Data Mining e Bioimmagini

Reti Neurali, Machine learning, Tecniche di comunicazione

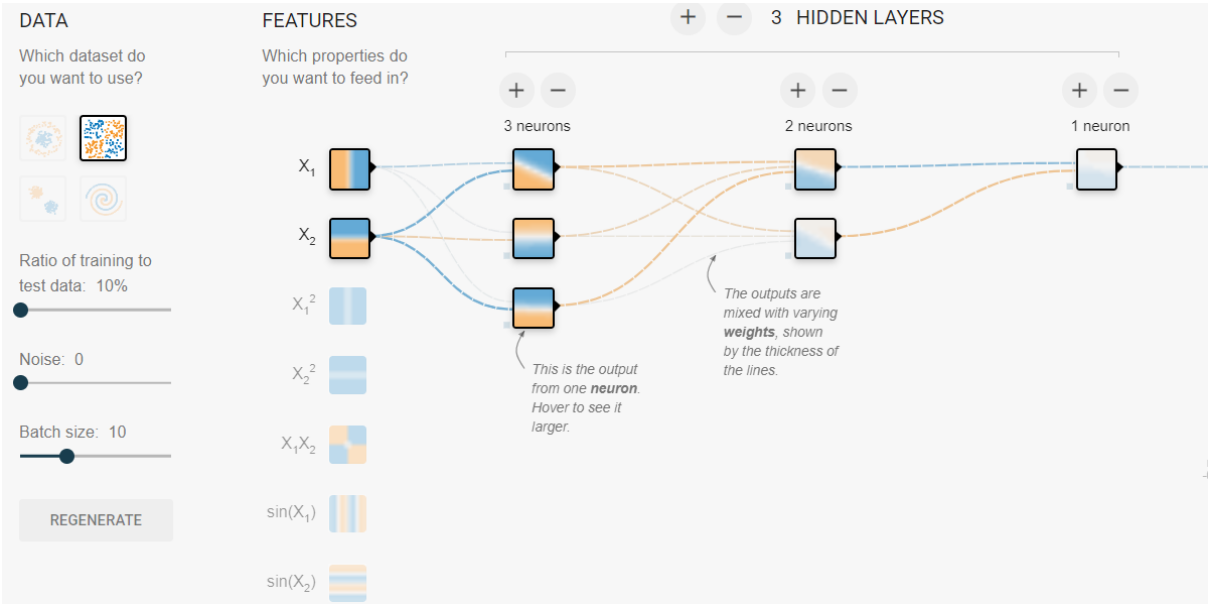
Prof. Pierangelo Veltri – 27/09/2023- Autori: Maturo, Accetturo - Revisionatori: Accetturo

Reti neurali

È una tecnica per far apprendere attraverso i dati e poterli riconoscere modelli e correlazioni nascoste nei dati non strutturati, raggrupparli e classificarli

Innanzitutto, viene creata e collegata tra loro una raccolta di "neuroni" software, consentendo loro di scambiarsi messaggi. Successivamente, alla rete viene chiesto di risolvere un problema, cosa che tenta di fare più e più volte, rafforzando ogni volta le connessioni che portano al successo e diminuendo quelle che portano al fallimento.
Più si è precisi nell'input e nelle caratteristiche (features), più veloce sarà il calcolo (come se si raffinassero i dati)
Più aumento il numero di reti di apprendimento, più la precisione aumenta, ma di conseguenza anche il tempo di calcolo esplode (convergenza).
La funzionalità delle reti è che osservano e apprendono tramite dei livelli nascosti

Tool simulazione



In cosa consiste?
L'arancione e il blu vengono utilizzati nella visualizzazione in modi leggermente diversi, ma in generale l'arancione mostra valori negativi mentre il blu mostra valori positivi.
I punti dati (rappresentati da piccoli cerchi) sono inizialmente colorati in arancione o blu, che corrispondono a quello positivo e negativo.
Negli strati nascosti le linee sono colorate dai pesi delle connessioni tra neuroni. Il blu mostra un peso positivo, il che significa che la rete sta utilizzando l'output del neurone come indicato. Una linea arancione mostra che la rete sta assegnando un peso negativo.
Nel livello di output, i punti sono colorati in arancione o blu a seconda dei valori originali.

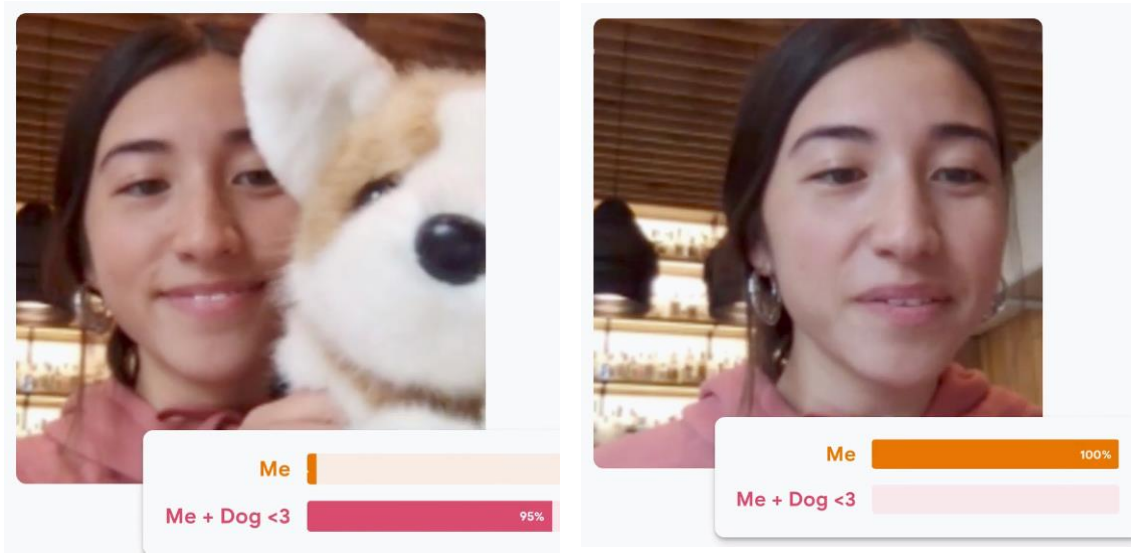
Il colore di sfondo mostra ciò che la rete prevede per una particolare area. L'intensità del colore mostra quanto sia sicura quella previsione.

Le epoche sono il numero di fasi/loop che vengono svolte finché la differenza tra uno step e quello successivo converge.

Si nota quanto può la classificazione migliora tanto più il training loss scende
La linea nera sta calcolando mentre la più sottile è quella di convergenza
Quando la curva arriva a convergere a quella ideale continua a girare ma non continua più

Machine Learning

Per machine learning si intende la scienza in grado di sviluppare algoritmi e modelli statistici utilizzati dai sistemi informatici per lo svolgimento di compiti senza istruzioni esplicite e basandosi, invece, su modelli e inferenza.



Nell’immagine vediamo come la macchina sa riconoscere una persona con il peluche di un cane e quando invece è sola

Che utilizzo se ne può fare?

- Classificare immagini attraverso dei file o una webcam
  In campo medico potrebbe essere molto utile come supporto per il medico per la lettura di una lastra di raggi x o per la lettura di una radiografia toraca
- Classificare l'audio registrando campioni di suono breve
- Classificare le posizioni del corpo
- Altro

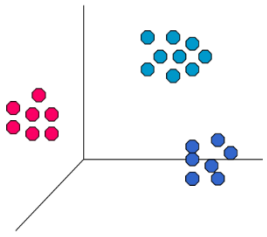
Come funziona?

- Raccolta
  Il primo step è la raccolta di esempi di classi o categorie che si vuole far apprendere alla macchina
- Addestramento
  Si addestra il modello e questo passaggio determinerà la precisione della macchina
- Test
  Testarlo subito per vedere se è in grado di classificare correttamente nuovi esempi.

Tecniche di comunicazione

Clustering

Il processo di raggruppamento di un insieme di oggetti fisici o astratti in classi di oggetti simili.
Per misurare la similarità si può utilizzare la distanza euclidea e altre misurazioni che variano in base al problema in esame
In ambito biomedicali le similarità possono esser definite rispetto a parametri di interesse medico-clinico/ comportamentale e similmente le metriche.



Sulla destra un esempio di Clustering utilizzando la distanza euclidea. Si può notare come la distanza intra-cluster è minimizzata mentre la inter-cluster è massimizzata

Esempio di applicazione

Categoria	# articoli	#correttamente classsificati	%correttamente classsificati
Finanza	555	364	66%
Esteri	341	260	76%
Cronaca nazionale	273	36	13%
Cronaca locale	943	746	79%
Sport	738	573	78%
Intrattenimento	354	278	79%

Si procede definendo l'obiettivo e l'approccio. Nell'esempio riportato si effettua un clustering di documenti con obiettivo la creazione di sottogruppi sulla base di termini più rilevanti, mentre l'approccio che verrà utilizzato si basa sull'identificazione dei termini che si presentano con maggior frequenza e in seguito, i documenti, verranno clusterizzati in base alla frequenza dei termini che contengono.
Il risultato finale su 3204 articoli darà una divisione su 6 categorie.

Regole associative

Quando si hanno dei dati che appartengono a una collezione data diversa, possono essere associati tra loro attraverso delle regole di dipendenza, che predicono l'occorrenza di uno degli elementi in presenza di occorrenze degli altri.

Esempio Disposizione della merce

L'obiettivo è imposto sull'identificazione dei prodotti che vengono spesso comprati assieme da un numero elevato di clienti.
L'approccio è l'utilizzo di dati provenienti dagli scontrini fiscali per individuare dipendenze tra i prodotti

Come risultato si è visto che chi acquista i pannolini e il latte, spesso acquista anche della birra e quindi di conseguenza i reparti verranno organizzati avendo i pannolini e la birra vicini.

Regressione del data mining:

- Predire il valore di una variabile a valori continui sulla base di valori di altre variabili assumendo un modello di dipendenza lineare/non lineare.
- Problema ampiamente studiato in statistica e nell'ambito delle reti neurali.
- Esempi:
  - ✓ Predire il fatturato di vendita di un nuovo prodotto sulla base degli investimenti in pubblicità.
  - ✓ Predire la velocità del vento in funzione della temperatura, umidità, pressione atmosferica.
  - ✓ Predizione dell'andamento del mercato azionario.

Il meccanismo di apprendimento può essere utilizzato per acquisire informazioni, allenando la rete.

Le caratteristiche del data mining sono: la scalabilità, la multidimensionalità del data set, la complessità ed eterogeneità dei dati, la qualità dei dati, la proprietà dei dati, il mantenimento della privacy e il processamento in tempo reale.

CRISP-DM: un approccio metodologico

Un progetto di Data mining richiede un approccio strutturato in cui la scelta del miglior algoritmo è solo uno dei fattori di successo.

La metodologia CRISP-DM è una delle proposte maggiormente strutturate per definire i passi fondamentali di un progetto di Data Mining.

Le sei fasi del ciclo di vita non sono strettamente sequenziali. È spesso necessario tornare su attività già svolte.

1) Comprensione del dominio applicativo: capire gli obiettivi del progetto dal punto di vista dell'utente, tradurre il problema dell'utente in un problema di data mining e definire un primo piano di progetto;

2) Comprensione dei dati: raccolta preliminare dei dati finalizzata a identificare problemi di qualità e a svolgere analisi preliminari che permettano di identificarne le caratteristiche salienti;

3) Preparazione dei dati: comprende tutte le attività necessarie a creare il dataset finale: selezione di attributi e record, trasformazione e pulizia dei dati;

4) Creazione del modello: diverse tecniche di data mining sono applicate al dataset anche con parametri diversi al fine di individuare quella che permette di costruire il modello più accurato;

5) Valutazione del modello e dei risultati: il modello/i ottenuti dalla fase precedente sono analizzati al fine di verificare che siano sufficientemente precisi e robusti da rispondere adeguatamente agli obiettivi dell'utente;

6) Deployment: il modello costruito e la conoscenza acquisita devono essere messi a disposizione degli utenti. Questa fase può quindi semplicemente comportare la creazione di un report oppure può richiedere di implementare un sistema di data mining controllabile direttamente dall'utente;

DATA MINING: DATA

Una **distribuzione di frequenza** (o **tabella di frequenza**) mostra come i dati si distribuiscono tra diverse categorie di valori (o classi), elencando le categorie e, per ognuna il numero di osservazioni corrispondente, detto frequenza. Il modo più semplice per rappresentare dei dati in una forma più o meno analizzabile è quello sotto forma tabellare.

Per esempio: avendo a disposizione una serie di cartelle cliniche con una serie di informazioni, messe in termini di rappresentazione verticale, e in ogni punto di incrocio è inserito un valore che rappresenta la frequenza di quell'item. Questa rappresentazione riporta dunque dei valori numerici che danno delle informazioni sulla frequenza di utilizzo di un reparto.

Record data

Dati costituiti da una raccolta di record, ciascuno di cui consiste in un insieme fisso di attributi.

Dati relativi alle transazioni

Rappresentano un tipo speciale di dati, dove ogni transazione coinvolge un insieme di elementi.

Consideriamo ad esempio un negozio di alimentari: l'insieme dei prodotti acquistati da un cliente durante un giro di shopping costituisce una transazione, mentre i singoli prodotti acquistati sono gli articoli.

TID	*Record*
1	Pane, Coca Cola, Latte
2	Birra, Pane
3	Birra, Coca Cola, Pannolini, Latte
4	Birra, Pane, Pannolini, Latte
5	Birra, Pannolini, Latte

Graph Data

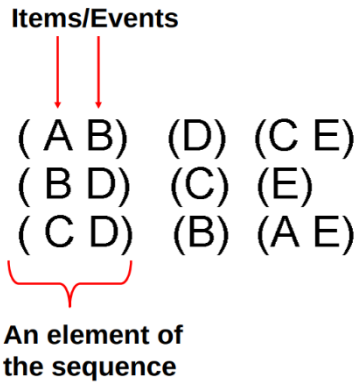
È possibile rappresentare i dati con un graph data, cioè prendendo le informazioni e rappresentandole secondo degli item e delle correlazioni. È inoltre possibile avere una rappresentazione a grafo banale sulle pagine web, per esempio. È possibile avere una correlazione tra informazioni che sono rappresentate mediante dei nodi (esempio: reparti collegati tra di loro, lo stesso medico collegato su reparti o la correlazione tra medici).

Ordered Data

È possibile avere dei dati che prendono le transazioni e le mettono in sequenza (una sequenza di informazioni), cioè una cosa associata ad un'altra.

Esempi:

Sequenze di transazioni



Dati di sequenza genomica

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Data Matrix

Si prendono le informazioni, si correlano tra di loro e il punto di incrocio della matrice dà qualche informazione. È possibile avere una matrice tridimensionale in cui si ha un'informazione, una presenza e un'altra informazione. Per esempio, la terza dimensione può essere il tempo; in questo caso il dato rappresenta un'informazione che è una combinazione dei tre dati. È possibile avere matrici n-dimensionali, dove il punto rappresenta la rappresentatività della combinazione di diversi altri punti.

L'insieme di dati può essere rappresentato da una matrice m per n, dove ci sono m righe, una per ogni oggetto, ed n colonne, una per ogni attributo.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Spatio-Temporal Data

Un'altra informazione è lo Spatio-Temporal Data, cioè si prendono le informazioni, si caratterizzano ed è necessario conoscere quel dato rappresentato in quell'intervallo di tempo. Per esempio, l'analisi epidemiologica dei dati.

Misura di distanza

La prima informazione si basa sulla tipologia dei dati, per la seconda informazione si prende il dataset e si rappresenta. È necessario definire la misura di distanza (definire delle metriche), cioè sapere se una cosa è simile ad un'altra.