4. Machine learning system

lunedì 28 ottobre 2024 18:03

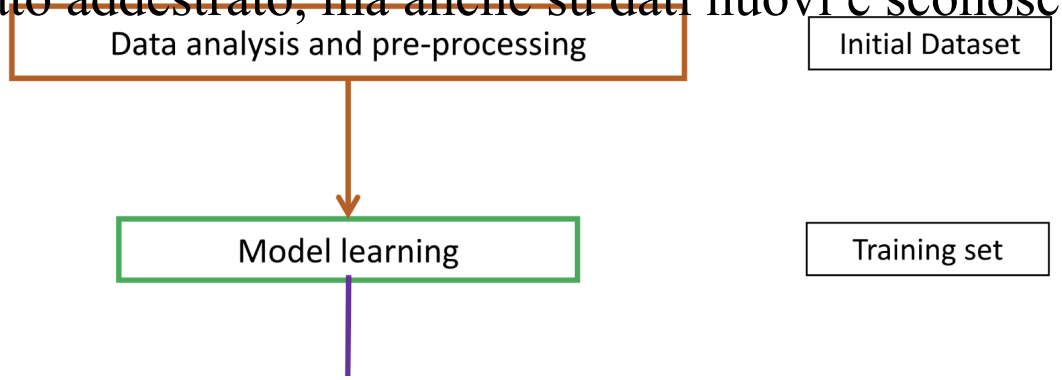
Un sistema di Machine Learning si basa su tre elementi princ
rappresentazione, ottimizzazione e valutazione.

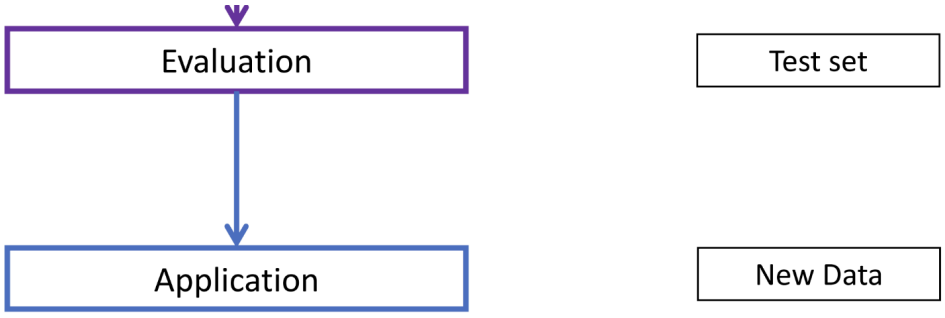
Rappresentazione: consiste nell'identificare lo spazio delle ip
modello e decidere quali caratteristiche (o features) utilizzare
rappresentare i dati. In altre parole, si sceglie il modo in cui i
presentati al modello per l'apprendimento.

Ottimizzazione: si tratta di scegliere il metodo per addestrare
Ad esempio, possono essere utilizzati metodi come la discesa
(Gradient Descent) o la ricerca golosa (greedy search) per trc
parametri migliori che ottimizzano le prestazioni del modello

Valutazione: implica la scelta di una funzione di valutazione
funzione di punteggio o funzione obiettivo) per distinguere u
modello da uno meno buono. Questa fase serve per capire qu
modello sta imparando e come si comporta sui dati.

Obiettivo: L'obiettivo finale di un sistema di Machine Learni
generalizzare bene dai dati visti durante l'addestramento (il tr
dati non visti, ovvero ai nuovi esempi. Ciò significa che il mc
dovrebbe essere in grado di fare previsioni accurate non solo
è stato addestrato, ma anche su dati nuovi e sconosciuti.





L'analisi e la pre-elaborazione dei dati sono fasi fondamentali perché i dati reali spesso sono "sporchi":

Incompleti: significa che il valore di alcuni attributi è mancante o che alcuni attributi interessanti sono completamente assenti.

Inaccurati: i dati contengono valori errati derivanti da osservazioni imprecise o parziali.

Il principio di GIGO (Garbage In, Garbage Out) si applica qui: se si immettono dati di scarsa qualità, anche i risultati saranno di scarsa qualità.

ella fase di pre-elaborazione dei dati (pre-processing) si svolgono diverse operazioni per migliorare la qualità dei dati e prepararli per l'analisi. Ecco una spiegazione dei principali passaggi:

Pulizia dei dati (Cleaning of data):

Rimozione degli outlier: eliminare i valori anomali che si discostano notevolmente dalla maggior parte dei dati, poiché possono influenzare negativamente l'analisi.
Rimozione del rumore: eliminare i valori casuali o distorsioni nei dati che possono alterare i risultati.

Rimozione dei duplicati: eliminare le voci ripetute nei dati, che potrebbero causare sovrastime o errori nell'analisi.
Modifica dei dati (Changing data):

Discretizzazione (Discretize): trasformare variabili continue in variabili discrete, ad esempio suddividendo un intervallo numerico in intervalli categorizzati (buckets).
Aggregazione (Aggregate): combinare i dati per ridurre la loro dimensione o per evidenziare tendenze. Ad esempio, sommare i dati giornalieri per ottenere dati settimanali.
Normalizzazione e riscalamento (Normalization and re-scaling): ridimensionare i valori dei dati per far sì che rientrino in un certo intervallo (ad esempio, tra 0 e 1), migliorando la performance di molti algoritmi di machine learning.Creazione di nuovi attributi: generare nuove variabili a partire dai dati esistenti, per rendere l'analisi più ricca e completa. Ad esempio, si può calcolare una variabile che rappresenta il rapporto tra due variabili esistenti.
DEF: Gli outlier (o valori anomali) sono dati che si discostano in modo significativo dalla maggior parte degli altri valori in un insieme di dati.

ESERCIZIO SPIEGAZIONE

14. Siano x = [117, 122, 138, 132, 147, 138, 121, 159, 121, 90], i valori rilevati della velocità di 10 auto transitano su una strada provinciale. Si calcoli il valore della mediana, della dispersione, del primo, secondo e ter Calcolare altresì lo scarto interquartile, la media aritmetica e gli eventuali outliers.

N è pari.

Dopo aver ordinato i dati in moto crescente:

B =

90 117 121 121 122 132 138 138 147 159

Mediana = 127
Dispersione = $159 - 90 = 69$
$Q_1 = 10 * 0.25 = 2.5 \rightarrow Q_1 = X_3 = 121$
$Q_2 = 10 * 0.50 = 5.0 \rightarrow Q_2 = (X_5 + X_6) = 127$
$Q_3 = 10 * 0.75 = 7.5 \rightarrow Q_3 = X_8 = 138$

Interquartile = IQR = Q_3-Q_1 = 17
Media=128.5

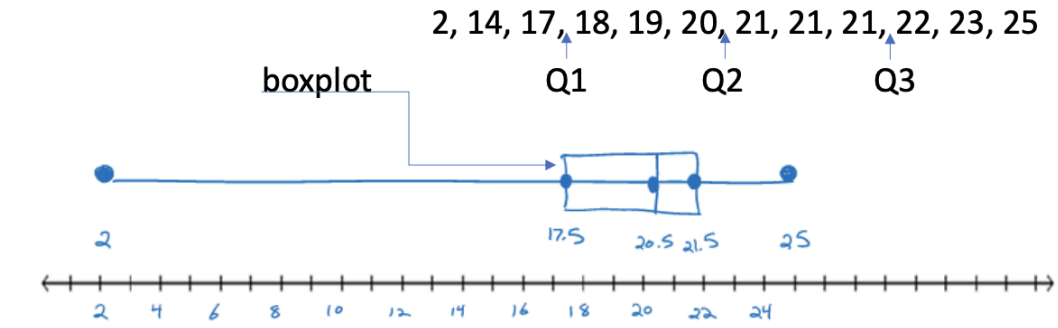
I valori limite sono:

Out_1 = Q1-1.5*IQR = 95.5
Out_2 = Q3+1.5*IQR = 163.5

per cui l'unico outlier è 90.

QUARTILES:
Q2 is equal to the median of the data set.
Q1 is the median of the values that are below Q2
Q3 is the median of the values that are above Q2

Quindi si contano le virgole e si prendono i valori a sinistra d



$Q_2 = \frac{20}{}$

$Q_1 = \frac{17}{}$

$Q_3 = \frac{21}{}$

IQR = Interquartile = Q3 - Q1

QUI ESSENZIALMENTE PRENDE LA MEDIANA Q2 (MEDIANA 6+7 VALORE)/2

POI PRENDE LA MEDIANA Q1 DEL 3 E 4 VALORE E POI FA DIVISO 2

PRENDE A GRUPPI DI MULTIPLI DI 2 E SI CALCOLA Q1 Q2 Q3

IQR = Q3 - Q1 = 21,5 - 17,5 = 4

TUTTO QUELLO CHE E' A SINISTRA O A DESTRA DEI VALORI LIMITE RISPETTIVAMENTE CON IL SEGNO SONO CONSIDERATI OUTLIER

$V_{ALORE\ LIMITE} = Q1 - 1.5 \cdot IQR = 11.5$

POICHE 2<11.5 è UN OUTLIER

$V_{L2} = Q3 + 1.5 \cdot IQR = 27,5$ NON CI SONO OUTLIERS

FEATURE SELECTION

La Feature Selection è una tecnica utilizzata per selezionare s caratteristiche (o feature) che sono importanti per un modello learning. Spesso ci sono molte caratteristiche disponibili, alcu quali potrebbero essere ridondanti o non rilevanti, e questo pu modello meno efficiente e più difficile da interpretare.

1. Domain experts: Coinvolgere esperti nel dominio specifico identificare le feature più rilevanti.
2. Filter: Misurare l'importanza di ciascuna feature per discri classi. Alcuni metodi includono Information Gain, Entropy e Information.
3. Wrappers: Metodo iterativo che cerca di trovare un buon s di feature valutando ogni sottoinsieme possibile.
4. Dimensionality reduction: Ridurre la dimensionalità dei da utilizzando tecniche come PCA (Principal Component Analy (Singular Value Decomposition).