Data and Web Mining

Raccolta domande teoriche

Quali sono le differenze tra gli algoritmi di supervised learning e quelli di unsupervised learning?

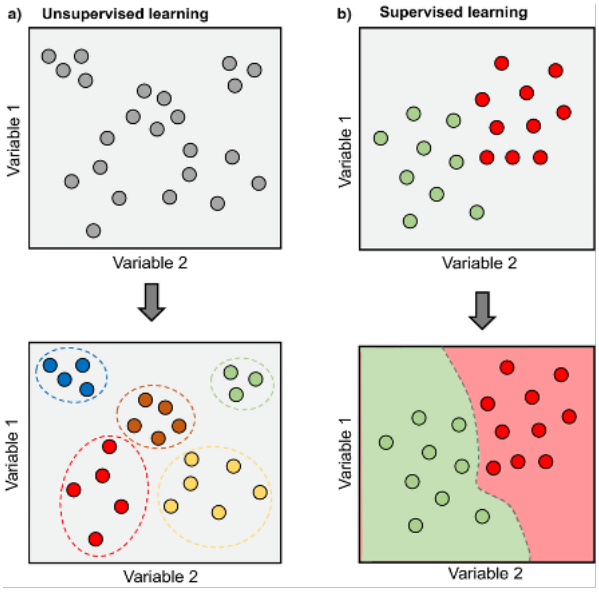
Supervised Learning:

- Obiettivo: Predire un'etichetta di output basata su una o più caratteristiche di input.
- Esempi di Applicazioni: Diagnosi mediche, riconoscimento vocale, riconoscimento di immagini, previsione del prezzo delle azioni.
- Valutazione del Modello: È possibile valutare la precisione del modello confrontando le previsioni del modello con le etichette vere.
- Complessità Computazionale: Tendenzialmente più elevata rispetto all'apprendimento non supervisionato, a causa della necessità di addestrare il modello con un grande numero di esempi etichettati.
- Overfitting: È un rischio maggiore in quanto il modello potrebbe adattarsi troppo ai dati di addestramento e perdere la capacità di generalizzare su dati nuovi e non visti.
- Esempio algoritmi: Decision Trees, Support Vector Machines, Neural Networks, Linear Regression.

Unsupervised Learning:

- Obiettivo: Esplorare la struttura sottostante o le relazioni tra le variabili nei dati.
- Esempi di Applicazioni: Segmentazione del mercato, organizzazione di grandi biblioteche di documenti, compressione di immagini.
- Valutazione del Modello: È più difficile valutare l'efficacia del modello, poiché non ci sono etichette vere con cui confrontare le previsioni o i raggruppamenti del modello.
- Complessità Computazionale: Tendenzialmente meno elevata rispetto all'apprendimento supervisionato, ma può variare a seconda dell'algoritmo e del numero di dati.
- Scoperta di Conoscenza: È particolarmente utile quando non si conoscono le etichette o quando si desidera scoprire relazioni non note tra i dati.
- Esempio algoritmi: K-Means, Hierarchical Clustering, DBSCAN, t-SNE.

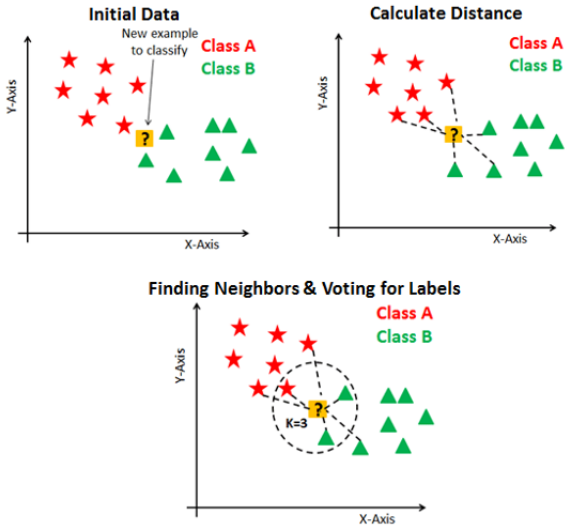
In generale, gli algoritmi di apprendimento supervisionato sono utilizzati per problemi di classificazione e regressione, mentre gli algoritmi di apprendimento non supervisionato sono utilizzati per problemi di clustering e riduzione della dimensionalità.



Descrivere l'algoritmo K-nn e spiegare cosa succede cambiando il parametro k

L'algoritmo K-nn, o K-Nearest Neighbors, è un metodo di apprendimento supervisionato utilizzato sia per la classificazione che per la regressione. Il suo funzionamento è intuitivo: dato un nuovo punto da classificare o da cui prevedere un valore, l'algoritmo identifica i k punti più vicini nel dataset di addestramento e assegna la classe o il valore medio di questi vicini al nuovo punto.
Funzionamento:

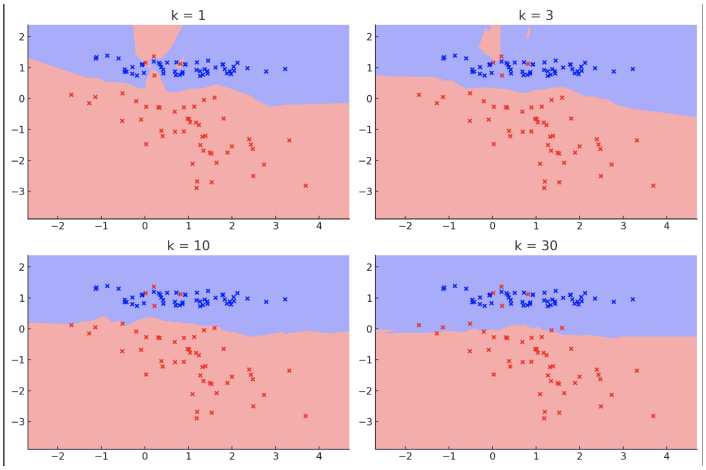
- Initial Data: Si parte con un dataset di addestramento dove ogni punto è etichettato con una classe o un valore.
- Calculate Distance: Per ogni nuovo punto, si calcola la distanza da tutti i punti nel dataset di addestramento. La distanza euclidea è comunemente utilizzata, ma altre metriche di distanza possono essere applicate a seconda del contesto.
- Finding Neighborhood & Voting for Labels: Si selezionano i k punti più vicini e, in base al task:
  - Classificazione: Si assegna la classe più frequente tra i k vicini (majority voting).
  - Regressione: Si assegna la media dei valori dei k vicini.



Il valore di k è cruciale. Un k troppo piccolo rende l'algoritmo sensibile al rumore, mentre un k troppo grande lo rende insensibile alle variazioni locali. È comune utilizzare un numero dispari per k in task di classificazione per evitare pareggi nel voting. Inoltre, è possibile attribuire pesi diversi ai vicini in base alla loro distanza dal nuovo punto, dando più importanza ai punti più vicini.

Per migliorare le prestazioni, è consigliabile scalare le feature in modo che abbiano tutte lo stesso range di variazione, utilizzando tecniche come MinMax Scaler o StandardScaler. Questo perché le feature con variazioni più ampie potrebbero dominare quelle con variazioni più ridotte nel calcolo delle distanze.

L'algoritmo K-nn è semplice ed efficace, soprattutto con dati numerici e quando si dispone di una metrica di distanza appropriata. Tuttavia, ha un costo computazionale elevato, soprattutto con dataset di grandi dimensioni, poiché richiede il calcolo della distanza da tutti i punti del dataset per ogni nuovo punto.



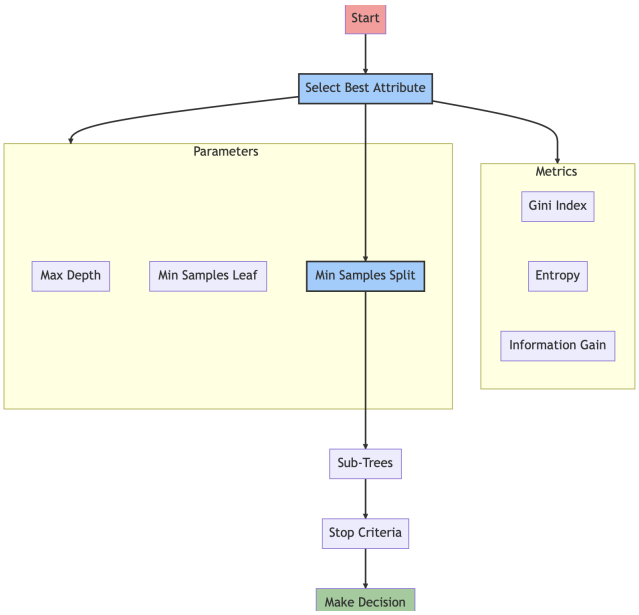
L'immagine mostra come l'algoritmo K-nn varia con diversi valori di k, illustrando l'effetto di overfitting con k piccoli e di underfitting con k grandi. Nello specifico abbiamo ogni grafico che rappresenta un diverso valore di k e mostra come l'algoritmo classifica i punti in due classi diverse (colori diversi nello sfondo) basandosi sui punti del dataset (punti colorati nel grafico).

- Nel primo grafico, con $k = 1$, si nota un evidente overfitting, con una frontiera di decisione molto irregolare che segue strettamente i punti del dataset.
- Nel secondo grafico, con $k = 3$, la frontiera di decisione è leggermente più liscia ma ancora abbastanza irregolare.

- Nel terzo grafico, con $k = 10$, la frontiera di decisione è più liscia e generalizzata, ma potrebbe ancora catturare bene la struttura dei dati.
- Nel quarto grafico, con $k = 30$ la frontiera di decisione è molto semplice e liscia, il che potrebbe indicare un potenziale underfitting, dove il modello non cattura adeguatamente la struttura sottostante dei dati.

Descrivere l'algoritmo per la costruzione di un Decision Tree e spiegare le metriche e i vari parametri

Gli Alberi di Decisione sono modelli di apprendimento supervisionato utilizzati per problemi di classificazione e regressione. Un albero di decisione divide ricorsivamente lo spazio degli input in regioni omogenee, per poi assegnare una classe (o un valore, in caso di regressione) alla regione in cui cade un nuovo input.



Algoritmo di Costruzione: la costruzione di un albero di decisione inizia con la radice, che contiene l'intero dataset di addestramento. Il dataset viene poi diviso in sottoinsiemi omogenei basandosi su un attributo e un valore di soglia. Questo processo si ripete ricorsivamente su ogni sottoinsieme, creando nuovi nodi, fino a quando tutti i dati in un nodo appartengono alla stessa classe o fino a quando non sono soddisfatti altri criteri di arresto.

Criteri di Divisione: i criteri di divisione, come l'Entropia e l'Indice Gini, misurano l'impurità di un nodo. Un nodo è puro quando tutti i suoi dati appartengono alla stessa classe. Il miglior attributo e valore di soglia da utilizzare per dividere un nodo sono quelli che riducono al massimo l'impurità.

- **Entropia:** Misura dell'incertezza o del disordine in un nodo.

$$\text{Entropia}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

- **Guadagno di Informazione:** Differenza tra l'entropia del nodo padre e la somma ponderata delle entropie dei nodi figli.

$$\text{Guadagno}(S, A) = \text{Entropia}(S) - \sum_{v \in \text{Val}(A)} \frac{|S_v|}{|S|} \text{Entropia}(S_v)$$

Criteri di Arresto: alcuni dei criteri di arresto includono la profondità massima dell'albero, il numero minimo di campioni per nodo e il numero minimo di campioni per foglia.

Metriche di Valutazione: le metriche di valutazione, come l'Accuracy, la Precision, la Recall e l'F1 Score, aiutano a quantificare le prestazioni di un modello di classificazione.

- **Accuracy:** Rapporto tra le previsioni corrette e il totale delle previsioni.
- **Precision:** Rapporto tra i veri positivi e la somma dei veri positivi e dei falsi positivi.
- **Recall:** Rapporto tra i veri positivi e la somma dei veri positivi e dei falsi negativi.
- **F1 Score:** Media armonica tra precision e recall.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Parametri del Modello: alcuni parametri chiave che influenzano la costruzione dell'albero di decisione sono la profondità massima dell'albero, il numero minimo di campioni richiesti per dividere un nodo interno e il numero minimo di campioni richiesti per essere presenti in un nodo foglia.

Cosa cambia in un decision Tree utilizzato in un task di classificazione a un decision Tree utilizzato per un task di regressione?

Un *Decision Tree* è un modello di apprendimento supervisionato utilizzato sia per problemi di classificazione che di regressione. La principale differenza tra i due tipi di alberi risiede nella natura della variabile obiettivo e nelle metriche di errore utilizzate per effettuare le divisioni.

Decision Tree per Classificazione
Nel contesto della classificazione, diverse metriche possono essere utilizzate per valutare la qualità delle divisioni:

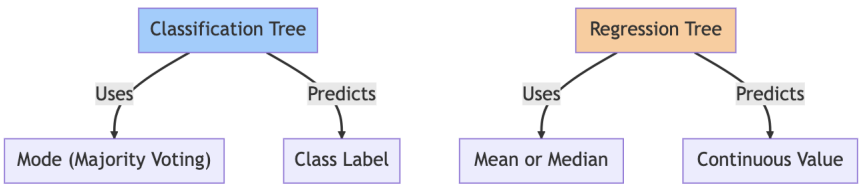
- **Classification Error:**
$$\text{Error} = 1 - \max(p_1, p_2, \ldots, p_m)$$
dove p_i rappresenta la proporzione di campioni appartenenti alla classe i in un nodo.
- **Information Gain:**
$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Val}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$
dove l'entropia è definita come:
$$\text{Entropy}(S) = -\sum_{i=1}^{m} p_i \log_2 p_i$$
- **Gain Ratio:**
$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{SplitInformation}(S, A)}$$
con:
$$\text{SplitInformation}(S, A) = -\sum_{v \in \text{Val}(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$
- **Gini Index:**
$$\text{Gini}(S) = 1 - \sum_{i=1}^{m} p_i^2$$
dove p_i è la proporzione di campioni appartenenti alla classe i in un nodo.

Decision Tree per Regressione
Nel contesto della regressione, l'obiettivo è prevedere un valore continuo piuttosto che una classe. La metrica di errore comunemente utilizzata per la divisione è il *Mean Squared Error* (MSE):

$$\text{MSE}(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2$$

dove y_i è il valore obiettivo del campione i e $\bar{y}$ è la media dei valori obiettivo in S.

La principale differenza nella struttura dell'albero tra classificazione e regressione risiede nelle foglie: in un albero di regressione, una foglia contiene la media dei valori obiettivo dei campioni in essa, mentre in un albero di classificazione, contiene la classe maggioritaria.



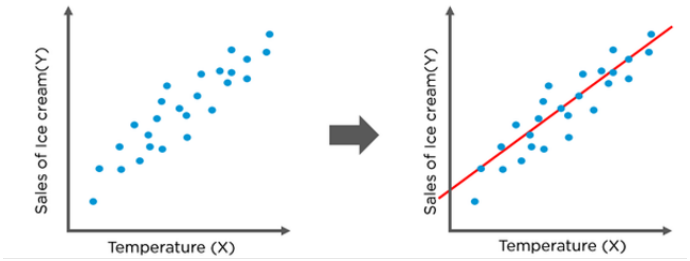
Descrivere un algoritmo efficiente per costruire un decision Tree

Un algoritmo efficiente per costruire un Decision Tree può essere realizzato utilizzando un approccio iterativo e una coda di priorità. L'algoritmo proposto si basa sulla massimizzazione del gain informativo a ogni passo per decidere il miglior attributo su cui effettuare lo split. Ecco i passi fondamentali dell'algoritmo:

```
1. Inizializza albero con nodo radice contenente tutto il training set
   Inizializza coda di priorità PQ

2. Per ogni nodo N nell'albero:
       Per ogni possibile split S di N:
           Calcola gain informativo di S
           Inserisci (N, S, gain) in PQ

3. Finché PQ non è vuota:
       Estrai (N, S, gain) da PQ con il massimo gain
       Esegui split S su nodo N
       Crea nodi figli destro e sinistro

4. Ripeti i passi 2 e 3 per ogni nuovo nodo creato
   Finché non sono soddisfatti i criteri di arresto

5. Se PQ è vuota o sono raggiunti i criteri di arresto:
       Termina algoritmo
```

Questo approccio consente una costruzione efficiente e ottimizzata dell'albero, garantendo che, ad ogni passo, si scelga lo split che massimizza il gain informativo. Tuttavia, è importante considerare gli iper-parametri e i criteri di arresto, in quanto possono influenzare significativamente la struttura dell'albero finale, potenzialmente portando a overfitting o underfitting del modello rispetto ai dati di addestramento.

Descrivere la regressione lineare e le metriche



La *regressione lineare* è un modello statistico ampiamente utilizzato, particolarmente adatto per task di regressione. Il modello è apprezzato per la sua semplicità, interpretabilità e adattabilità ai dati.
Questo approccio mira a trovare la retta, o in generale un iperpiano, che meglio approssima la distribuzione dei dati nel piano cartesiano. La forma della retta in due dimensioni è data da:

$$y = mx + b$$

dove:

- y è la variabile dipendente (risposta),
- x è la variabile indipendente (predittore),
- m è il coefficiente angolare (pendenza della retta),
- b è l'intercetta (punto in cui la retta interseca l'asse y).

L'obiettivo della regressione lineare è trovare i valori di m e b che minimizzano l'errore quadratico medio (MSE) tra i valori osservati e quelli predetti da modello:

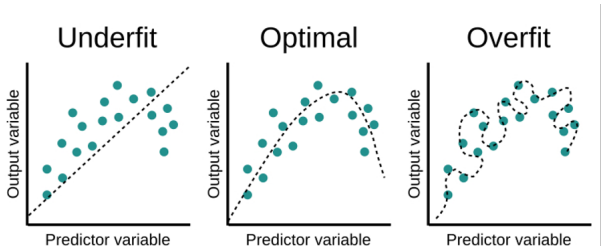
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

In alcuni casi, un modello lineare potrebbe non essere sufficientemente espressivo per catturare la complessità dei dati. In tali situazioni, si può ricorrere alla regressione polinomiale, che modella la relazione tra la variabile indipendente x e la variabile dipendente y come un polinomio di grado d:

$$y = a_d x^d + a_{d-1} x^{d-1} + \cdots + a_1 x + a_0$$

Tuttavia, aumentare eccessivamente il grado del polinomio può portare a overfitting, soprattutto quando il grado del polinomio è uguale al numero di punti nel dataset, rendendo il modello troppo complesso e incapace di generalizzare bene su dati non visti.

Cos'è l'overfitting?



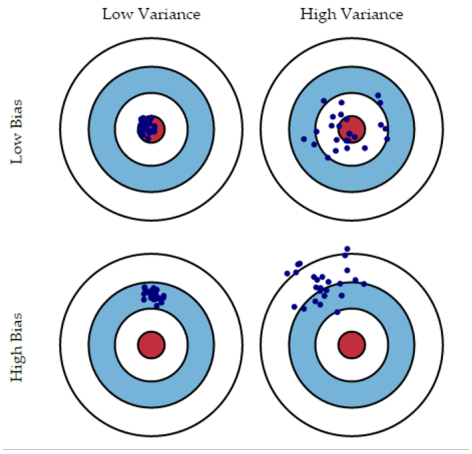
L'overfitting è un fenomeno critico in apprendimento automatico che si verifica quando un modello, addestrato eccessivamente bene sui dati di training, apprende non solo le relazioni sottostanti tra le variabili, ma anche il rumore presente nei dati. Questa eccessiva adattabilità ai dati di training impedisce al modello di generalizzare efficacemente su dati non visti, compromettendo così la sua utilità pratica. L'overfitting può essere attribuito a vari fattori, tra cui la complessità eccessiva del modello, l'eccessivo numero di parametri, o la scarsità dei dati di addestramento disponibili. In pratica, l'overfitting si manifesta con un'elevata accuratezza sui dati di training, ma con prestazioni scadenti su dati non visti o su un set di validazione.

Differenza tra bias e variance?

Il bias e la varianza sono due aspetti fondamentali dell'errore di previsione in un modello di apprendimento automatico.

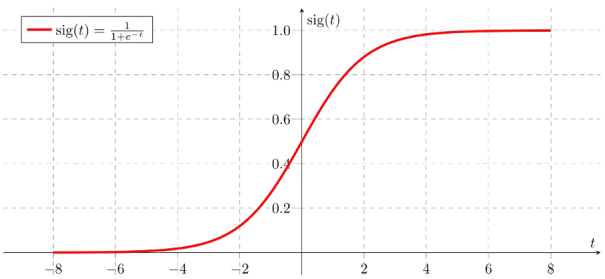
- Il bias misura quanto le previsioni del modello differiscono, in media, dal valore reale. Un alto bias generalmente indica che il modello è troppo semplice, non riuscendo a catturare la struttura sottostante dei dati, risultando in previsioni sistematicamente errate. Questo fenomeno è noto anche come underfitting.
- La varianza, invece, quantifica quanto le previsioni del modello sono sensibili a fluttuazioni nei dati di addestramento. Un modello con alta varianza è spesso troppo complesso, adattandosi eccessivamente ai dati di addestramento, inclusi il rumore e gli outlier, e risulta in un'incapacità di generalizzare bene su nuovi dati, un fenomeno noto come overfitting.

L'obiettivo nella costruzione di modelli di apprendimento automatico è trovare un equilibrio ottimale tra bias e varianza, minimizzando sia l'errore sistematico che la sensibilità alle fluttuazioni nei dati di addestramento, per costruire un modello che generalizzi efficacemente su dati non visti.



Descrivere la logistic regression

La *Regressione Logistica* è un algoritmo di apprendimento supervisionato utilizzato per problemi di classificazione binaria. Esso modella la probabilità che un'osservazione appartenga a una specifica classe utilizzando una funzione sigmoide, che mappa un input reale a un valore tra 0 e 1.



La probabilità che la variabile dipendente Y sia 1, è modellata come:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}$$

L'addestramento si basa sulla minimizzazione della *Log Loss*:

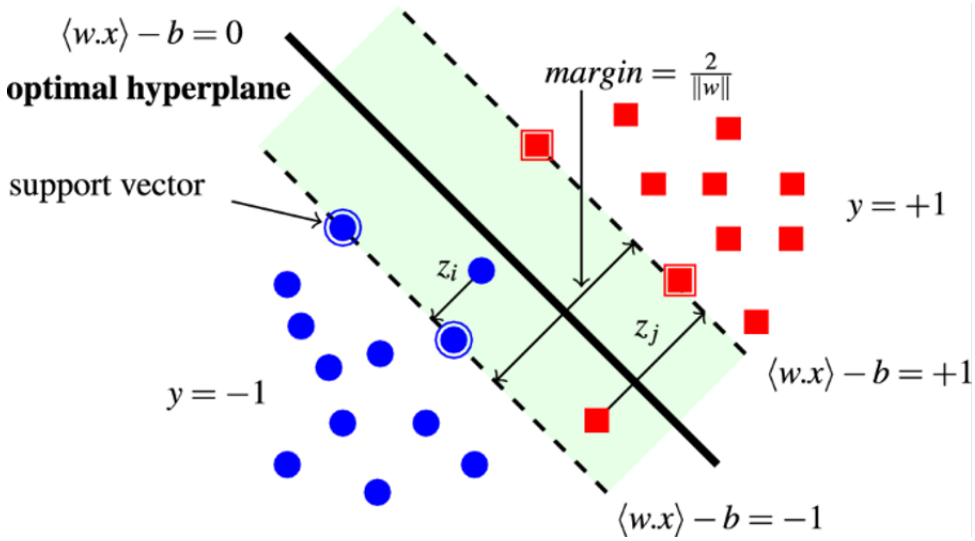
$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

mediante tecniche di ottimizzazione come la discesa del gradiente.

Valutazione e Interpretazione: Il modello, una volta addestrato, può essere utilizzato per classificare nuove osservazioni e i suoi coefficienti possono essere interpretati per comprendere l'importanza delle variabili indipendenti nel modello. La previsione del modello è considerata positiva quando il valore restituito dalla sigmoide è superiore a una determinata soglia (di solito 0,5), altrimenti è considerata negativa.

Limitazioni e Applicazioni: Nonostante la sua assunzione di linearità e altre limitazioni, la regressione logistica è ampiamente utilizzata in vari campi per la sua semplicità e interpretabilità.

Descrivere SVM, il suo funzionamento e le metriche



Le *Support Vector Machines* (SVM) sono uno strumento potente nel campo dell'apprendimento automatico, utilizzato principalmente per la classificazione, ma anche per la regressione. L'idea alla base di SVM è abbastanza semplice: immagina di avere dei dati che appartengono a due categorie diverse e di voler trovare la "linea" che li separa meglio.

In SVM, questa "linea" è chiamata iperpiano, e il "meglio" significa che l'iperpiano è il più lontano possibile dai punti più vicini di ogni categoria, chiamati *support vectors*. Se pensiamo ai dati come a punti in uno spazio, l'iperpiano è una sorta di "pavimento" che divide lo spazio in due parti, ognuna corrispondente a una categoria. L'iperpiano è definito come:

$$\vec{w} \cdot \vec{x} - b = 0$$

dove $\vec{w}$ è il vettore normale all'iperpiano e b è il termine di bias.

Il problema di ottimizzazione di base per un SVM è:

$$\min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|^2$$

soggetto a:

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 \quad \forall i$$

dove y_i sono le etichette di classe e $\vec{x}_i$ sono i vettori di caratteristiche. Solitamente è possibile avere un margine soft o un margine hard, ovvero possiamo permetterci di fare alcuni errori di misclassificazione lasciando che magari alcune istanze oltrepassino la retta, oppure non permettere che ci siano errori nelle predizioni. Infatti, in presenza di rumore o di dati non linearmente separabili, si introduce una variabile di slack ξ_i e un parametro di regolarizzazione C per permettere alcune violazioni del margine:

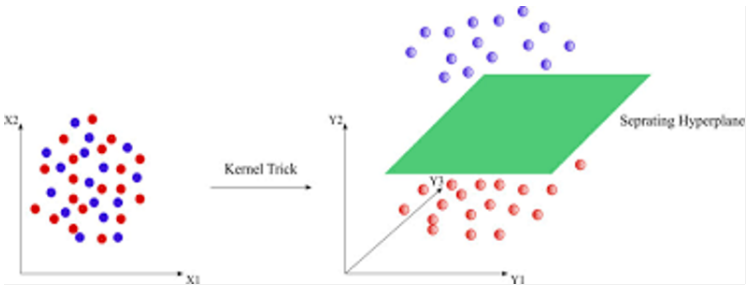
$$\min_{\vec{w},b,\xi} \frac{1}{2}\|\vec{w}\|^2 + C\sum_i \xi_i$$

soggetto a:

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i \quad \text{e} \quad \xi_i \geq 0 \quad \forall i$$

Kernel Trick: Per dati non linearmente separabili, SVM può essere esteso mediante l'utilizzo di funzioni kernel, che mappano implicitamente i dati in uno spazio di dimensione superiore in cui possono diventare linearmente separabili. Un kernel popolare è il kernel gaussiano (RBF):

$$K(\vec{x}, \vec{x}') = e^{-\gamma\|\vec{x}-\vec{x}'\|^2}$$

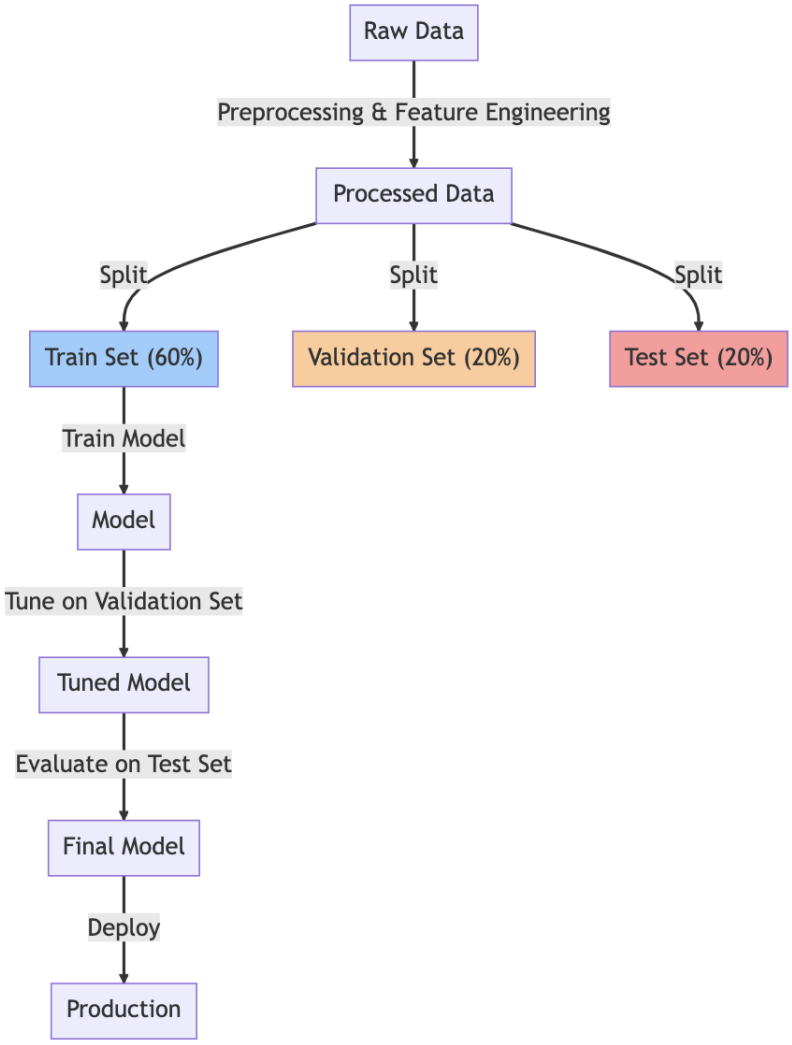


Cosa sono train validation e test

Nel campo del machine learning, il processo di sviluppo del modello richiede la divisione del dataset in tre sottoinsiemi distinti: train, validation e test. Questi sottoinsiemi hanno ruoli diversi nel processo di sviluppo e valutazione del modello.

- Train Set: Il train set è utilizzato per addestrare il modello. In questa fase, il modello apprende le relazioni e le patterns presenti nei dati, ottimizzando i suoi parametri per minimizzare l'errore nelle predizioni.
- Validation Set: Il validation set è utilizzato per valutare la performance del modello durante la fase di addestramento e per effettuare il tuning degli iperparametri. Questo set permette di identificare eventuali problemi come l'overfitting e di selezionare il modello migliore.
- Test Set: Il test set è utilizzato per valutare la performance del modello finale. Questa valutazione fornisce un'indicazione di come il modello si comporterà su dati non visti, rappresentando quindi una stima dell'errore di generalizzazione.

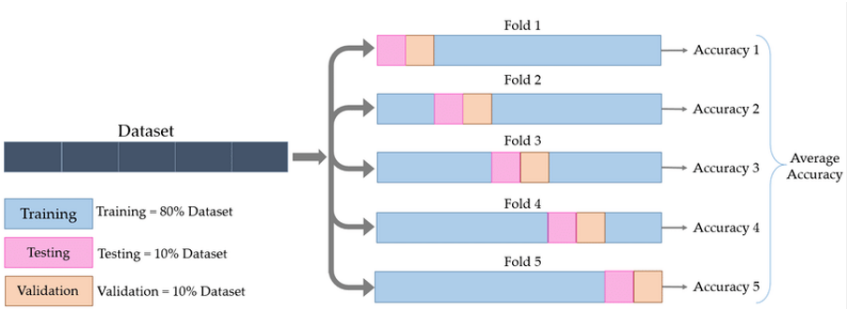
Una suddivisione comune dei dati è 60% train, 20% validation e 20% test. Tuttavia, è possibile utilizzare tecniche come la k-fold cross-validation per ottimizzare l'utilizzo dei dati disponibili, permettendo al modello di apprendere da diverse combinazioni dei dati.



Spiegare k-fold cross-validation

La k-fold cross-validation è un metodo utilizzato per valutare la performance di un modello di apprendimento automatico su un insieme di dati. Il metodo consiste nel dividere i dati in k "gruppi" di uguali dimensioni e quindi addestrare il modello su k-1 gruppi e testarlo sull'ultimo gruppo. Ciò viene ripetuto k volte, in modo che ogni gruppo venga utilizzata almeno una

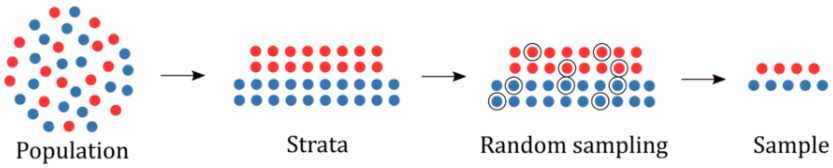
volta come set di dati di test. Alla fine, i risultati delle prove verranno combinati per ottenere una valutazione più precisa della performance del modello tramite la media.



Per esempio, se k = 5, i dati verranno divisi in 5 gruppi e il modello verrà addestrato su 4 di questi gruppi e testato sull'ultimo. Ciò verrà ripetuto 5 volte, utilizzando un gruppo diverso come set di dati di prova ogni volta. Alla fine, i risultati delle 5 prove verranno combinati per ottenere una valutazione più precisa della performance del modello tramite la media.

Il vantaggio principale della k-fold cross-validation è che utilizza tutti i dati disponibili per l'addestramento e la valutazione del modello. Ciò significa che non si perde alcun dato prezioso e si ottiene una valutazione più precisa della performance del modello rispetto ad altri metodi di validazione come la semplice divisione in training/test set. Inoltre, la k-fold cross-validation permette di valutare la robustezza del modello e di individuare eventuali problemi di overfitting o underfitting. Infatti, se un modello soffre di overfitting, i risultati sui set di dati di prova diverranno peggiori man mano che si effettuano più prove.

Cos'è lo stratified sampling e quando si utilizza



Lo stratified sampling è una tecnica di campionamento che mira a garantire che ogni sottogruppo della popolazione sia rappresentato adeguatamente nel campione finale. Questa tecnica è particolarmente utile quando la popolazione è eterogenea e presenta diverse sottopopolazioni o strati.

Nello stratified sampling, la popolazione totale viene divisa in diversi strati, o gruppi, in base a specifiche caratteristiche o attributi, come l'età, il genere, il livello di istruzione, ecc. Una volta identificati gli strati, vengono selezionati campioni casuali da ciascuno strato in modo proporzionale alla loro presenza nella popolazione totale. Questo processo garantisce che ogni strato sia rappresentato nel campione finale.

Vantaggi

- Rappresentatività: Lo stratified sampling garantisce una rappresentazione equa di tutti gli strati della popolazione, migliorando l'accuratezza e la precisione delle stime statistiche.
- Riduzione dell'Errore di Stima: La stratificazione riduce la varianza e l'errore di stima, poiché ogni strato è più omogeneo rispetto alla popolazione totale.
- Analisi Specifica degli Strati: Permette analisi più dettagliate e specifiche per ciascun strato, facilitando lo studio di sottopopolazioni specifiche.

Lo stratified sampling è particolarmente utile quando:

- dati sono eterogenei: Quando la popolazione è composta da diversi sottogruppi o strati con caratteristiche diverse.
- Presenza di sottopopolazioni di interesse: Quando è necessario analizzare specifiche sottopopolazioni o strati all'interno dei dati.
- I dati sono sbilanciati: Quando esiste uno sbilanciamento significativo tra le diverse categorie o strati nei dati, lo stratified sampling aiuta a ottenere un campione più bilanciato e rappresentativo.

Esempio: se si dispone di un dataset di pazienti e si vuole analizzare l'incidenza di una certa malattia, ma la prevalenza della malattia è diversa tra uomini e donne, si potrebbe utilizzare lo stratified sampling per garantire che il campione finale contenga un numero adeguato di uomini e donne, permettendo così analisi più accurate e affidabili.

Cosa significa fare tuning dei parametri?

Questo processo mira a trovare la configurazione ottimale degli iperparametri di un modello, ovvero quei parametri che non vengono appresi durante l'addestramento, ma che influenzano significativamente le prestazioni del modello.

L'obiettivo principale del tuning dei parametri è migliorare la capacità del modello di generalizzare su dati non visti, ottimizzando una metrica di performance specifica, come l'accuratezza, la precisione, il recall, l'F1-score, l'AUC-ROC, o la log-loss, a seconda del problema in questione.

Il tuning dei parametri si basa sull'uso di un set di validazione, su cui vengono testate diverse combinazioni di iperparametri, evitando così il rischio di overfitting sul set di test. Le principali tecniche di tuning sono:

- Grid Search: Esplora sistematicamente tutte le combinazioni possibili di valori degli iperparametri definiti a priori. È efficace ma computazionalmente costoso.
- Random Search: Esplora combinazioni casuali di valori degli iperparametri, offrendo un buon compromesso tra efficacia e costo computazionale.
- Ottimizzazione Bayesiana: Utilizza modelli probabilistici per guidare la ricerca dei valori ottimali, risultando spesso più efficiente delle tecniche precedenti.

Il tuning dei parametri è fondamentale per:

- Ottimizzare le Prestazioni: Trovare la configurazione ottimale degli iperparametri può significare la differenza tra un modello mediocre e un modello eccellente.
- Evitare Overfitting e Underfitting: Un tuning adeguato può aiutare a bilanciare la capacità del modello di apprendere dai dati e di generalizzare su nuovi dati.
- Adattare il Modello al Problema Specifico: Ogni problema ha le sue peculiarità, e il tuning permette di adattare il modello alle specifiche esigenze del problema.

Cos' è un classificatore baseline?

Un classificatore baseline, o modello baseline, è un modello semplice e fondamentale utilizzato tendenzialmente come punto di partenza nel processo di sviluppo del modello. Serve come riferimento per valutare le prestazioni di modelli più complessi e avanzati. L'obiettivo è superare le prestazioni del classificatore baseline con modelli più sofisticati.

In Task di Classificazione: In un task di classificazione, un classificatore baseline è spesso un modello che predice sempre la classe più frequente nel dataset di addestramento, indipendentemente dalle caratteristiche dell'input. Questo tipo di modello è anche conosciuto come "ZeroR" o "Majority Class Classifier".

In Task di Regressione: In un task di regressione, il classificatore baseline è generalmente un modello che predice sempre la media (o la mediana) del target nel dataset di addestramento, senza considerare il valore delle variabili indipendenti. Questo è spesso chiamato "Mean" o "Median Predictor".

Importanza del Classificatore Baseline:

- Punto di Partenza: Fornisce un punto di partenza semplice e intuitivo nel processo di modellazione.
- Benchmark: Serve come benchmark per valutare se modelli più complessi offrono miglioramenti significativi.
- Valutazione delle Prestazioni: Aiuta a stabilire un minimo livello di accettabilità per le prestazioni del modello.
- Rapidità e Semplicità: È veloce da implementare e richiede poco sforzo computazionale, permettendo una valutazione rapida del problema.

Come funziona il modello naive bayes

Si tratta di una tecnica semplice per costruire classificatori. Nonostante l'assenza di un algoritmo univoco per allenare tali classificatori, il principio comune su cui si fondano è la presupposizione che ciascuna feature sia completamente indipendente dalle altre. Questo modello prende il nome dal *Teorema di Bayes*, formalizzato come segue:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Dove:

- $P(A|B)$ è la probabilità a posteriori di A dato B.
- $P(B|A)$ è la probabilità a priori, ovvero la probabilità di osservare B dato A.

- $P(A)$ e $P(B)$ sono le probabilità marginali di A e B rispettivamente.

Nel contesto del *machine learning*, il problema si traduce nel trovare la classe C_k che massimizza la probabilità a posteriori, ovvero:

$$C_k = \arg\max_k P(C_k|\text{feature}_1, \text{feature}_2, \ldots, \text{feature}_n)$$

Sotto l'assunzione di indipendenza condizionale tra le features, si ottiene:

$$P(\text{feature}_1, \text{feature}_2, \ldots, \text{feature}_n|C_k) = \prod_{i=1}^{n} P(\text{feature}_i|C_k)$$

Pertanto, la predizione del modello sarà la classe che massimizza la probabilità a posteriori. Questo è fondamentalmente l'approccio adottato dai classificatori Naive Bayes nel *machine learning*.

Spiegare il criterio di splitting basato su Information Gain Index

DA FARE

Spiegare i metodi principali per processare testo

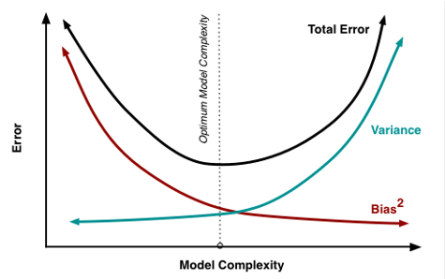
Lo scopo del text-processing nasce dalla necessità di poter salvare e computare input testuali con il minor numero di risorse possibili, mantenendo le stesse informazioni.

Alcuni metodi per fare ciò sono:

- Tokenization: si tratta il testo come una serie di token, che possono essere parole frasi o paragrafi;
- Lemming: si rimuovono i suffissi
- STemming: si mantiene la forma base delle parole, previa analisi morfologica;
- Rimozione delle "stopword": si rimuovono tutti i termini che rappresentano punteggiature, congiunzioni, articoli e preposizioni che non hanno un contributo significativo al valore del testo;
- Normalization: elimina sistematicamente lettere maiuscole e/o segni di punteggiatura dal testo, per ridurre il numero di caratteri che si possono incontrare;
- Aggiunta di informazioni semantiche: in pratica come l'identificazione di entità, relazioni o azioni presenti nel testo

Un altra tecnica è l'impiego di shingling, min-hashing o locally sensitive hashing (LSH) spiegato nelle domande

Spiegare bias-variance decomposition



La *decomposizione di bias-variance* è fondamentale in apprendimento automatico, permettendo di analizzare l'errore di previsione di un modello. L'errore totale può essere scomposto in *bias*, *varianza*, ed *errore irriducibile*.

Componenti dell'Errore

- **Bias:** Misura la differenza tra le previsioni del modello e i valori reali. Alto bias causa *underfitting*.
- **Varianza:** Misura le variazioni delle previsioni per diversi insiemi di dati di addestramento. Alta varianza causa *overfitting*.
- **Errore Irriducibile:** Errore intrinseco, dovuto al rumore nei dati, che non può essere ridotto.

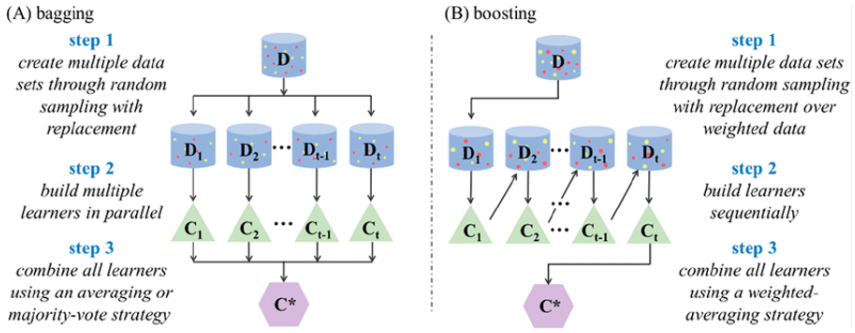
Formula dell'Errore Totale:
L'errore totale E può essere rappresentato come:

$$E = \text{Bias}^2 + \text{Varianza} + \text{Errore Irriducibile}$$

Trade-off Bias-Varianza: Esiste un trade-off tra bias e varianza. L'obiettivo è minimizzare l'errore totale attraverso tecniche come la regolarizzazione e l'ottimizzazione degli iperparametri, bilanciando bias e varianza.

Soluzioni:
Le soluzioni includono la selezione delle caratteristiche, la regolarizzazione, e l'ottimizzazione degli iperparametri per trovare un equilibrio ottimale e ridurre l'errore complessivo.

Spiegare quando è utile usare bagging(descriverlo) e quando invece è utile usare boosting(descriverlo)



Bagging e Boosting sono due tecniche di ensemble utilizzate per migliorare la performance di un modello di apprendimento automatico.

- Bagging: Il Bagging è una tecnica che consiste nel creare più copie del modello originale, ognuna addestrata su un sottoinsieme casuale dei dati di addestramento. Il risultato finale è un insieme di modelli che possono essere utilizzati insieme per fare previsioni, come la media o la moda delle previsioni dei singoli modelli in base al task richiesto. È particolarmente utile per modelli ad alto rischio di overfitting, come gli alberi di decisione profondi.
- Boosting: Il Boosting è una tecnica che consiste nell'addestrare una serie di modelli in sequenza, dove ogni modello cerca di correggere gli errori del modello precedente. Il risultato finale è un insieme di modelli che lavorano insieme per fare previsioni, come la somma ponderata delle previsioni dei singoli modelli. È particolarmente utile per modelli a basso rischio di overfitting, come i modelli lineari.

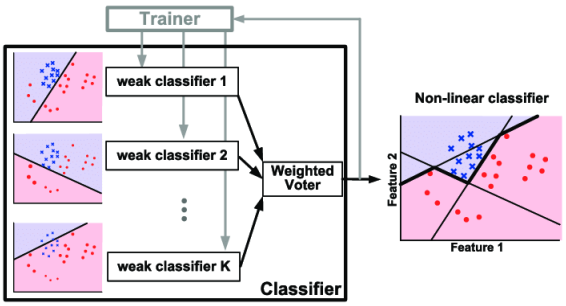
Scelta della Tecnica: La scelta tra bagging e boosting dipende dalla natura del problema e dalla tipologia del modello. Se il modello soffre di alta varianza, il bagging potrebbe essere la scelta migliore. Se il modello ha un alto bias, il boosting potrebbe essere più appropriato.

Compromesso Bias-Varianza: Entrambe le tecniche cercano di ottimizzare il compromesso bias-varianza, ma lo fanno in modi diversi. Il bagging mira a ridurre la varianza senza aumentare il bias, mentre il boosting mira a ridurre il bias senza aumentare eccessivamente la varianza.

Spiegare adaboost

AdaBoost, o Adaptive Boosting, è un algoritmo di boosting particolarmente efficace e popolare. L'obiettivo di AdaBoost è di combinare i punti di forza di molti modelli deboli per creare un modello forte ed accurato. Qui di seguito è spiegato più dettagliatamente il funzionamento di AdaBoost:

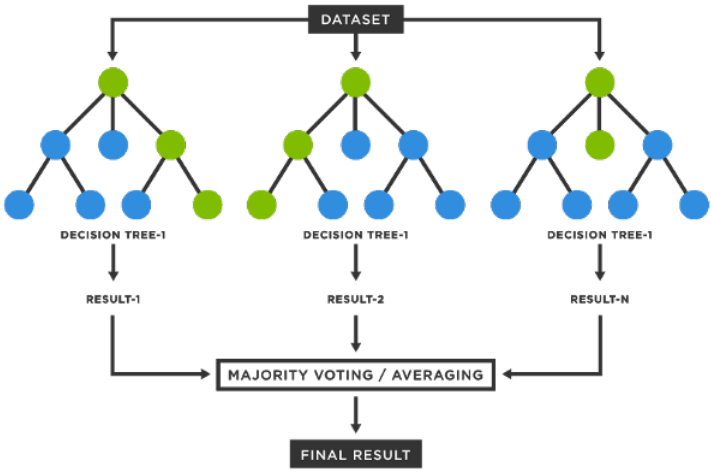
1. Inizializzazione dei Pesi: Ogni osservazione nel dataset inizialmente ha lo stesso peso, $1/n$, dove n è il numero totale di osservazioni.
2. Creazione del Modello Debole: AdaBoost crea un modello debole, solitamente un albero di decisione con un solo livello (stump).
3. Calcolo dell'Errore: L'errore del modello è calcolato come la somma dei pesi associati alle osservazioni classificate in modo errato.
4. Calcolo dell'Importanza del Modello: AdaBoost calcola l'importanza del modello debole in base al suo errore.
5. Aggiornamento dei Pesi: I pesi delle osservazioni sono aggiornati in modo che le osservazioni classificate in modo errato ricevano più peso, mentre quelle classificate correttamente ricevono meno peso.
6. Iterazione: Il processo è ripetuto, creando e aggiungendo modelli deboli al modello complessivo finché non si raggiunge un numero predeterminato di modelli o finché il modello complessivo non classifica perfettamente il training set.
7. Creazione del Modello Finale: Il modello finale è una combinazione ponderata dei modelli deboli, in cui il peso di ogni modello è determinato dalla sua accuratezza.



Vantaggi di AdaBoost:

- È in grado di ridurre il bias e la varianza.
- È efficace con dataset sbilanciati.
- Può essere utilizzato con vari tipi di modelli di apprendimento.

Spiegare random forest

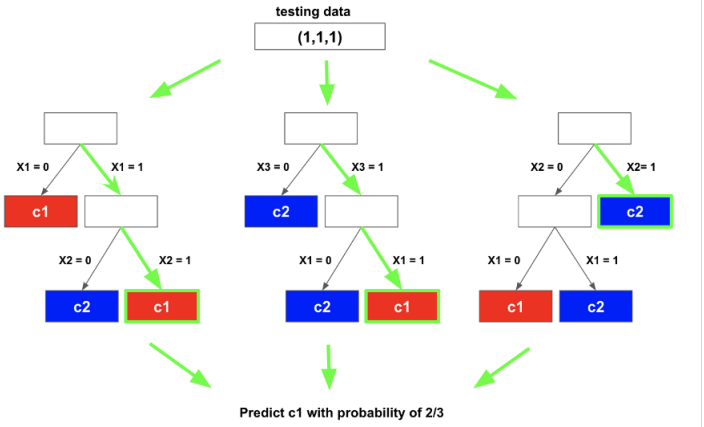


Random Forest è un algoritmo molto utilizzato di bagging, ovvero si utilizza per cercare di abbassare la varianza.
L'idea generale di un algoritmo di boosting solitamente è:

1. si prende il dataset originale e da questo si estrae tramite la tecnica di bootstrap un campione (campionamento con rimpiazzamento) e la dimensione del campione pescato è uguale alla dimensione del dataset iniziale;
2. Dopo si allenano i diversi modelli in parallelo sui vari campioni pescati e la predizione sarà in base al task che dobbiamo risolvere: una media di tutte le previsioni dei vari modelli nel caso di regressione e la moda nel caso invece di classificazione.

Ciò per cui differisce la random forest è che è importante che i modelli siano il più possibile indipendenti, per questo motivo random forest oltre a fare il procedimento descritto, durante la creazione di questi alberi che devono essere fully grown (con basso bias e elevata varianza) si cerca di renderli il più diversi possibili utilizzando la random input selection ovvero si utilizza un sottoinsieme di features diverse su ogni modello che si allena in modo tale che ogni albero sia il più possibile diverso dagli altri

Come è possibile usare la random forest per stimare la similarità



Per stimare la similarità tra due istanze con una Random Forest, si osserva in quali foglie delle diverse alberi le due istanze finiscono. Se due istanze finiscono frequentemente nelle stesse foglie attraverso i diversi alberi della foresta, si può inferire che sono simili tra loro. Questo perché le istanze che percorrono gli stessi cammini e finiscono nelle stesse foglie hanno caratteristiche simili, secondo i criteri di divisione degli alberi.
Procedura:

- Addestramento della Foresta: Addestra una Random Forest sul tuo set di dati.
- Percorso delle Istanze: Per ogni istanza, registra le foglie in cui cade in ogni albero della foresta.
- Calcolo della Similarità: Per ogni coppia di istanze, calcola la percentuale di alberi in cui cadono nella stessa foglia. Una percentuale più alta indica una maggiore similarità.

Questa misura di similarità può essere utilizzata in vari contesti, come il clustering, la riduzione della dimensionalità, o come feature in altri modelli di machine learning.
Vantaggi

- Robustezza: La Random Forest è resistente agli outlier e può gestire bene le variabili categoriche e continue.
- Interpretabilità: La similarità basata sulla Random Forest può fornire intuizioni intuitive, poiché si basa sulla struttura degli alberi di decisione.

Esempio: considera due istanze A e B. Se, in una Random Forest di 100 alberi, A e B cadono nella stessa foglia in 85 alberi, lo score di similarità sarà 85%.

Come è possibile usare la random forest per identificare gli outliers?

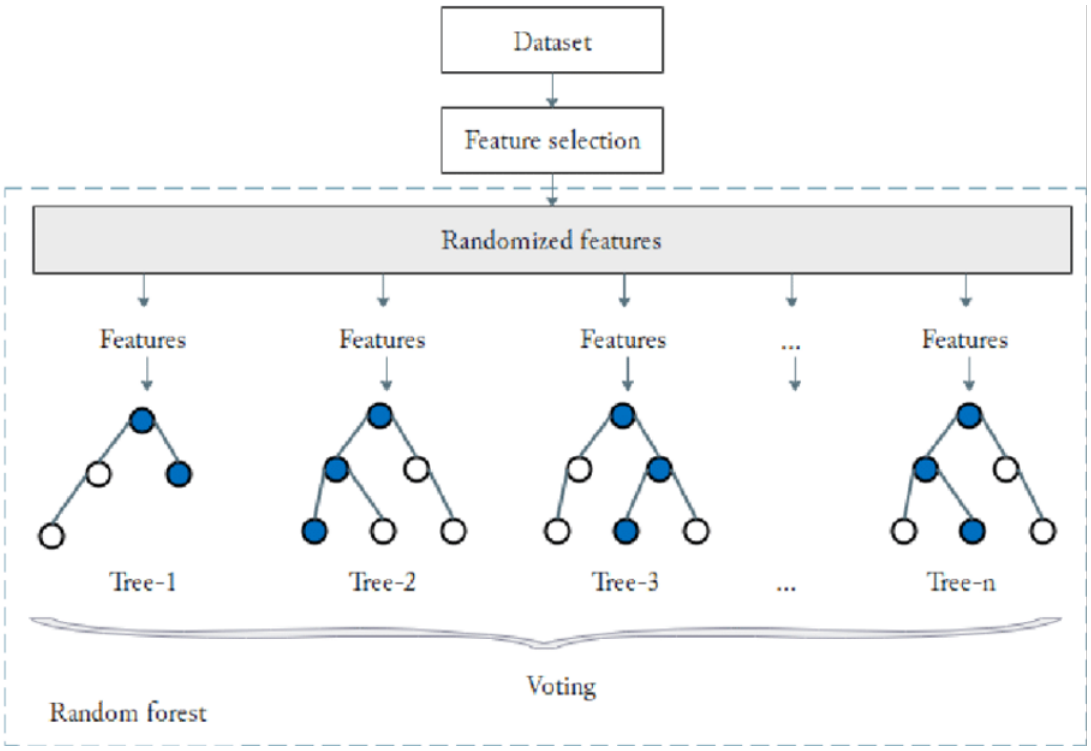
Per identificare gli outliers si può utilizzare un procedimento simile a quelli visto sopra ovvero dato un punto per capire se è un outlier si può misurare il suo score come outlier misurando la sua dissimilarità dal resto dei punti come 1 / la similarità calcolata dalla random forest, ovviamente più sarà alto lo score più sarà probabile che sia un outlier.

Come è possibile usare la random forest per rimpiazzare i valori mancanti?

È possibile usare una random forest per rimpiazzare i valori mancanti infatti grazie ad essa possiamo per un valore mancante di una determinata feature trovarlo.
Il procedimento descritto brevemente si basa su calcolare la similarità basata sulla random forest dell'istanza dove manca il valore di una determinata feature con tutto il resto delle istanze e per ognuna di queste si prende dove è presente il valore della future che vogliamo sistemare e si fa una media pesata sulla base della similarità calcolata dalla random forest.

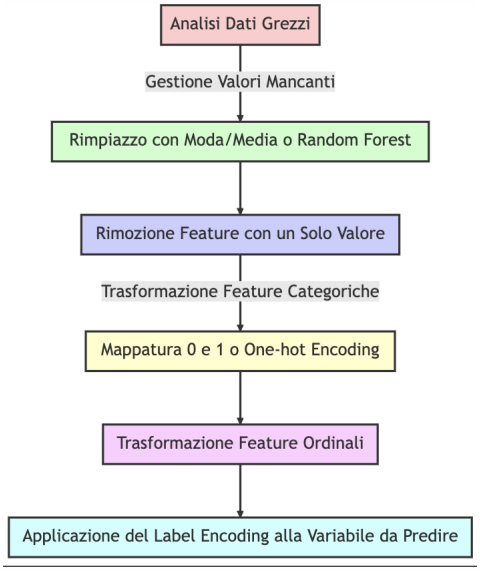
Questo procedimento può essere iterato più volte sui vari valori mancanti del dataset. Solitamente un altro procedimento utilizzato è quello di rimpiazzare un valore numerico con la media dei valori che assume quella feature oppure se categorico la moda.

Come è possibile fare feature selection con la random forest?

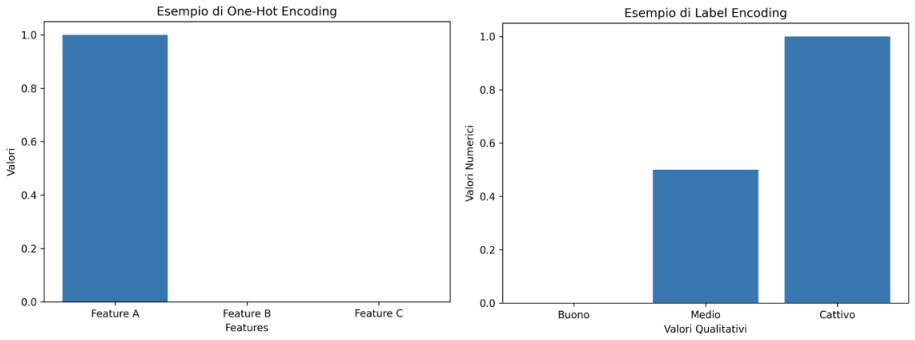


È possibile inoltre utilizzare la random forest per fare feature selection e quindi evitare il problema della dimensionalità. Infatti grazie alla random forest, è possibile salvare durante la fase di creazione dei diversi alberi quali sono stati gli split (feature e treshold) che hanno portato a un maggior guadagno, in tal modo poi è possibile ordinare le varie feature in base proprio all'importanza determinata dal gain a splittare, e sulla base di questo le feature meno importanti si può decidere di scartare le feature meno importanti e di riapplicare il procedimento più volte fino a che non abbiamo ottenuto un numero di feature ridotto.

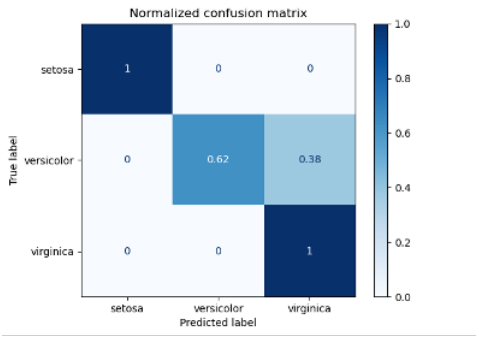
Come si esegue il feature engineering



Il feature engineering è una delle parti più importanti durante la creazione di un modello predittivo. È la fase in cui dai dati raccolti (grezzi) si analizzano e si trasformano cercando di tenere solamente quelli che si ritengono importanti e rimuovendo gli outliers. Innanzitutto solitamente si osserva se sono presenti valori mancanti nel dataset e solitamente vengono rimpiazzati con la moda nel caso sia una feature categorica o con la media se è una feature numerica (se si vuole è anche possibile utilizzare una random forest per stimare i valori mancanti).
Dopo aver sistemato i valori mancanti si osserva se ci sono feature contenenti solo un valore: in questo caso si possono scartare perché non forniscono informazioni interessanti, e poi si passa alla trasformazione delle feature categoriche dato che la maggior parte degli algoritmi di machine learning funzionano bene con valori numerici.
Se la feature è categorica e ha solamente due valori si può rimappare tra 0 e 1, altrimenti se ha k valori si può utilizzare la tecnica del one-hot encoding ovvero per ogni valore che assume la feature categorica che vogliamo trasformare viene aggiunta una colonna binaria dove con 1 indichiamo la presenza di quel valore e con 0 l'assenza. Inoltre sono presenti anche le feature ordinali che assumono ad esempio valori come (buono, medio, cattivo) solitamente queste vengono rimappate numericamente mantenendo l'ordine qualitativo ad esempio (0, 0.5, 1).
Inoltre per la variabile da predire se è categorica si può applicare il label encoding, ovvero si rimappa in id numerici.



Cos'è la confusion matrix



La Confusion Matrix è un elemento chiave nella valutazione dei modelli di classificazione nell'apprendimento supervisionato. Essa fornisce una rappresentazione tabulare delle performance del modello, confrontando le classi reali con quelle predette. In una matrice binaria, si distinguono True Positive (TP), True Negative (TN), False Positive (FP) e False Negative (FN), che rappresentano rispettivamente le classificazioni corrette e quelle errate di ciascuna classe.

In una classificazione multiclasse, la matrice si espande per includere queste categorie per ogni classe. La diagonale principale rappresenta le classificazioni corrette, mentre gli elementi fuori dalla diagonale indicano errori. Un modello ideale avrebbe solo valori non nulli sulla diagonale.

Dalla Confusion Matrix si derivano metriche cruciali come Accuratezza, Precisione, Recall e F1 Score, che permettono di valutare l'efficacia del modello. Questa matrice è particolarmente utile per identificare le aree di debolezza del modello e ottimizzare le sue performance, essendo rappresentabile anche visivamente attraverso un heatmap, facilitando così l'interpretazione dei risultati.

Quali sono le principali metriche di prestazione di un modello

Le metriche per stabilire le prestazioni di un modello dipendono fortemente dal tipo di task che tale modello deve portare a termine.

- Classificazione:
  - **Accuracy**: La percentuale di predizioni corrette prodotte dal modello.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

    Esempio: Se il modello ha correttamente predetto 80 su 100 campioni, l'accuracy è del 80%.
  - **Precision**: Indica la proporzione di identificazioni positive che sono effettivamente corrette.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

    Esempio: Se il modello ha identificato 50 positivi, ma solo 40 sono veri positivi, la precision è 0.8.
  - **Recall**: Mostra la proporzione di effettivi positivi che sono stati identificati correttamente.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

    Esempio: Se ci sono 50 positivi reali e il modello ne ha identificati 40, il recall è 0.8.
  - **F1-Score**: Media armonica tra Precision e Recall.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

    Esempio: Con Precision=0.8 e Recall=0.8, l'F1-Score è 0.8.
  - **AUC-ROC**: Area sotto la curva ROC. Un valore di 1.0 indica una classificazione perfetta.(vedi domanda successiva)
  - **Log-Los**: Misura la performance di un modello di classificazione.

$$\text{Log-Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))]$$

    Esempio: Un modello con log-loss inferiore ha migliori performance.
- Regressione:

– **MSE**: Calcola la media dei quadrati degli errori.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Esempio: Differenza tra i valori osservati e quelli predetti.

– **R-Squared**: Indica la percentuale di varianza della variabile dipendente spiegata dal modello.

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

Esempio: Un R^2 di 0.8 indica che l'80% della varianza è spiegata dal modello.

Spiegare cos'è la ROC curve (e AUC) / Come si valuta un classificatore binario

La curva ROC (Receiver Operating Characteristic) è uno strumento grafico fondamentale per valutare la capacità di un modello di classificazione binaria di distinguere tra le classi positive e negative. Essa traccia il Tasso di Veri Positivi (True Positive Rate, TPR) contro il Tasso di Falsi Positivi (False Positive Rate, FPR) a vari livelli di soglia di classificazione.
TPR (Sensibilità o Recall): È la proporzione di osservazioni positive reali che sono correttamente identificate dal modello.

True Positive Rate (TPR) o Sensibilità:

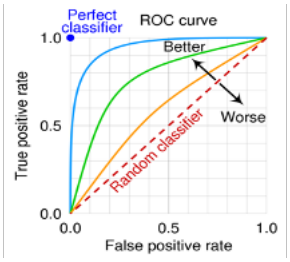
$$TPR = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Dove TP sono i Veri Positivi e FN sono i Falsi Negativi. Rappresenta la proporzione di osservazioni positive reali che sono correttamente identificate dal modello. FPR (1 - Specificità): È la proporzione di osservazioni negative reali che sono erroneamente identificate come positive.

False Positive Rate (FPR) o $1 -$ Specificità:

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Dove FP sono i Falsi Positivi e TN sono i Veri Negativi. Rappresenta la proporzione di osservazioni negative reali che sono erroneamente identificate come positive.



La linea diagonale rappresenta la performance di un classificatore casuale; un modello utile si trova sopra di essa.
Un punto in alto a sinistra del grafico indica un basso FPR e un alto TPR, il che è ideale.
L'area sotto la curva ROC, o AUC, è un indicatore della qualità del modello. Un AUC di 1.0 indica un modello perfetto, mentre un AUC di 0.5 indica un modello non informativo, equivalente a un lancio di moneta.

L'AUC è un indicatore numerico della capacità del modello di distinguere tra classi positive e negative. Valori di AUC più vicini a 1 indicano un migliore equilibrio tra Sensibilità e Specificità, mentre un AUC di 0.5 suggerisce che il modello non ha capacità discriminante.

Valutazione del Classificatore:
AUC elevata: Indica che il modello ha una buona capacità di distinguere tra le classi. AUC bassa: Suggerisce che il modello ha difficoltà a distinguere tra le classi. AUC = 0.5: Il modello non è in grado di distinguere tra le classi, equivalente a un classificatore casuale. AUC = 1.0: Il modello è in grado di classificare perfettamente tutte le osservazioni. Esempio: Un modello con un AUC di 0.8 è generalmente considerato buono, ma potrebbe comunque beneficiare di ottimizzazioni per ridurre ulteriormente gli errori di classificazione. Un modello con un AUC di 0.9 o superiore è considerato eccellente.

Overfitting vs underfitting

Overfitting si verifica quando un modello di apprendimento automatico è troppo complesso rispetto ai dati di addestramento e inizia a memorizzare i dettagli delle singole osservazioni, invece di generalizzare le tendenze generali. Ciò può causare un alto rendimento sui dati di addestramento, ma un rendimento scarso sui dati di test o di validazione.
Underfitting si verifica quando un modello è troppo semplice rispetto ai dati di addestramento e non è in grado di catturare le tendenze nei dati. Ciò può causare un basso rendimento sia sui dati di addestramento che sui dati di test o di validazione.
In generale, il modello ideale dovrebbe essere complesso abbastanza da catturare le tendenze nei dati, ma non così complesso da memorizzare i dettagli delle singole osservazioni.

A cosa servono le association rules

Lorem ipsum

Spiegare apriori

Lorem ipsum

Spiegare fp growth

Lorem ipsum

Spiegare cos'è k-shingles

K-Shingles è una tecnica che permette di trovare documenti simili a una query di documenti in modo efficiente sia in termini di tempo che di spazio, superando la necessità di una corrispondenza esatta delle stringhe. Questo metodo è particolarmente utile per confrontare la similarità tra documenti basandosi sulla sintassi.
Un documento è considerato come una stringa di caratteri, e i suoi k-shingles sono definiti come l'insieme di tutte le possibili sottostringhe di lunghezza k presenti nel documento. La similarità tra due documenti può quindi essere calcolata utilizzando la similarità di Jaccard tra i loro shingles.

Calcolo della Similarità: La similarità di Jaccard tra due insiemi è definita come:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

dove A e B sono gli insiemi di shingles dei due documenti.
Scelta del Valore di K: La scelta del valore di k è cruciale:

- Valori bassi di k aumentano la probabilità di trovare shingles comuni, aumentando la similarità.
- Valori alti di k aumentano il potere discriminante dei shingles.
- Un valore di k compreso tra 5 e 9 è spesso considerato ottimale, a seconda della lunghezza dei documenti.

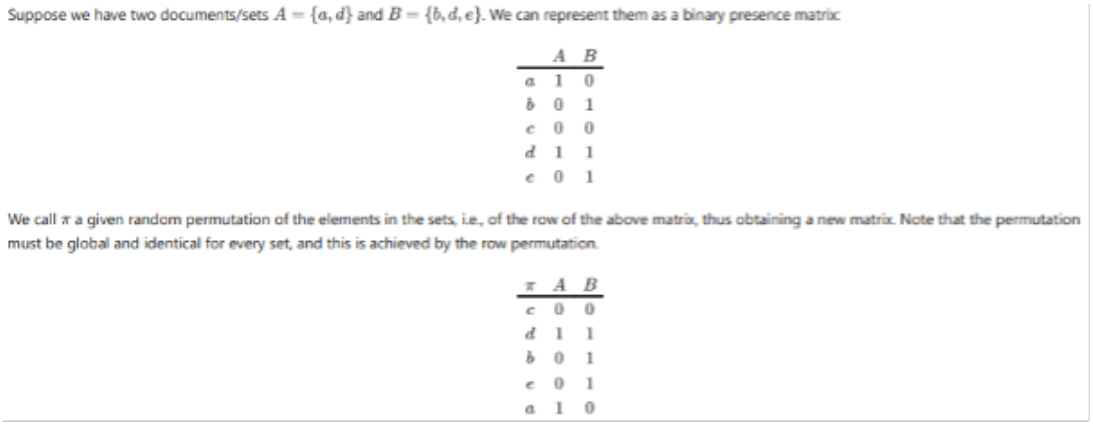
Numero di Shingles: Il numero di shingles per un documento d è calcolato come:

$$\#\text{shingles}(d) = |d| - k + 1$$

dove $|d|$ è la lunghezza del documento in caratteri.
Ottimizzazione: Data la potenziale onerosità computazionale della strategia di k-shingles, sono state sviluppate tecniche di ottimizzazione come il Min-Hashing e il Locality-Sensitive Hashing (LSH) per ridurre ulteriormente il costo computazionale nella comparazione di documenti.

Spiegare il Min-Hashing

La *Min-Hashing* è una tecnica avanzata utilizzata per rappresentare in modo efficiente e compatto i documenti, mantenendo la capacità di calcolare accuratamente la similarità tra di essi. Questa tecnica è particolarmente utile quando si desidera trovare documenti simili a un documento di query in modo efficiente sia in termini di tempo che di spazio, superando i limiti della corrispondenza esatta delle stringhe.
Dati due documenti, A e B, il min-hash di A è definito come il più piccolo elemento dell'insieme A dopo l'applicazione di una permutazione casuale π. Importante notare che questa permutazione è globale e applicata in modo uniforme a tutti i documenti nell'insieme di dati.



Suppose we have two documents/sets $A = \{a, d\}$ and $B = \{b, d, e\}$. We can represent them as a binary presence matrix:

	A	B
a	1	0
b	0	1
c	0	0
d	1	1
e	0	1

We call π a given random permutation of the elements in the sets, i.e., of the row of the above matrix, thus obtaining a new matrix. Note that the permutation must be global and identical for every set, and this is achieved by the row permutation.

π	A	B
c	0	0
d	1	1
b	0	1
e	0	1
a	1	0

Matematicamente, se definiamo $\min(\pi A) = d$ e $\min(\pi B) = d$, allora il min-hash fornisce un mezzo per calcolare la similarità in quanto la probabilità che i due min-hashes siano uguali è equivalente alla similarità di Jaccard dei due documenti, ovvero:

$$P(\min(\pi A) = \min(\pi B)) = J(A, B)$$

Questa tecnica offre vantaggi significativi rispetto al metodo k-shingles. Mentre salvare un singolo shingle è certamente più efficiente dal punto di vista della memoria rispetto a salvare l'intero documento, il metodo k-shingles non fornisce informazioni sulla misura in cui due documenti non simili differiscono. Questa lacuna informativa può essere colmata utilizzando la tecnica di min-hash in combinazione con la *Local Sensitivity Hashing (LSH).*

Spiegare Min-Hash Signatures e LSH

LOOK HERE
Min-Hash Signatures e *LSH* (Locality Sensitive Hashing) sono tecniche avanzate utilizzate per superare le limitazioni della tecnica min-hashing, permettendo di calcolare efficientemente la similarità tra documenti. Quando due documenti non sono simili, la tecnica min-hashing non fornisce informazioni sulla misura in cui differiscono. Per superare questa limitazione, si possono utilizzare più min-hash.
Dando m permutazioni, $\pi_1, \pi_2, \ldots, \pi_m$, per ogni documento A si possono calcolare gli m min-hashes di tali permutazioni ottenendo il vettore

$$\min(\pi A_1), \min(\pi A_2), \ldots, \min(\pi A_m)$$

, cioè la min-hash signature. Questa signature può essere usata per stimare la similarità tra due documenti A, B con la formula:

$$J(A, B) \approx \frac{k}{m}$$

dove k è il numero di min-hashes corrispondenti ottenuti da m permutazioni, e $J(A, B)$ è la distanza di Jaccard tra i documenti A, B.
LSH interviene per risolvere il problema della ricerca e selezione veloce, dividendo gli m min-hashes in b gruppi con $r = \frac{m}{b}$ elementi ognuno. Ogni sottosequenza di r min-hashes viene concatenata e sottoposta nuovamente a hash per ottenere un nuovo hash noto come super-signature (o super-shingle). Queste super-signatures sono usate come chiavi per accedere alla tabella hash.

Spiegare sim-hashing

Sim-Hashing è una tecnica avanzata che permette di rappresentare documenti in uno spazio multidimensionale e di calcolare la similarità del coseno tra di loro. Dati due documenti, rappresentati in uno spazio bidimensionale, essi sono separati da un angolo θ. Utilizzando un vettore random r (o un iperpiano r in spazi di dimensione maggiore), si possono calcolare le probabilità che i due documenti cadano dalla stessa parte di r e che r cada tra i due documenti, permettendo così di stimare la cosine-similarity tra i documenti.
L'algoritmo di Sim-Hashing funziona come segue:

1. Scegliere randomicamente r come un vettore o iperpiano disegnato uniformemente.
2. Dato un documento A, calcolare il prodotto scalare $r \cdot A$ per determinare su quale lato dell'iperpiano A si trova, assegnando 1 se positivo, altrimenti 0.
3. Ripetere i passi da 1 a 2 per m volte per calcolare una firma di m bit.
4. Dati due documenti, calcolare la distanza di Hamming tra le loro signatures per stimare il loro angolo e quindi la cosine similarity.

Spiegare le misure di qualità dei recommender systems

I *Recommender Systems* sono sistemi di filtraggio delle informazioni che mirano a prevedere la preferenza dell'utente e suggerire prodotti, servizi o informazioni pertinenti. La loro qualità è valutata secondo diverse misure:

1. **Efficienza nella costruzione del modello:** Valuta il costo computazionale associato al processamento dei dati e alla costruzione del Recommender System, includendo l'analisi necessaria per generare i dati utilizzati.
2. **Efficienza nella generazione dei suggerimenti:** Misura il costo computazionale delle raccomandazioni a run-time, ovvero il tempo e le risorse necessarie per generare suggerimenti in tempo reale.
3. **Serendipità delle raccomandazioni:** Le raccomandazioni devono essere nuove, non banali, e inaspettate, aiutando gli utenti a scoprire nuovi oggetti e ad esplorare nuovi interessi.
4. **Adattamento al problema della "partenza a freddo":** Valuta come il modello si comporta quando incontra nuovi utenti o nuovi items, per i quali ha poche o nessuna informazione precedente.

Queste misure permettono di valutare l'efficacia e l'efficienza dei Recommender Systems, garantendo che siano in grado di fornire suggerimenti pertinenti e utili agli utenti, anche in presenza di informazioni limitate.

Spiegare le varie tecniche usate in recommender system

Si identificano 3 tipi di recommender systems che utilizzano, ognuno, delle strategie diverse: Content Based, User-Based e Item-Based.

[0] Content Based: Il filtraggio basato sul contenuto si focalizza sulle caratteristiche degli items. Se un utente ha mostrato interesse per un particolare tipo di item, il sistema raccomanderà items con caratteristiche simili.
Formula di Similarità:
La similarità tra due items x e y è data dal coseno dell'angolo tra i loro vettori di caratteristiche:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

Dove:

- x e y sono rappresentazioni vettoriali degli items basate sulla frequenza dei termini.
- $tf(t, x)$ è la frequenza del termine t nell'item x.
- $idf(t)$ è l'inverso della frequenza di t nell'insieme di items I, che sminuisce il peso delle parole comuni e aumenta il peso delle parole rare.

Pro e Contro:

- **Pro**: Facile da implementare e non richiede dati su altri utenti.
- **Contro**: Può essere limitato nella capacità di scoprire nuovi interessi dell'utente. Soffre di un problema di "Cold-Start" parziale perché ha bisogno di un certo grado di interazione dell'utente con almeno un item.

[1] User-Based: Questo approccio cerca di trovare utenti simili all'utente target e raccomanda items che questi utenti simili hanno apprezzato. **Formula di Similarità:**
La similarità tra due utenti u e v è spesso calcolata usando il coefficiente di correlazione di Pearson:

$$\rho(u, v) = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2 \sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}}$$

Dove:

- $r_{u,i}$ è il rating dato dall'utente u all'item i.
- $\bar{r}_u$ è la media dei ratings dell'utente u.

Pro e Contro:

- **Pro**: Può fornire raccomandazioni molto personalizzate.
- **Contro**: Costoso computazionalmente, può soffrire di scarsità di dati e ha un forte problema di "Cold-Start" per nuovi utenti.

[2] Item-Based:Invece di misurare la similarità tra utenti, l'item-based CF misura la similarità tra items basandosi sulle valutazioni degli utenti. **Formula di Similarità:**
La similarità tra due items i e j può essere calcolata come:

$$\text{sim}(i, j) = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_u)(r_{u,j} - \bar{r}_u)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_u)^2 \sum_{u \in U}(r_{u,j} - \bar{r}_u)^2}}$$

Dove:

- $r_{u,i}$ è il rating dato dall'utente u all'item i.
- $\bar{r}_u$ è la media dei ratings dell'utente u.

Pro e Contro:

- **Pro**: Generalmente più stabile e meno costoso computazionalmente rispetto al filtraggio basato sull'utente.
- **Contro**: Può essere meno personalizzato e, come il filtraggio basato sull'utente, può soffrire di problemi di "Cold-Start".

Quali sono le principali misure di similarità?

Esistono diverse misure di similarità tra le quali citiamo:

1. Minkowski Distance: Dati due oggetti n-dimensionali x e y con $x = [x_1, x_2, \ldots, x_n]$ e $y = [y_1, y_2, \ldots, y_n]$:

$$d(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^q \right)^{\frac{1}{q}}$$

2. Distanza Euclidea: È un caso particolare della distanza di Minkowski con $q = 2$:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

3. Distanza di Manhattan: È un caso particolare della distanza di Minkowski con $q = 1$:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$$

4. Distanza di Chebyshev: Con q tendente all'infinito:

$$d(X,Y) = \max_i |x_i - y_i|$$

5. Jaccard Similarity: Utilizzata nell'analisi dei documenti tra due set A e B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

6. Cosine Similarity: Utilizzata spesso per confrontare vettori:

$$\cos(A,B) = \frac{A \cdot B}{\|A\|\|B\|}$$

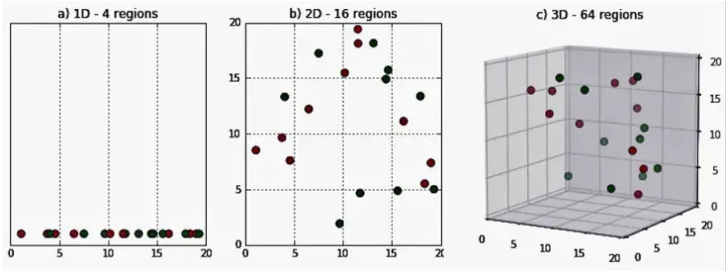
7. Pearson Correlation: Misura la correlazione lineare tra due variabili e varia tra -1 e 1:

$$\text{corr}(A,B) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{A})^2 \sum_{i=1}^{n}(b_i - \bar{B})^2}}$$

   Dove $\bar{A}$ e $\bar{B}$ sono le medie delle variabili A e B.

8. Trasformata di Fourier: Utilizzata per analizzare le frequenze in serie temporali. La trasformata di Fourier trasforma una funzione del tempo (segnale) in una funzione delle frequenze.

Spiegare il problema della dimensionalità

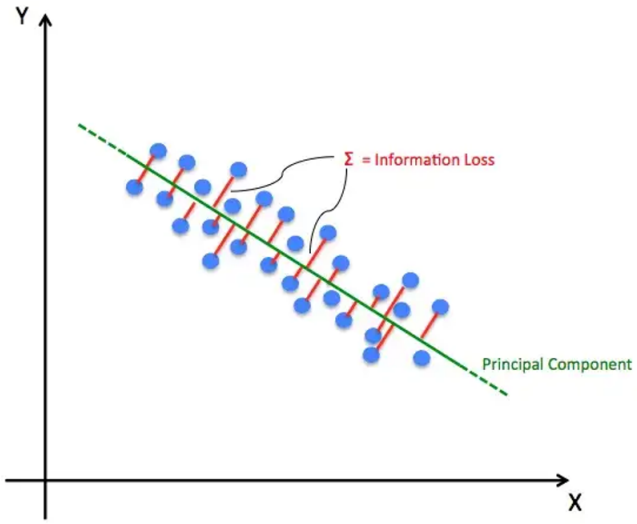


Il problema della dimensionalità è un concetto chiave nell'ambito del machine learning e della statistica. Questo problema emerge principalmente quando si tratta di grandi set di dati con molte variabili o dimensioni.
La maggior parte degli algoritmi di machine learning valuta misure di similarità o distanza tra i dati. Tuttavia, con l'aumento delle dimensioni, queste misure tendono a perdere il loro significato. Questo fenomeno è spesso chiamato *"maledizione della dimensionalità"*.
Aumentando la dimensionalità, anche se solo leggermente, l'area (o volume, in dimensioni superiori) dello spazio dei dati aumenta esponenzialmente. Di conseguenza, i dati diventano molto sparsi e le distanze tra le osservazioni tendono a convergere, rendendo difficile distinguere tra loro. Questo deteriora la performance di molti algoritmi, specialmente quelli che si basano su concetti di distanza, come k-NN (k-Nearest Neighbors).
Per esempio, con un'elevata dimensionalità (ad es. > 10), le misure di distanza come la distanza euclidea diventano meno efficaci. Le distanze tra i punti tendono a diventare quasi uguali, rendendo difficile per l'algoritmo distinguere tra diverse osservazioni.
La soluzione a questo problema passa attraverso la riduzione della dimensionalità. Ci sono due principali approcci:

1. **Feature Selection**: Questo metodo consiste nel selezionare solo un sottoinsieme delle feature originali, eliminando quelle meno rilevanti o ridondanti.
2. **Feature Extraction**: Tecniche come l'Analisi delle Componenti Principali (PCA) sono usate per trasformare le originali dimensioni ad alta dimensionalità in un nuovo set di dimensioni a bassa dimensionalità che catturano la maggior parte della varianza dei dati. In pratica, PCA identifica le "direzioni" in cui i dati variano di più e li proietta su queste direzioni.

In sintesi, la riduzione della dimensionalità non solo migliora le performance degli algoritmi di machine learning, ma può anche fornire intuizioni sui dati, rendendo le variabili più interpretabili.

Come è possibile risolvere il problema della dimensionalità tecniche



Il problema della dimensionalità rappresenta una sfida significativa nel campo del machine learning, specialmente quando si lavora con set di dati ad alta dimensionalità. Mentre l'alta dimensionalità può fornire una ricchezza di informazioni, può anche rendere difficile l'analisi dei dati e l'addestramento degli algoritmi.
Una delle tecniche più efficaci e comunemente utilizzate per affrontare il problema della dimensionalità è la PCA (Analisi delle Componenti Principali).
La PCA è una tecnica di riduzione della dimensionalità che si basa sull'estrazione delle direzioni (o componenti) che massimizzano la varianza nei dati. In altre parole, PCA identifica le "direzioni" nel set di dati dove esiste la maggior varianza o dispersione dei dati e utilizza queste direzioni come nuovi assi, chiamati componenti principali.
Il primo componente principale rappresenta la direzione di massima varianza, il secondo componente rappresenta la direzione di seconda massima varianza, ortogonale al primo, e così via.
L'idea di base della PCA è rappresentare il set di dati originale, che potrebbe avere molte dimensioni, in un nuovo spazio a dimensione ridotta, mantenendo la maggior parte della varianza originale dei dati.
Tuttavia, è importante notare che, riducendo la dimensionalità, alcune informazioni vengono perse. Il compromesso fondamentale della PCA è tra riduzione della dimensionalità e conservazione dell'informazione.

Discutere come stimare la distanza di Jaccard tra due insieme tramite MinHashing

Lorem ipsum

Discutere brevemente cosa si intende per dimensionality curse

Lorem ipsum

Discuterela rappresentazionevettorialedeidocumentibasatato tfidf

Cosa significa fare clustering

Fare clustering significa trovare cluster all'interno del dataset, ossia trovare delle "somiglianze" tra i dati in base a determinate caratteristiche e raggruppare i dati similari in clusters.

Si definisce cluster una collezione di oggetti, ossia un subset del dataset originario con le seguenti caratteristiche:

- gli oggetti nello stesso cluster sono simili (o correlati) tra loro;
- gli oggetti in cluster differenti sono dissimili (o incorrelati) tra loro.

Il clustering è un processo categorizzato come unsupervised learning, dove non si hanno classi predefinite e l'algoritmo apprende tramite osservazione.
E' tipicamente usato per ottenere informazioni sui dati o come pre-processo dei dati per altri algoritmi, anche di supervised learning, allo scopo di migliorare le performance di questi ultimi.

Descrivere i vari tipi di clustering

Il clustering, o raggruppamento, è una tecnica di apprendimento non supervisionato che mira a dividere un insieme di dati in gruppi, o cluster, di oggetti simili. Ci sono vari tipi di clustering che possono essere categorizzati in base ai seguenti aspetti:

1. Similarity Measures
   - **Distanza**: Il clustering si basa sulla distanza fisica tra gli oggetti. Es. k-means utilizza la distanza euclidea.
   - **Connettività**: Basato su links o connessioni tra gli oggetti. Es. DBSCAN considera la densità e la connettività.
2. Clustering Space
   - **Full space**: Riguarda tutte le features.
   - **Sub space**: Riguarda un subset delle features. Questo è particolarmente utile quando alcune dimensioni sono irrilevanti o rumorose.
3. Partitioning Criteria
   - **Single level**: Una normale divisione tra clusters.
   - **Hierarchical**: Si ha una gerarchia di clusters. Gli algoritmi come Agglomerative Clustering funzionano in questo modo.
4. Separation of Clusters
   - **Esclusivo**: Un oggetto appartiene solo ed esclusivamente ad un cluster. Es. k-means.
   - **Non esclusivo**: Un oggetto può appartenere a più clusters con un certo grado di appartenenza. Es. Fuzzy C-means.

Diversi algoritmi di clustering sono stati sviluppati basandosi su questi aspetti.

Approcci al clustering

Il clustering è una tecnica di apprendimento non supervisionato utilizzata per raggruppare dati simili in insiemi o "cluster". Esistono vari approcci al clustering, e la scelta di quale utilizzare dipende dalla natura dei dati e dal problema specifico. Di seguito sono descritti i principali approcci:

1. Partitioning Approach: In questo approccio, si tenta di suddividere direttamente l'insieme di dati in un numero predeterminato di cluster. Gli algoritmi basati su questo approccio iniziano con una divisione iniziale dei dati e quindi iterativamente riorganizzano i cluster per ottimizzare un certo criterio, come la somma dei quadrati degli errori (SSE). **Metodi principali**:
   - **k-means**: Assegna ogni punto al cluster il cui centroide è il più vicino.
   - **k-medoids**: Simile al k-means, ma utilizza punti effettivi del dataset come centroidi.
2. Hierarchical Approach: Gli algoritmi gerarchici costruiscono una decomposizione gerarchica dei dati basata su misure di distanza. Questi algoritmi possono essere agglomerativi (partendo da singoli punti e combinandoli) o divisivi (partendo da un singolo cluster e suddividendolo). **Metodi principali**:
   - **HAC (Hierarchical Agglomerative Clustering)**: Inizia con ogni punto come un cluster separato e li fonde iterativamente.
   - **HDC (Hierarchical Divisive Clustering)**: Inizia con tutti i punti in un cluster e li divide iterativamente.
3. Density-Based Approach: Questi algoritmi cercano regioni dello spazio dei dati che hanno una densità di punti superiore rispetto alle regioni circostanti. I punti in regioni a bassa densità sono solitamente considerati come rumore o punti limite. **Metodi principali**:
   - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Raggruppa insieme punti che sono vicini in termini di densità e può identificare cluster di forma arbitraria.
4. Altri Approcci: Esistono anche altri approcci, come:
   - **Grid-based**: Divide lo spazio dei dati in un numero finito di celle formando una griglia e quindi effettua il clustering sulle celle invece che sui dati.
   - **Model-based**: Assume che i dati siano generati da una miscela di diversi modelli e cerca di identificare questi modelli.
   - **Frequent-pattern based**: Basato su tecniche di mining di pattern frequenti.
   - **User-guided**: L'utente fornisce informazioni o vincoli per guidare il processo di clustering.
   - **Link-based**: Utilizzato per dati che possono essere rappresentati come grafi, come reti sociali o pagine web.

Come funziona k-means

Il K-means è uno degli algoritmi di clustering più popolari e si basa sulla seguente procedura:

1. Scegliere casualmente k punti come centroidi iniziali.
2. Assegnare ogni punto al centroide più vicino usando la distanza euclidea.
3. Calcolare il nuovo centroide di ogni cluster come la media dei punti che appartengono a quel cluster.
4. Ripetere i passi 2 e 3 fino a convergenza, ovvero quando i centroidi non cambiano più.

Nota: Questo metodo è particolarmente efficace per dataset che presentano cluster sferici.

Come funziona k-means++

K-means++ è una variante del K-means che migliora la scelta dei centroidi iniziali. Al posto di sceglierli completamente a caso, K-means++ seleziona i centroidi iniziali in modo che siano distanti tra loro. Il processo è il seguente:

1. Scegliere casualmente un punto come primo centroide.
2. Per i successivi $k-1$ centroidi, scegliere un punto con una probabilità proporzionale alla sua distanza quadrata dal punto più vicino già scelto come centroide.
3. Procedere con l'algoritmo K-means standard.

Come funziona k-medoid (PAM)

K-medoids, noto anche come Partitioning Around Medoids (PAM), è simile al K-means ma utilizza punti effettivi del dataset come centroidi (chiamati medoidi). La procedura è:

1. Scegliere casualmente k punti del dataset come medoidi iniziali.
2. Assegnare ogni punto al medoide più vicino.
3. Per ogni medoide e per ogni punto non medoide, scambiare il medoide con il punto e calcolare il costo totale (somma delle distanze tra i punti e il loro medoide). Se lo scambio riduce il costo, accettarlo.
4. Ripetere i passi 2 e 3 fino a convergenza.

Spiegare come funziona HAC e le varie misure (vedi anche complessità)

Hierarchical Agglomerative Clustering (HAC) Il Hierarchical Agglomerative Clustering (HAC) è una metodologia di clustering che costruisce una gerarchia di cluster in un approccio bottom-up. Di seguito è riportato un'esposizione dettagliata del funzionamento di HAC, delle misure di linkage utilizzate, e della complessità computazionale associata. **Funzionamento dell'Algoritmo HAC** L'algoritmo HAC inizia trattando ogni punto dati come un cluster singolo e poi, in ogni iterazione, fonde i due cluster più vicini fino a quando non rimane che un solo cluster o fino a quando non si raggiunge un certo criterio di terminazione. La procedura può essere descritta come segue:

1. **Inizializzazione**: Ogni punto dati è inizialmente considerato come un cluster singolo, quindi si hanno $|D|$ cluster, dove D è l'insieme dei dati.
2. **Iterazione**:
   (a) Trovare la coppia di cluster C_i e C_j che sono i più vicini secondo una certa misura di similarità o distanza.
   (b) Unire C_i e C_j in un nuovo cluster.
   (c) Aggiornare la matrice di similarità/distanza per riflettere la fusione.
3. **Terminazione**: Continuare l'iterazione finché il numero di cluster $|C|$ non è ridotto a 1, o fino a quando non si raggiunge un altro criterio di terminazione.

A livello di strutture dati, l'algoritmo utilizza una matrice di similarità/distanza S di dimensione $|D| \times |D|$. Quando due cluster sono fusi, una delle colonne e una delle righe corrispondenti nella matrice S vengono aggiornate per riflettere la fusione, mentre l'altra colonna e l'altra riga vengono invalidate.

Misure di Linkage Le misure di linkage determinano la distanza tra due cluster, e ci sono diverse misure comuni utilizzate in HAC:

- **Single Linkage**: La distanza tra due cluster è data dalla distanza minima tra qualsiasi punto in un cluster e qualsiasi punto nell'altro cluster:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

- **Complete Linkage**: La distanza tra due cluster è data dalla distanza massima tra qualsiasi punto in un cluster e qualsiasi punto nell'altro cluster:
$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

- **Average Linkage**: La distanza tra due cluster è data dalla distanza media tra tutti i punti in un cluster e tutti i punti nell'altro cluster:
$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

- **Centroid Linkage**: La distanza tra due cluster è data dalla distanza tra i centroidi dei cluster:
$$d(C_i, C_j) = d(\text{centroid}(C_i), \text{centroid}(C_j))$$

- **WARD Linkage**: La distanza tra due cluster è data dall'incremento della somma dei quadrati degli errori (SSE) risultante dalla fusione dei cluster:
$$d(C_i, C_j) = \text{SSE}(C_i \cup C_j) - (\text{SSE}(C_i) + \text{SSE}(C_j))$$

Complessità Computazionale La complessità computazionale dell'algoritmo HAC dipende dalla struttura dati utilizzata per memorizzare e aggiornare la matrice di similarità/distanza, e dalla misura di linkage utilizzata.

- Nel caso generale, la complessità è $O(n^3)$ a causa dell'aggiornamento della matrice di similarità/distanza in ogni iterazione.
- Utilizzando una struttura dati min-heap, la complessità può essere ridotta a $O(n^2 \log n)$.
- Nel caso del single-linkage, con un'implementazione efficiente, la complessità può essere ulteriormente ridotta a $O(n^2)$.

HAC è un algoritmo flessibile in termini di forma del cluster, ma è sensibile al rumore e agli outliers che possono influenzare negativamente la qualità del clustering.

A cosa serve il dendrogramma

Si tratta di un grafico ad albero dove sull'asse delle ordinate è riportata la "distanza" tra i cluster e sull'asse orizzontale vengono riportati i vari dati in ingresso.

In questo diagramma, inoltre, le righe verticali corrispondono ad un cluster, quelle orizzontali ad operazioni di unione (se si usa la versione agglomerativa dell'algoritmo, che si legge dal basso verso l'alto) o di divisione (se si usa la versione divisiva dell'algoritmo, che si legge dall'alto verso il basso).
Quindi, il dendrogramma rappresenta una sorta di "memoria" dei cluster e visualizza come questi cambiano in funzione della distanza massima che vogliamo applicare ai nostri dati, distanza che funzionerà come una soglia.

Spiegare clustering divisivo

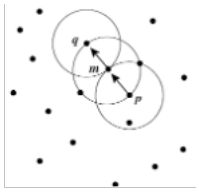
HDC - Hierarchical Divisive Clustering è un approccio di clustering gerarchico divisivo, che non richiede di definire a priori il numero di clusters, in cui inizialmente tutti i punti nel dataset appartengono a un singolo cluster e la divisione in clusters ulteriori viene eseguita in modo ricorsivo man mano che si scende nella gerarchia, attuando quindi una strategia top-down.
Gli step che segue l'algoritmo sono i seguenti:

- Inizialmente, tutti i punti nel dataset appartengono a un singolo ed unico cluster;
- Partiziona il cluster in 2 cluster, il meno simile tra di loro;
- Ad ogni iterazione, il cluster più eterogeneo viene diviso in due cluster;
- Procede con le iterazioni fino a che tutti gli oggetti sono nel loro cluster o comunque fino a terminazione anticipata (es.: valore soglia di numero di clusters formati).

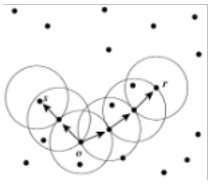
Spiegare DBScan

DBScan è un algoritmo di clustering basato sulla densità che opera identificando le regioni di alta densità di punti separati da regioni di bassa densità. L'algoritmo utilizza due parametri: un raggio ϵ e un numero minimo di punti MinPts. Di seguito sono descritti i concetti chiave e il funzionamento dell'algoritmo DBScan.
Concetti Chiave

1. **Core Point (Punto Centrale)**: Un punto p è un core point se ci sono almeno MinPts punti entro una distanza ϵ da p, ossia se $|N_\epsilon(p)| \geq \text{MinPts}$ dove $N_\epsilon(p) = \{q \in D \mid \text{dist}(q, p) \leq \epsilon\}$.
2. **Density Reachability (Raggiungibilità della Densità)**: Un punto p è detto "density reachable" da un punto q se p è un core point e q appartiene a $N_\epsilon(p)$, e se esistono punti $p_1, p_2, \ldots, p_n$ con $p_1 = p$ e $p_n = q$ tale che p_{i+1} sono core point e appartengono al vicinato di p.



3. **Density Connectivity (Connettività della Densità)**: Un punto s è density-connected al punto r se esiste un punto o tale che s è density reachable da o e r è density reachable da o.



Algoritmo Siano $X = x_1, x_2, \ldots, x_n$ l'insieme dei punti. DBScan richiede la definizione a priori di ϵ e MinPts. L'algoritmo procede come segue:

1. Seleziona un punto di partenza arbitrario che non è stato visitato.
2. Estrae l'intorno di questo punto usando ϵ (tutti i punti che si trovano all'interno della distanza ϵ sono considerati vicini).
3. Se ci sono abbastanza vicini attorno a questo punto, il processo di raggruppamento inizia e il punto viene contrassegnato come visitato, altrimenti il punto viene etichettato come rumore.
4. Se un punto è parte del cluster, anche il suo intorno ϵ è parte del cluster, e gli steps dal punto (2) vengono ripetuti per tutti i punti dell'intorno ϵ. Questo viene ripetuto finché non vengono determinati tutti i punti nel cluster.
5. Ripeti dal punto 1 fino a quando tutti i punti sono stati visitati.

Vantaggi e Svantaggi

- **Vantaggi**:
  - Non richiede di specificare a priori il numero di cluster.
  - È in grado di identificare il rumore durante il processo di clustering.
  - È in grado di trovare cluster di dimensioni e forme arbitrarie.
- **Svantaggi**:
  - Fallisce nel caso di cluster a densità variabile.

Spiegare la valutazione del clustering intrinseca ed estrinseca / Cos'è il silhouette coefficient

Valutazione del Clustering La valutazione del clustering può essere effettuata in due modi principali: intrinseca ed estrinseca. **Valutazione Intrinseca** La valutazione intrinseca è utilizzata quando non si dispone di una verità fondamentale. Questo tipo di valutazione si basa sulla misurazione della similarity intra-class e della dissimilarity inter-class.

- **Intra-class Similarity**: Misura quanto vicino è un oggetto agli altri oggetti all'interno dello stesso cluster. Una similarity intra-class elevata è desiderabile.

$$\text{Intra-class Similarity} = \frac{1}{|C_h|} \sum_{o_i \in C_h} \text{sim}(o_i, o_j)$$

  dove C_h è il cluster, o_i e o_j sono oggetti nel cluster, e $\text{sim}(o_i, o_j)$ è una funzione di similarity tra o_i e o_j.

- **Inter-class Dissimilarity**: Misura la dissimilarity tra un oggetto e gli oggetti in altri cluster. Una dissimilarity inter-class elevata è desiderabile.

$$\text{Inter-class Dissimilarity} = \frac{1}{|C| - 1} \sum_{C_k \neq C_h} \text{dissim}(o_i, o_j)$$

  dove C è l'insieme di tutti i cluster, C_k è un cluster diverso da C_h, e $\text{dissim}(o_i, o_j)$ è una funzione di dissimilarity tra o_i e o_j.

Il **Silhouette Coefficient** è una metrica derivata da queste misure, definita come segue:

$$s = \frac{b - a}{\max(a, b)}$$

dove a è la intra-class similarity e b è la inter-class dissimilarity. Il Silhouette Coefficient assume valori tra -1 e 1, ed è un indice della qualità del clustering.

Valutazione Estrinseca La valutazione estrinseca è utilizzata quando si dispone di una verità fondamentale, come un piccolo dataset etichettato manualmente. Una tabella di contingenza è costruita per confrontare il clustering ottenuto con la verità fondamentale.

- **Rand Statistic**:

$$R = \frac{TP + TN}{TP + TN + FP + FN}$$

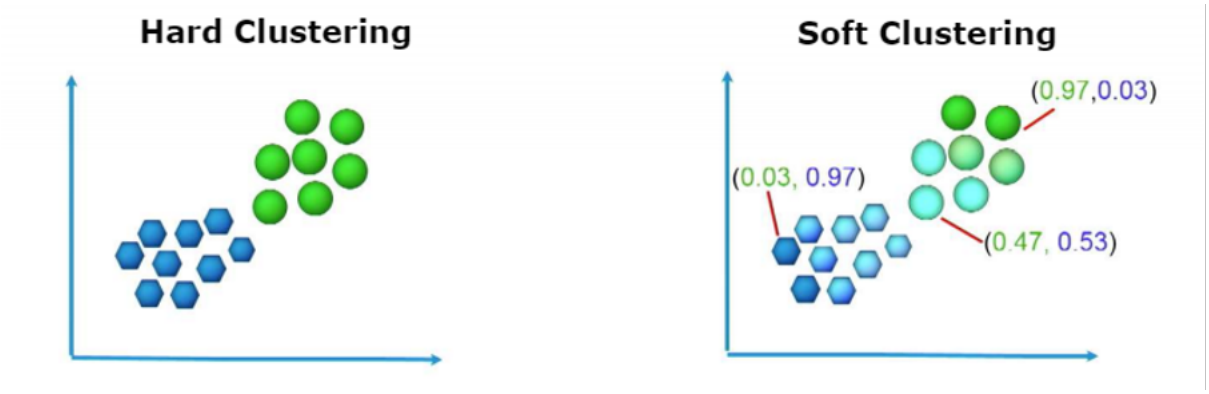
  dove TP, TN, FP, e FN sono i valori di True Positive, True Negative, False Positive e False Negative rispettivamente.

- **Jaccard Coefficient**:

$$J = \frac{TP}{TP + FP + FN}$$

  Il Jaccard Coefficient ignora i valori di True Negative ed è spesso considerato più accurato rispetto alla Rand Statistic.

Cos'è l'hard clustering e il soft clustering?



L'hard clustering e il soft clustering sono due diverse tecniche utilizzate per l'assegnazione degli oggetti ai cluster in un algoritmo di clustering.
In hard clustering, ogni oggetto viene assegnato a un solo cluster. In altre parole, un oggetto appartiene completamente ad un solo cluster e non appartiene a nessun altro. Gli esempi di algoritmi di clustering hard sono K-Means, Hierarchical clustering, DBSCAN.
In soft clustering, ogni oggetto viene assegnato a più di un cluster, ma con diverse appartenenze. In altre parole, un oggetto può appartenere parzialmente ad uno o più cluster. Gli esempi di algoritmi di clustering soft sono Fuzzy C-Means ed altri.
In generale, l'hard clustering è più semplice da implementare e più facile da interpretare rispetto al soft clustering, ma può essere meno adatto per i dati con classi sovrapposte o per i dati che non possono essere facilmente assegnati a un singolo cluster.
Il soft clustering, d'altra parte, è più flessibile e adatto per i dati con classi sovrapposte, ma può essere più complesso da implementare e da interpretare.

Spiegare fuzzy C-means.

Fuzzy C-means (FCM) Fuzzy C-means è un metodo di soft clustering che permette ad un oggetto di appartenere parzialmente ad uno o più cluster, con un determinato valore w_{ij} di appartenenza, detto peso. Tale peso w_{ij} rappresenta la probabilità che l'oggetto x_i appartenga al cluster C_j, e assume valori nell'intervallo $[0, 1]$. La somma dei pesi di ogni oggetto è uguale a 1, e ogni cluster C_j contiene almeno un punto con peso non nullo, senza contenere tutti i punti con peso pari a 1.
Algoritmo L'algoritmo Fuzzy C-means prevede i seguenti passi:

1. **Inizializzazione**: Assegnazione iniziale dei pesi agli oggetti, eventualmente in modo casuale, rispettando il vincolo che la somma dei pesi di ogni oggetto sia 1.
2. **Iterazione**: Fino a quando i centroidi non cambiano più, o la somma dei quadrati degli errori (SSE) è minore di una soglia t:
   (a) Calcolo del centroide per ogni cluster usando la fuzzy pseudo-function, in modo da minimizzare il fuzzy SSE:

$$c_j = \frac{\sum_{i=1}^{n} w_{ij}^p x_i}{\sum_{i=1}^{n} w_{ij}^p}$$

   dove p è un parametro che controlla il grado di "fuzziness" della partizione. Con $p = 1$ l'algoritmo si comporta come k-means, con $p = 2$ la formula si semplifica, e con p alto la partizione diventa più fuzzy, come se non ci fosse alcun cluster.

(b) Calcolo della partizione fuzzy, ossia assegnazione dell'oggetto all'insieme con il peso più alto e aggiornamento dell'i-esimo peso per il j-esimo cluster:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d(x_i, c_j)}{d(x_i, c_k)} \right)^{\frac{2}{m-1}}}$$

dove d è una funzione di distanza, c è il numero di cluster, e m è un parametro che controlla il grado di "fuzziness" della partizione.

(c) Calcolo del fuzzy SSE:

$$\text{Fuzzy SSE} = \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{p} d(x_i, c_j)^2$$

Considerazioni L'algoritmo Fuzzy C-means fornisce informazioni sul grado di appartenenza degli oggetti ai cluster, offrendo una rappresentazione più ricca rispetto al k-means tradizionale. Tuttavia, è computazionalmente più intensivo, richiedendo il calcolo dei pesi e dei centroidi fuzzy in ogni iterazione.

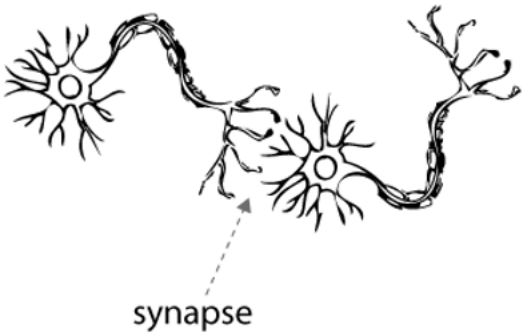
Spiegare SOM(self organizing map)

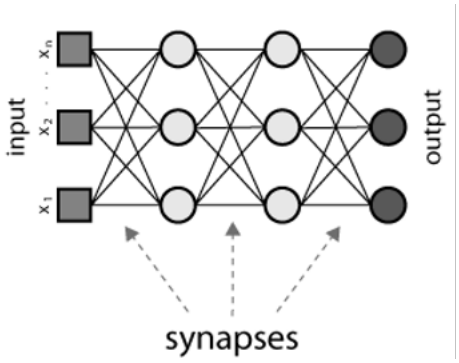
SOM (Self Organizing Map) è un algoritmo di clustering basato sui centroidi che rafforza la relazione di neighborhood dei centroidi risultanti, trovando un insieme di centroidi "reference vectors" allo scopo di assegnare ogni oggetto al centroide che fornisce la miglior approssimazione per quell'oggetto.
Una volta inizializzati i centroidi, l'algoritmo ripete i seguenti passi:

- seleziona il prossimo oggetto
- determina il centroide più vicino all'oggetto
- aggiorna il centroide e i suoi vicini, in un specifico neighborhood

Tali passi vengono ripetuti fino al raggiungimento di un valore soglia o fino a quando i centroidi non cambiano più.
Dopodiché, ogni oggetto viene assegnato al suo centroide più vicino e vengono ritornati in output i centroidi e i clusters.
Tale algoritmo risulta essere una generalizzazione del K-means e fornisce data visualization.

Spiegare cos'è un ANN e come funziona





Una rete neurale artificiale (ANN) è un modello matematico ispirato al funzionamento del cervello umano. Consiste in una serie di nodi, chiamati neuroni, che sono collegati tra loro mediante delle connessioni chiamate pesi o sinapsi.
Ogni neurone riceve un input dai neuroni connessi a esso, effettua una semplice operazione matematica su di esso e quindi passa il risultato all'uscita del neurone. Le connessioni tra i neuroni hanno pesi associati ad essi che possono essere regolati durante il processo di apprendimento.
Ogni strato di neuroni può inoltre essere caratterizzato da una funzione di attivazione che viene applicata a ogni neurone prima di passare il risultato all'uscita.

Spiegare le varie funzioni di attivazione

Esistono diverse funzioni di attivazione utilizzate nei diversi layer di un'ANN:

- La funzione di attivazione di tipo sigmoidale (chiamata anche logistica) che produce un output compreso tra 0 e 1, ed è spesso utilizzata per la classificazione binaria.
- La funzione di attivazione ReLU (Rectified Linear Unit) che produce un output uguale all'input se è maggiore di zero, altrimenti produce zero. È spesso utilizzata nei layer intermedi della rete.
- La funzione di attivazione Tanh (Tangente iperbolica), simile alla funzione sigmoidale, ma produce un output compreso tra -1 e 1.
- La funzione di attivazione softmax, utilizzata per la classificazione multipla, essa trasforma gli output dei neuroni in una distribuzione di probabilità, in modo che la somma degli output sia uguale a 1.

Cos'è la loss function per un ANN

La funzione di perdita (loss function) in una rete neurale artificiale (ANN) è una funzione matematica utilizzata per misurare la differenza tra l'output desiderato della rete e l'output effettivo della rete per un determinato set di input.
La funzione di perdita è utilizzata durante il processo di addestramento della rete, perché aiuta a misurare quanto la rete è "lontana" dal produrre gli output desiderati per un determinato set di input. In base al valore della funzione di perdita, i pesi delle connessioni tra i neuroni della rete vengono regolati, in modo che la rete produca output più precisi per gli input.
Esistono diverse funzioni di perdita, alcune delle quali sono utilizzate per problemi di classificazione, mentre altre sono utilizzate per problemi di regressione. Ad esempio, la funzione di perdita log-loss è spesso utilizzata per i problemi di classificazione binaria, mentre la funzione di perdita mean squared error (MSE) è spesso utilizzata per i problemi di regressione.
In generale, la scelta della funzione di perdita dipende dal tipo di problema che si vuole risolvere e dalle caratteristiche dei dati di input.

Cosa apprende un ANN

Una rete neurale apprende pattern di correlazione dai dati in input per poter fornire una predizione accurata in output.
In pratica ogni layer rappresenta un set di features, le quali dipendono da quelle del layer precedente in base ai pesi delle connessioni fra i due.
Il processo di apprendimento modifica proceduralmente i pesi fino a convergenza, in modo che i neuroni che si attivano (quelli che forniscono un valore più significativo) rappresentino in qualche modo i dati in input tramite le nuove feature calcolate dai vari layer e in modo che l'attivazione del layer finale sia il più simile all'output desiderato.
È da tenere a mente che "modifica proceduralmente i pesi fino a convergenza" non vuol dire che continua fino ad un totale overfit dei dati, bensì fino a che successivi allenamenti non comportano un miglioramento significativo delle prestazioni del modello.

Cos'è una rete convoluzionale e come funziona

Una rete neurale convolutionale (CNN) è un tipo specifico di rete neurale artificiale progettata per lavorare con dati che hanno una struttura spaziale, come immagini, video e audio. La caratteristica principale di una CNN è l'utilizzo di strati di convoluzione, che sono in grado di estrarre caratteristiche rilevanti dai dati di input.
I layer di convoluzione hanno un set di filtri (kernel) che scorrono sui dati di input, calcolando il prodotto scalare tra i valori dei pixel dell'immagine e i pesi del filtro.
Proceduralmente, per ogni posizione nella quale si può applicare il filtro (ovvero le posizioni nelle quali è completamente contenuto nel tensore in input) si calcola il prodotto scalare fra il filtro e i pixel sotto ad esso. Il risultato di questo calcolo viene salvato in un tensore di output, che ha dimensioni minori rispetto a quelle del tensore in input. Questo processo viene ripetuto per ogni filtro, generando un tensore di output per ogni filtro. I risultati di tutti i filtri vengono concatenati in un unico tensore di output, che viene utilizzato come input per il layer successivo.

A cosa servono i filtri

I filtri in una rete neurale convolutiva (CNN) sono utilizzati per estrarre caratteristiche specifiche dai dati in input. Essi vengono fatti scorrere sull'immagine (o qualsiasi altro tipo di dato strutturato) in modo da catturare gli eventuali pattern.
In pratica si tratta di matrici/tensori composti da alcuni pesi. Per ogni posizione possibile viene eseguito il prodotto scalare fra il filtro stesso e la porzione di dati sottostante.
Il risultato sarà un singolo valore, chiamato anche attivazione del filtro o del kernel.
L'insieme di tutte le attivazioni del filtro per l'input processato compone una nuova matrice/tensore detta anche feature map. Se per un singolo layer esistono più filtri, il risultato sarà un vettore di matrici/tensori i quali rappresenteranno i diversi pattern riconosciuti dai vari filtri.

Per quali task può essere usata un ANN e invece una convolutional network?

Una classica rete neurale (ANN) può essere utilizzata per scopi di classificazione o regressione, mentre una rete neurale convolutiva (CNN) è progettata per gestire dati più strutturati, come ad esempio immagini, video o audio. Tramite il processo di convoluzione riesce ad estrarre informazioni dall'input in modo più efficace rispetto ad una rete neurale classica, il che la rende ottima per scopi di riconoscimento di immagini o pattern in immagini/video/audio.

Come risolvere l'overfitting in un ANN

In una rete neurale, l'overfitting si verifica nel caso in cui la rete ha imparato eccessivamente bene i dati in input al punto da esserne troppo aderente. Per risolvere il problema esistono varie tecniche:

- Ridurre la complessità del modello, ad esempio rimuovendo layer o riducendo il numero dei neuroni per un determinato layer.
- Utilizzare un set di dati di validazione, monitorando il modello durante l'addestramento, non appena si rileva overfitting si termina l'allenamento. In pratica usa una tecnica detta early-stopping, la quale interrompe immediatamente l'allenamento una volta che si verifica una diminuzione di prestazioni del modello dovute ad un ulteriore allenamento sui dati.

- Dropout, in pratica elimina casualmente alcuni neuroni durante l'addestramento, in modo che la rete non si concentri eccessivamente sui tali
- Data augmentation, utilizzato soprattutto per reti convolutive, si tratta di generare nuovi dati da quelli già presenti tramite trasformazioni come flip, rotate, skew e translate in modo da modificare leggermente i dati senza renderli irriconoscibili e per poter allenare la rete su pattern non troppo specifici.

Vedi web search and ranking

Spieghiamo una delle misure più usate, Il Normalized Discounted Cumulative Gain (NDCG) è una misura utilizzata per valutare l'efficacia di un sistema di classificazione. Viene comunemente utilizzato nella valutazione dei sistemi di ricerca e del recupero di informazioni. L'NDCG è una variazione della misura Discounted Cumulative Gain (DCG), dove la classificazione ideale è normalizzata in modo che abbia un valore massimo di 1. La valutazione NDCG si ottiene comparando la classificazione prevista degli elementi con la classificazione ideale e prendendo la somma cumulativa scontata dei punteggi di rilevanza degli elementi nella classificazione prevista, diviso per la somma cumulativa scontata dei punteggi di rilevanza nella classificazione ideale. Un punteggio NDCG più alto indica un sistema di classificazione migliore.

Raccolta esercizi pratici

Dato un dataset, trovare la radice del DT usando GINI Index

Lorem ipsum

Dato un dataset, trovare la radice del DT usando Information Gain

Lorem ipsum

Dato il seguente training set, usare un classificatore Bayesiano per predire la classe"PlayTennis" nel test set

Lorem ipsum

Dato un dataset, predire la classe di decisione per la nuova istanza specificata

Lorem ipsum

Data una matrice Transaction ID - Items, trovare l'item set di candidati X con A-Priori Algorithm, considerando 2 come supporto minimo. Calcolare anche la confidenza degli elementi dei candidati risultanti, con confidenza minima pari al 60%

Lorem ipsum

Data una matrice Transaction ID - Items, trovare il Frequent Pattern con l'algoritmo FP_Growth

Lorem ipsum

Filtri convoluzionali

Lorem ipsum