

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	2
1.1	Panoramica dell'algoritmo K-Nearest Neighbors (KNN)	2
1.2	Importanza e applicazioni del KNN	3
1.3	Obiettivi dell'articolo	3
2	Fondamenti Teorici del KNN	3
2.1	Definizione e concetto di base	3
2.2	Definizione matematica formale	3
2.3	Scelta del parametro K	3
2.4	Metriche di distanza	3
2.4.1	Distanza Euclidea	3
2.4.2	Distanza di Manhattan	3
2.4.3	Distanza di Minkowski	3
2.4.4	Altre metriche di distanza	3
3	Proprietà Matematiche e Analisi Teorica	3
3.1	La maledizione della dimensionalità	3
3.2	Complessità computazionale	3
3.3	Trade-off bias-varianza nel KNN	3
3.4	Interpretazione probabilistica del KNN	3
3.5	Comportamento asintotico e convergenza	3
4	Analisi Teorica	3
4.1	La maledizione della dimensionalità	3
4.2	Complessità computazionale	3
4.3	Trade-off bias-varianza nel KNN	3
4.4	Interpretazione probabilistica del KNN	3
4.5	Comportamento asintotico e convergenza	3

1 Introduzione

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

L'algoritmo K-Nearest Neighbors (KNN) è un metodo di apprendimento supervisionato utilizzato sia per problemi di classificazione che di regressione. La sua essenza risiede nel principio di vicinanza: gli oggetti simili tendono a trovarsi vicini nello spazio delle caratteristiche. Questa caratteristica rende il KNN intuitivo e semplice da implementare, pur essendo potente in molte applicazioni pratiche.

KNN è un metodo basato sulla prossimità, il che significa che, al momento della previsione per un nuovo dato, l'algoritmo cerca i K punti di addestramento più vicini (i "vicini") e utilizza le loro informazioni per fare la previsione. Per i problemi di classificazione, KNN assegna l'etichetta più comune tra i vicini; per i problemi di regressione, calcola la media dei valori dei vicini.

Un aspetto fondamentale del KNN è la scelta del parametro K, che rappresenta il numero di vicini da considerare. La scelta di K influisce significativamente sulla performance dell'algoritmo: un K troppo piccolo può rendere il modello sensibile al rumore (overfitting), mentre un K troppo grande può diluire la precisione del modello (underfitting).

Un altro elemento critico del KNN è la metrica di distanza utilizzata per determinare la vicinanza tra i punti. Le metriche di distanza più comuni includono la distanza euclidea, la distanza di Manhattan e la distanza di Minkowski, ognuna delle quali ha proprietà diverse che possono influenzare i risultati in base alla natura dei dati.

Nonostante la sua semplicità, KNN presenta alcune sfide, in particolare riguardo alla gestione di grandi dataset e alla sensibilità alla dimensionalità dei dati. Tuttavia, grazie alla sua natura non parametrica e alla facilità di implementazione, rimane un metodo popolare e ampiamente utilizzato in molte applicazioni di machine learning.

1.2 Importanza e applicazioni del KNN

1.3 Obiettivi dell'articolo

2 Fondamenti Teorici del KNN

2.1 Definizione e concetto di base

2.2 Definizione matematica formale

2.3 Scelta del parametro K

2.4 Metriche di distanza

2.4.1 Distanza Euclidea

2.4.2 Distanza di Manhattan

2.4.3 Distanza di Minkowski

2.4.4 Altre metriche di distanza

3 Proprietà Matematiche e Analisi Teorica

3.1 La maledizione della dimensionalità

3.2 Complessità computazionale

3.3 Trade-off bias-varianza nel KNN

3.4 Interpretazione probabilistica del KNN

3.5 Comportamento asintotico e convergenza

4 Analisi Teorica

4.1 La maledizione della dimensionalità

4.2 Complessità computazionale

4.3 Trade-off bias-varianza nel KNN

4.4 Interpretazione probabilistica del KNN

4.5 Comportamento asintotico e convergenza

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.