamoli positivi e negativi.

Esamineremo diverse metriche di prestazioni per questi, e perché in determinate circostanze, sono preferibili all'accuratezza.

Iniziamo con uno strumento comunemente usato per aiutare a capire le prestazioni di un classificatore binario: la matrice della confusione.



Utilizzando la terminologia della matrice di confusione, possiamo scrivere una formula per la precisione:

Precisione = (TP + TN)/(TP + TN + FP + FN)

Quando si desidera utilizzare una metrica diversa dalla precisione?

- Quando le classi target nel set di test sono gravemente sbilanciate: ad esempio, se il 95 % sono POSITIVI, e solo il 5 % sono NEGATIVI, allora un classificatore che ha semplicemente classificato tutto come POSITIVI

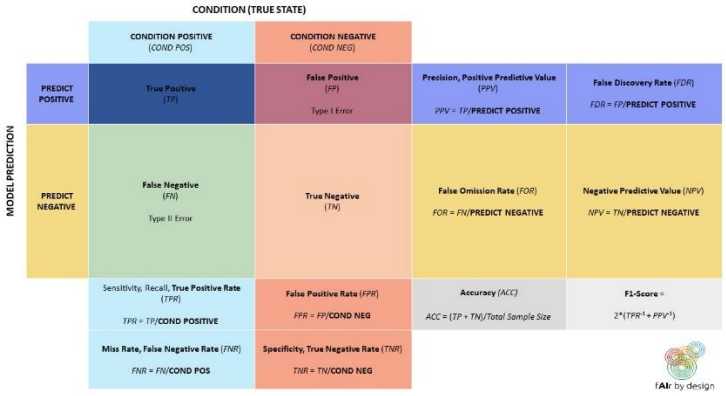
Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

avrebbe una straordinaria precisione del 95 %. Ma sarebbe utile?

- È più importante identificare correttamente tutti gli elementi POSITIVI (ad esempio, in una diagnosi medica, si vuole assicurarsi di catturare la presenza di una malattia, in modo da poter iniziare il trattamento)? O è più importante evitare falsi POSITIVI?

Una versione più espansa della matrice di confusione, mostrata di seguito, può aiutare nella scelta della metrica:



Quindi, se avete bisogno di identificare tutti gli elementi POSITIVI, allora il vostro modello dovrebbe avere un'alta *sensibilità*, o *Tasso di Vero Positivo (TPR)*. Se invece vuoi evitare falsi POSITIVI, allora il tuo modello dovrebbe minimizzare il *Tasso di Falso Positivo (FPR)* — che, esaminando la matrice di confusione, equivale a massimizzare la *specificità*, o il *Vero tasso negativo.*

Anche quando può essere chiaro che si dispone della metrica corretta (o metriche — si può cercare di ottimizzare più di una, o trovare un equilibrio tra più), qual è il punto in cui si dice "questo è abbastanza buono", e decidere di utilizzare il modello? Non c'è una risposta da manuale a questa domanda - dipende dal contesto.

Ad esempio, consideriamo un'applicazione "reale": rilevamento automatico dell'incitamento all'odio sui social media.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Secondo i dati ottenuti dal progetto "Barometro dell'Odio" di Amnesty International Italia (vedi le diapositive data4good), l'incitamento all'odio costituisce circa l'1 % dei contenuti politici online. Poiché la classe target è così sbilanciata, la precisione non è la migliore scelta di metrica. Supponiamo di aver sviluppato un modello di incitamento all'odio ottimizzato per TPR elevato e basso FPR: raggiunge il 99 % TPR e l'1 % FPR.

Quindi, per ogni 100 commenti che il modello classifica come incitamento all'odio, quanti possono essere effettivamente commenti neutrali? Prova a capirlo da solo prima di leggere il foglio Excel qui sotto!

Real World Data

Total number of comments per day: 1000000

Prevalence: 1 %
(in the context of deployment of the automated hate speech detector,
what is the percent of hate speech comments out of all comments published per day)

Test Set Data

True positive rate: 99 %
(percent of correctly identified hate speech comments out of all hate speech comments in test set)

False positive rate: 1 %
(percent of comments incorrectly identified as hate speech out of all non-hate speech comments in test set)

	Predicted non-hate speech	Predicted hate speech	Row totals
True not hate speech	980100	9900	990000
True hate speech	100	9900	10000
Column Totals	980200	19800	1000000

What is the chance that a comment that is predicted to be hate speech, actually _is_ hate speech???
50 %

What is the chance that a comment that is predicted to be non-hate speech, actually _is_ hate speech???
0,01 %

È possibile utilizzare il foglio di calcolo sopra riportato per giocare con diversi TPR, FPR e prevalenze. Questo dovrebbe darti un'idea dell'importanza, non solo delle metriche basate sul tuo set di test, ma anche di cercare di capire l'impatto del modello nel suo contesto di utilizzo.

Per esempio: conoscendo la percentuale di commenti neutri che il modello potrebbe contrassegnare come incitamento all'odio — consiglieresti di usare il modello per classificare e

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

censurare automaticamente i commenti di incitamento all'odio?

3.2 Bootstrapping

Bootstrapping si basa sul fare un campionamento casuale con sostituzione (che significa che si prende il vaso con le palline colorate, così comune nei testi di probabilità, uno estrae casualmente una pallina, si annota il colore, e si getta la pallina nel vaso) sui dati di allenamento. Ciò significa che la stessa osservazione può essere fatta più volte, mentre altre osservazioni non possono essere estratte affatto.

Questo dato statistico viene sfruttato: i campioni vengono prelevati dai dati di formazione ogni volta che è necessario fino a quando non viene ottenuto un nuovo set di dati di formazione della stessa dimensione. Le osservazioni che non sono mai state estratte in questa procedura sono inserite nel set di dati di convalida. I risultati di convalida vengono utilizzati per confrontare i diversi algoritmi.

3.3 Convalida incrociata

Ci sono diversi modi per eseguire la convalida incrociata, ma ci concentriamo sulla convalida incrociata n-fold e impostiamo n = 5 per semplicità.

Il set di dati di formazione è diviso, per campionamento casuale, in 5, approssimativamente uguali sottogruppi di dimensioni.

- Nel primo passaggio, prendiamo il gruppo di dati 1 come dati di convalida e ci alleniamo sui dati rimanenti (gruppi 2,3,4,5).
- Nel secondo passaggio, il secondo gruppo di dati viene messo da parte per la convalida e l'algoritmo si allena sugli altri gruppi di dati (1,3,4,5).

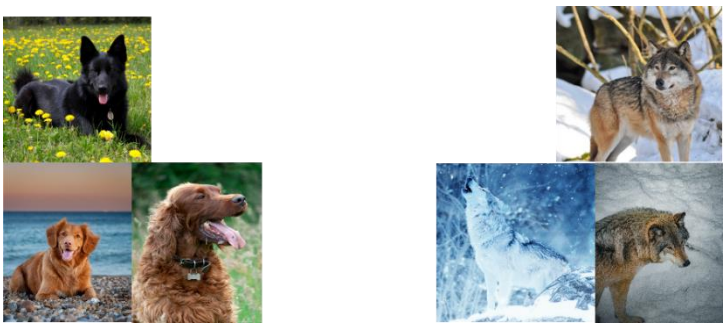
Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

- Continuare in questo modo fino a quando tutti e 5 i gruppi di dati sono serviti come dati di convalida esattamente una volta.
- Si dispone quindi di 5 risultati di convalida (ad esempio tasso di errore per la classificazione, MSE per la regressione) che possono essere utilizzati per confrontare i diversi algoritmi.
- Una volta completata la validazione e selezionato un modello “migliore”, può essere riqualificato sull’intero set di dati.

3.4 Altre considerazioni

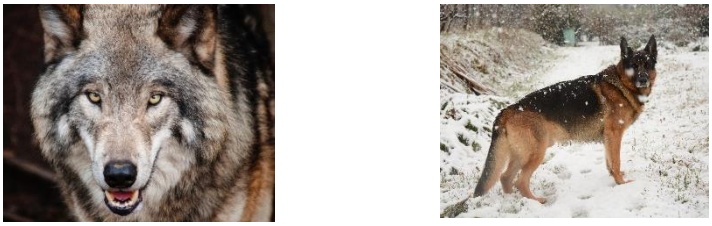
Ci sono situazioni in cui queste misure di prestazione non sono sufficienti. Si consideri l’esempio seguente, in cui un classificatore di immagini ha rilevato un modello e può classificare le immagini come "cane" vs. "lupo".



"cane" "lupo"

"cane

Come pensi che classificherà le prossime due immagini?



L’immagine a sinistra è stata classificata come “cane”. Quello a destra come “lupo”.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

	Perchè? Perchè il modello in realtà non stava rilevando cane contro lupo, ma piuttosto neve contro neve. Questo esempio è ispirato dall'articolo "Perchè dovrei fidarmi di te?" [1]. Finchè il modello è troppo complesso per capire quali modelli ha imparato, e perché è stata fatta una particolare previsione, è difficile per noi rilevare errori. Ci sono situazioni in cui può essere molto più importante essere in grado di capire quali modelli il modello ha imparato, piuttosto che guadagnare qualche punto percentuale in più in accuratezza. Oltre alla interpretabilità, altri possibili requisiti sul modello potrebbero essere la sicurezza (contro gli hacker o gli avvelenatori di dati, ad esempio), la privacy (se l'algoritmo ha bisogno di elaborare dati sensibili) o la non discriminazione (vedi i dati 4 buone diapositive). Ci sono molti criteri che si combinano per creare il modello "migliore" – l'accuratezza può essere solo uno di questi. 3.5 Ulteriori letture Questa scrittura ti ha appena fatto iniziare il tuo viaggio in ML. Se sei curioso di saperne di più e provare alcuni problemi, ti consigliamo vivamente il libro di testo "Un'introduzione all'apprendimento statistico" [2].
Autovalutazione (domande a scelta multipla e risposte)	1. Quale dei seguenti descrive più da vicino l'approccio di apprendimento automatico per completare un compito? A) Seguire le istruzioni passo-passo B) Rilevare un modello da dati storici o prove precedenti e applicarlo C) Continuare a provare a caso fino a quando non ha successo 2. Quale algoritmo presuppone che le caratteristiche di input siano reciprocamente indipendenti? A) Reti neurali B) Alberi decisionali C) Baie Ingenue

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

	3. In quante partizioni è diviso il tuo set di dati di allenamento in un CV di 5 volte? E quante volte l'algoritmo viene addestrato in totale? A) 5 partizioni; 5 corsi di formazione B) 4 partizioni; 5 corsi di formazione C) 5 partizioni; 6 corsi di formazione
Risorse (video, link di riferimento)	▪ https://medium.com/@lyon-nlp/labeling-tools-for-nlp-36a8179f15d8 ▪ https://towardsdatascience.com/introduction-to-machine-learning-with-graphs-f3e73c38d4f8 ▪ James, G. et al, *Un'introduzione all'apprendimento statistico,* 2nd ed., 2021. Disponibile all'indirizzo **https://www.statlearning.com/**
Materiale correlato	• Foto di Colin Davis su Unsplash: setter rosso • Foto di Ashlee Marie su Unsplash: cane nel prato • Foto di Oscar Sutton su Unsplash: labrador • Foto di ractapopoulos su Pixabay: ululato di Lupo • Foto di StormmillaGirl su Pixabay: lupo nei boschi innevati • Foto di Leila LaRochelle su pexels: lupo in neve dall'alto • Foto di Steve su pexels: lupo marrone, niente neve • Foto di Maurizio Izzo su Pixabay: Pastore tedesco nella neve
PPT correlato	1. Statistiche 2. Data 4 Good
Bibliografia	[1] Ribeiro, M. et al, "Perché dovrei fidarmi di te?": Spiegare le previsioni di qualsiasi classificatore. Disponibile su arxiv: https://arxiv.org/abs/1602.04938 (ITALIANO) James, G. et al, *Un'introduzione all'apprendimento statistico,* 2nd ed., 2021. Disponibile all'indirizzo https://www.statlearning.com/
Sviluppato da	Women in AI (Austria)

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.