K-Nearest Neighbors

Lorenzo Arcioni

19 giugno 2024

Sommario

In questo articolo, presentiamo un'analisi approfondita dell'algoritmo K-Nearest Neighbors (KNN), esaminandolo sia dal punto di vista teorico che pratico. L'algoritmo KNN è un metodo di apprendimento supervisionato utilizzato per la classificazione e la regressione, basato sul principio che oggetti simili sono vicini nello spazio delle caratteristiche. Iniziamo con una descrizione dettagliata dei fondamenti teorici del KNN, compresa la definizione formale, i criteri di scelta del parametro K e le metriche di distanza utilizzate per determinare la vicinanza tra i dati. Successivamente, esploriamo le sue proprietà matematiche e discutiamo l'impatto della dimensionalità dei dati e del rumore sulla sua performance. Attraverso un'analisi empirica, confrontiamo l'efficacia del KNN con altri algoritmi di machine learning, utilizzando dataset standard. Infine, esaminiamo le tecniche di ottimizzazione e miglioramento del KNN, come la normalizzazione dei dati e l'uso di pesi nei vicini, per aumentare la precisione e l'efficienza computazionale. Questo studio offre una visione completa del KNN, evidenziando i suoi punti di forza, le sue limitazioni e le situazioni in cui è più adatto.

Indice

1 Introduzione

1.1 Panoramica dell'algoritmo K-Nearest Neighbors (KNN)

L'algoritmo K-Nearest Neighbors (KNN) rappresenta un pilastro fondamentale nell'ambito dell'apprendimento automatico supervisionato, apprezzato per la sua semplicità concettuale e la sua efficacia in una vasta gamma di applicazioni. La sua filosofia si basa sul principio intuitivo che oggetti simili tendono a raggrupparsi nello stesso spazio delle caratteristiche. Questo approccio non parametrico permette a KNN di adattarsi a strutture dati complesse e a relazioni non lineari, senza fare assunzioni rigide sulla distribuzione dei dati.

1.2 Funzionamento di KNN

KNN opera determinando le etichette di classificazione o i valori di regressione per un nuovo punto, basandosi sulla vicinanza ai punti di addestramento. Il parametro chiave di KNN è K, che rappresenta il numero di "vicini" più prossimi da considerare durante la fase di predizione. Quando un nuovo dato deve essere classificato, KNN calcola la distanza (come vedremo, esistono varie tipologie di distanze) tra il dato da classificare e tutti i punti di addestramento, quindi seleziona i K punti più vicini. La classe o il valore di regressione del nuovo dato è determinato dalla classe maggiormente rappresentata o dalla media dei valori nei punti vicini, rispettivamente.

1.3 Potenzialità di KNN

KNN trova applicazione in numerosi settori grazie alla sua flessibilità e facilità di implementazione. Nei sistemi di raccomandazione, ad esempio, può suggerire prodotti o contenuti simili a quelli preferiti dall'utente sulla base dei gusti di altri utenti simili (vicini). Nell'analisi di immagini e nel riconoscimento di pattern, KNN può classificare nuove immagini confrontandole con esempi già noti. In ambito medico, può supportare la diagnosi confrontando i sintomi del paziente con casi storici simili.

1.4 Caratteristiche e Limitazioni

Una delle caratteristiche distintive di KNN è la sua interpretabilità. Le decisioni di classificazione o regressione si basano direttamente sulla vicinanza tra i dati, rendendo il processo decisionale trasparente e facilmente comprensibile. Tuttavia, KNN può essere sensibile al rumore nei dati e alla presenza di feature non rilevanti, il che può influenzare negativamente le previsioni. Inoltre, gestire grandi dataset con KNN può essere computazionalmente oneroso, poiché richiede il calcolo delle distanze tra il nuovo dato e tutti i punti di addestramento; sopratutto quando si ha a che fare con dataset dimensionalmente complessi.

1.5 Definizione e concetto di base

1.5.1 Dataset, feature e variabile target

Un dataset in machine learning è una collezione di dati organizzati in un formato strutturato. Ogni riga del dataset rappresenta un'osservazione, mentre ogni colonna rappresenta una caratteristica (feature) o la variabile target (etichetta). Le feature sono attributi che descrivono le osservazioni e possono essere di diversi tipi, come numeriche, categoriche o binarie. La variabile target è ciò che vogliamo predire utilizzando le feature.

Ad esempio, in un dataset per la predizione del prezzo delle case, le feature potrebbero includere la superficie della casa, il numero di camere, la posizione, l'anno di costruzione, ecc. La variabile target sarebbe il prezzo della casa.

1.5.2 Problema di classificazione

Consideriamo un esempio reale: la diagnosi precoce di malattie cardiache. Questo è un problema non banale che richiede l'uso del machine learning per essere risolto efficacemente. Il dataset potrebbe includere pazienti con varie caratteristiche cliniche misurate durante esami medici. Le feature potrebbero includere età, genere, pressione sanguigna, livelli di colesterolo, frequenza cardiaca massima, risultati di elettrocardiogrammi, e altre misure cliniche rilevanti. La variabile target sarebbe una variabile binaria che indica la presenza o l'assenza di una malattia cardiaca.

L'algoritmo KNN può essere utilizzato per classificare un nuovo paziente come "a rischio" o "non a rischio" di malattia cardiaca basandosi sui dati storici di altri pazienti. Quando un nuovo paziente entra per una valutazione, KNN calcola le distanze tra le caratteristiche cliniche del nuovo paziente e quelle dei pazienti nel dataset di addestramento. Seleziona i K pazienti più simili (i vicini più prossimi) e determina la classe del nuovo paziente in base alla maggioranza delle classi dei vicini.

Per esempio, se $K = 5$ e tra i 5 pazienti più vicini al nuovo paziente 3 hanno una malattia cardiaca e 2 no, KNN predice che il nuovo paziente è "a rischio" di malattia cardiaca. Questa previsione può aiutare i medici a prendere decisioni informate riguardo ulteriori test o trattamenti, dimostrando l'importanza e l'utilità del Machine Learning in contesti medici critici.

1.5.3 Problema di regressione

Un esempio di problema di regressione risolvibile con l'algoritmo K-Nearest Neighbors (KNN) è la previsione del valore di mercato delle proprietà immobiliari in una città. Questo problema richiede un approccio che si affida al machine learning per ottenere stime accurate e affidabili, data la moltitudine di fattori che influenzano i prezzi delle case.

Consideriamo un dataset che include informazioni dettagliate sulle proprietà immobiliari di una città. Le feature possono includere:

- Superficie della proprietà (in metri quadrati)
- Numero di camere da letto
- Numero di bagni
- Anno di costruzione
- Distanza dai servizi principali (scuole, ospedali, trasporti pubblici)
- Valutazioni della qualità del quartiere
- Prezzi recenti delle proprietà vicine

La variabile target in questo caso è il prezzo di vendita della proprietà.

Quando si vuole stimare il valore di una nuova proprietà, l'algoritmo KNN calcola la distanza tra le caratteristiche della nuova proprietà e quelle delle proprietà nel dataset di addestramento. Utilizzando una metrica di distanza (come la distanza euclidea), KNN identifica i K immobili più simili.

Ad esempio, se $K = 5$, KNN selezionerà le cinque proprietà più vicine alla nuova proprietà in termini di caratteristiche. Il prezzo stimato per la nuova proprietà sarà la media dei prezzi delle cinque proprietà più vicine.

$$\hat{y} = \frac{1}{K} \sum_{i \in \mathcal{N}_K(\mathbf{x})} y_i$$

Dove $\hat{y}$ è il prezzo stimato della nuova proprietà, K è il numero di vicini considerati, $\mathcal{N}_K(\mathbf{x})$ rappresenta l'insieme dei K vicini più prossimi e y_i è il prezzo di una delle proprietà vicine.

Questo approccio di regressione basato su KNN è particolarmente utile perché tiene conto della località spaziale e delle caratteristiche specifiche delle proprietà immobiliari. Inoltre, permette di adattarsi a variazioni non lineari e complesse nei dati, che sono comuni nel mercato immobiliare. La previsione accurata dei prezzi immobiliari è fondamentale per acquirenti, venditori, agenti immobiliari e investitori, rendendo KNN uno strumento prezioso in questo contesto.

1.6 Sfide e Ottimizzazioni

La "maledizione della dimensionalità" è una delle sfide principali di KNN, in quanto la performance dell'algoritmo può decadere significativamente con l'aumento della dimensionalità dei dati. Per mitigare questo problema, sono state sviluppate tecniche come la riduzione della dimensionalità e l'uso di strutture dati specializzate (come KD-Trees) per accelerare il calcolo delle distanze.

1.7 Applicazioni Avanzate e Ricerca

La ricerca attuale su KNN si concentra sulla sua integrazione con tecniche avanzate di machine learning, come l'apprendimento semi-supervisionato e il trasferimento di conoscenza, per migliorare ulteriormente la sua robustezza e la sua capacità predittiva in scenari complessi.

Concludendo questa introduzione, KNN rimane una scelta potente e versatile per molte applicazioni di machine learning grazie alla sua semplicità, flessibilità e interpretabilità. Nonostante le sfide associate, continua a essere ampiamente utilizzato come punto di partenza per problemi di classificazione e regressione, offrendo risultati affidabili e interpretazioni chiare in vari contesti applicativi.

1.8 Obiettivi dell'articolo

L'obiettivo principale di questo articolo è fornire una comprensione completa e dettagliata dell'algoritmo K-Nearest Neighbors (KNN) attraverso un'analisi sia teorica che pratica. In primo luogo, l'articolo presenterà i fondamenti teorici di KNN, spiegando il funzionamento dell'algoritmo, le metriche di distanza utilizzate (come la distanza euclidea e la distanza di Manhattan) e l'importanza della scelta del parametro K. Verranno, inoltre, discusse le implicazioni di queste scelte sulla performance dell'algoritmo.

In secondo luogo, verranno esaminate le proprietà matematiche di KNN, inclusa la complessità computazionale e le sfide legate alla "maledizione della dimensionalità". Saranno analizzati i trade-off tra bias e varianza per comprendere come ottimizzare le prestazioni di questo algoritmo.

In terzo luogo, l'articolo fornirà un'analisi empirica, confrontando KNN con altri algoritmi di machine learning su dataset standard. Questo confronto aiuterà a evidenziare i punti di forza e le limitazioni di KNN in scenari pratici.

Infine, l'articolo esplorerà le tecniche di ottimizzazione e miglioramento di KNN, come la normalizzazione dei dati, l'uso di KNN pesato e l'implementazione di strutture dati efficienti come KD-Trees.

In sintesi, l'articolo mira a offrire una visione completa e dettagliata dell'algoritmo KNN.

2 Fondamenti Teorici del KNN

2.1 Definizione matematica formale

Per formalizzare matematicamente l'algoritmo K-Nearest Neighbors (KNN), consideriamo un dataset di addestramento $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, dove $\mathbf{x}_i \in \mathbb{R}^d$ rappresenta un punto dati con d caratteristiche e $y_i \in \mathbb{R}$ (o $y_i \in \{1, \ldots, C\}$ per la classificazione) rappresenta l'etichetta associata.

2.1.1 Distanza tra punti dati

Per determinare i K vicini più prossimi, è necessario definire una metrica di distanza $d(\mathbf{x}, \mathbf{z})$ tra due punti dati $\mathbf{x}$ e $\mathbf{z}$. Le metriche comunemente utilizzate includono:

- **Distanza Euclidea:**
$$d(\mathbf{x}, \mathbf{z}) = ||\mathbf{x} - \mathbf{z}|| = \sqrt{\sum_{j=1}^{d} (x_j - z_j)^2}$$

- **Distanza di Manhattan:**
$$d(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{d} |x_j - z_j|$$

- **Distanza di Minkowski:**
$$d(\mathbf{x}, \mathbf{z}) = \left(\sum_{j=1}^{d} |x_j - z_j|^p \right)^{\frac{1}{p}}$$
dove p è un parametro positivo che determina la forma della distanza.

2.1.2 Classificazione

Nel contesto della classificazione, l'etichetta $\hat{y}$ di un nuovo punto dati $\mathbf{x}$ è determinata come segue:

1. Calcolare la distanza tra $\mathbf{x}$ e ogni punto dati $\mathbf{x}_i$ nel dataset di addestramento.
2. Identificare i K punti più vicini a $\mathbf{x}$ utilizzando la metrica di distanza scelta.
3. Assegnare a $\mathbf{x}$ l'etichetta di classe più frequente tra i K vicini più prossimi. Formalmente,
$$\hat{y} = \arg \max_{c \in \{1,\ldots,C\}} \sum_{i \in \mathcal{N}_K(\mathbf{x})} \mathbf{1}_{\{y_i = c\}}$$
dove $\mathcal{N}_K(\mathbf{x})$ denota l'insieme dei K vicini più prossimi di $\mathbf{x}$ e $\mathbf{1}_{\{y_i=c\}}$ è una funzione indicatrice che vale 1 se $y_i = c$ e 0 altrimenti.

2.1.3 Regressione

Per la regressione, il valore predetto $\hat{y}$ per un nuovo punto dati $\mathbf{x}$ è calcolato come la media dei valori dei K vicini più prossimi:
$$\hat{y} = \frac{1}{K} \sum_{i \in \mathcal{N}_K(\mathbf{x})} y_i$$

Questa definizione matematica formale fornisce una chiara comprensione del funzionamento di base dell'algoritmo KNN, sia per la classificazione che per la regressione.

Riferimenti bibliografici

[1] Reinhard Diestel. *Graph Theory*. Springer Berlin Heidelberg, 2017.

[2] J. A. Bondy and U. S. R. Murty. *Graph Theory*. Springer London, 2008.