**Tuần 6+7 ÁP DỤNG KNN CHO BÀI TOÁN PHÂN LOẠI VÀ BÀI TOÁN HỒI QUY**

Thuật toán KNN (K-Nearest Neighbors) là một thuật toán học máy có giám sát, được sử dụng rộng rãi trong các bài toán phân loại và hồi quy. Thuật toán này dựa trên giả định rằng các dữ liệu tương tự nhau sẽ có cùng nhãn.

* **Ứng dụng thuật toán KNN cho bài toán phân loại**

Trong bài toán phân loại, thuật toán KNN sẽ phân loại một điểm dữ liệu mới dựa trên nhãn của các điểm dữ liệu gần nó nhất. Số lượng điểm dữ liệu gần nhất được gọi là K.

Cách hoạt động của thuật toán KNN

Thuật toán KNN có thể được chia thành các bước sau:

Tập dữ liệu huấn luyện: Tập dữ liệu huấn luyện gồm các điểm dữ liệu đã biết nhãn.

Chọn K: Chọn số lượng điểm dữ liệu gần nhất (K).

Tính khoảng cách: Tính khoảng cách giữa điểm dữ liệu mới và các điểm dữ liệu trong tập dữ liệu huấn luyện.

Lựa chọn nhãn: Lựa chọn nhãn của điểm dữ liệu mới dựa trên nhãn của các điểm dữ liệu gần nhất.

Đặc điểm của thuật toán KNN

Ưu điểm:

Đơn giản và dễ hiểu.

Dễ triển khai và có thể áp dụng cho nhiều tập dữ liệu khác nhau.

Có thể giải quyết các bài toán phân loại phức tạp.

Nhược điểm:

Độ chính xác của mô hình phụ thuộc vào giá trị K.

Có thể gặp phải hiện tượng overfitting nếu K quá lớn.

Một số ví dụ về ứng dụng thuật toán KNN cho bài toán phân loại

Phân loại hình ảnh: Thuật toán KNN có thể được sử dụng để phân loại hình ảnh dựa trên các đặc điểm của hình ảnh. Ví dụ, thuật toán KNN có thể được sử dụng để phân loại ảnh người, ảnh động vật hoặc ảnh đồ vật.

Phân loại văn bản: Thuật toán KNN có thể được sử dụng để phân loại văn bản dựa trên nội dung của văn bản. Ví dụ, thuật toán KNN có thể được sử dụng để phân loại email là spam hay không spam, hoặc phân loại bài báo là tin tức hay quảng cáo.

Phân loại khách hàng: Thuật toán KNN có thể được sử dụng để phân loại khách hàng dựa trên các đặc điểm của khách hàng. Ví dụ, thuật toán KNN có thể được sử dụng để phân loại khách hàng là tiềm năng hay không tiềm năng, hoặc phân loại khách hàng là trung thành hay không trung thành.

Ví dụ:

Giả sử chúng ta có tập dữ liệu về màu sắc của hoa, với hai lớp: hoa hồng và hoa cúc. Tập dữ liệu bao gồm các đặc điểm sau:

Chiều dài cánh hoa: Số lượng pixel trong hình ảnh của cánh hoa.

Chiều rộng cánh hoa: Số lượng pixel trong hình ảnh của cánh hoa.

Chúng ta có thể sử dụng thuật toán KNN để xây dựng một mô hình phân loại hoa hồng và hoa cúc. Để làm điều này, chúng ta cần thu thập dữ liệu về hoa hồng và hoa cúc, chuẩn hóa dữ liệu và phân chia dữ liệu thành tập huấn luyện và tập kiểm tra.

Tiếp theo, chúng ta cần chọn giá trị k. Chúng ta có thể bắt đầu với giá trị k = 3 và tăng giá trị k cho đến khi chúng ta đạt được độ chính xác mong muốn.

Cuối cùng, chúng ta có thể sử dụng tập kiểm tra để đánh giá mô hình KNN. Nếu mô hình đạt được độ chính xác cao, chúng ta có thể sử dụng mô hình để phân loại hoa hồng và hoa cúc.

Ví dụ

Giả sử chúng ta có tập dữ liệu huấn luyện gồm 10 điểm dữ liệu, mỗi điểm dữ liệu có 2 thuộc tính:

Thuộc tính 1 Thuộc tính 2 Lớp

0 0 A

1 1 A

2 2 A

3 3 A

4 4 B

5 5 B

6 6 B

7 7 B

8 8 B

Để phân loại điểm dữ liệu mới có thuộc tính 1 = 9 và thuộc tính 2 = 9, ta thực hiện các bước sau:

Tính khoảng cách giữa điểm dữ liệu mới và 10 điểm dữ liệu trong tập huấn luyện.

Chọn ra 5 điểm dữ liệu gần nhất với điểm dữ liệu mới.

Trong 5 điểm dữ liệu gần nhất, có 4 điểm dữ liệu thuộc lớp A và 1 điểm dữ liệu thuộc lớp B.

Do đó, điểm dữ liệu mới sẽ được gán nhãn lớp A.

Ví dụ:

Giả sử chúng ta có tập dữ liệu đào tạo sau:

| Điểm dữ liệu | Nhãn |

|---|---|

| [1, 2] | A |

| [3, 4] | B |

| [5, 6] | C |

| [7, 8] | D |

Tập dữ liệu này có 4 điểm dữ liệu, mỗi điểm dữ liệu có hai thuộc tính và một nhãn. Chúng ta có thể chọn tham số K là 3.

Để phân loại điểm dữ liệu mới [9, 10], chúng ta sẽ tìm ba điểm dữ liệu gần nhất với điểm dữ liệu này. Các điểm dữ liệu gần nhất với điểm dữ liệu [9, 10] là [7, 8], [5, 6] và [3, 4]. Nhãn của ba điểm dữ liệu này là D, C và B. Do đó, chúng ta sẽ dự đoán nhãn của điểm dữ liệu [9, 10] là B.

Một số ứng dụng thực tế của thuật toán KNN cho bài toán hồi quy:

Dự đoán giá nhà

Dự đoán doanh số bán hàng

Dự đoán chi phí sản xuất

Dự đoán điểm thi

Dự đoán thời tiết

Ví dụ:

Giả sử chúng ta có tập dữ liệu sau về giá trị của một căn nhà và số phòng ngủ:

Số phòng ngủ Giá trị (USD)

2 100.000

3 200.000

4 300.000

5 400.000

Chúng ta muốn sử dụng KNN để dự đoán giá trị của một căn nhà có 6 phòng ngủ.

Để làm được điều này, chúng ta cần chọn một giá trị K. Giá trị K càng lớn thì kết quả dự đoán càng ít bị ảnh hưởng bởi các điểm dữ liệu ngoại lệ. Trong ví dụ này, chúng ta sẽ chọn K = 3.

Tiếp theo, chúng ta sẽ tìm ra 3 điểm dữ liệu gần nhất với điểm dữ liệu có 6 phòng ngủ trong tập dữ liệu huấn luyện.

Số phòng ngủ Giá trị (USD)

5 400.000

4 300.000

3 200.000

Giá trị trung bình của 3 điểm dữ liệu này là 300.000 USD. Do đó, chúng ta dự đoán giá trị của một căn nhà có 6 phòng ngủ là 300.000 USD.

* **Ứng dụng KNN cho bài toán hồi quy**

Trong bài toán hồi quy, thuật toán KNN được sử dụng để dự đoán giá trị đầu ra y của một điểm dữ liệu mới x dựa trên các điểm dữ liệu gần nhất với x trong tập huấn luyện. Thuật toán này hoạt động như sau:

1. Tính khoảng cách giữa điểm dữ liệu mới x và các điểm dữ liệu trong tập huấn luyện.
2. Chọn K điểm dữ liệu gần nhất với x.
3. Tính giá trị trung bình của các điểm dữ liệu K này để dự đoán giá trị đầu ra y của điểm dữ liệu mới x.

Ví dụ

Giả sử chúng ta có tập dữ liệu sau:

|  |  |
| --- | --- |
| x | y |
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |

drive\_spreadsheetXuất sang Trang tính

Chúng ta muốn sử dụng thuật toán KNN để dự đoán giá trị đầu ra y của điểm dữ liệu mới x = 4.

1. Tính khoảng cách giữa điểm dữ liệu mới x = 4 và các điểm dữ liệu trong tập huấn luyện.

distance(1, 4) = 3

distance(2, 4) = 2

distance(3, 4) = 1

distance là khoảng cách giữa hai điểm dữ liệu

1. Chọn K = 2 điểm dữ liệu gần nhất với x = 4.

K = 2

sorted\_distances = [distance(1, 4), distance(2, 4)]

sorted\_distances.sort()

sorted\_distances = [1, 2]

closest\_points = [1, 2]

sorted\_distances là một danh sách các khoảng cách giữa điểm dữ liệu mới x và các điểm dữ liệu trong tập huấn luyện, được sắp xếp theo thứ tự tăng dần.

1. Tính giá trị trung bình của các điểm dữ liệu K này để dự đoán giá trị đầu ra y của điểm dữ liệu mới x.

y\_hat = (2 + 3) / 2

y\_hat = 2.5

Trong bài toán hồi quy, y\_hat là giá trị trung bình của các điểm dữ liệu K gần nhất với x.

Vậy, thuật toán KNN dự đoán giá trị đầu ra y của điểm dữ liệu mới x = 4 là 2.5.

Ưu điểm của KNN trong bài toán hồi quy

* Đơn giản, dễ hiểu và dễ triển khai.
* Không cần giả định gì về phân phối của dữ liệu.
* Có thể được sử dụng để hồi quy các dữ liệu có số lượng đặc trưng lớn.
* Có thể được sử dụng để hồi quy các dữ liệu không tuyến tính.

Nhược điểm của KNN trong bài toán hồi quy

* Độ chính xác phụ thuộc vào số lượng láng giềng được sử dụng (K).
* Độ phức tạp phụ thuộc vào số chiều của không gian đặc trưng.
* Không thể học các đặc trưng phức tạp của dữ liệu.

Các phương pháp cải thiện hiệu quả của KNN trong bài toán hồi quy

* Thử nghiệm các giá trị K khác nhau: Chọn giá trị K phù hợp với tập dữ liệu.
* Sử dụng các thước đo khoảng cách khác nhau: Chọn thước đo khoảng cách phù hợp với tập dữ liệu.
* Sử dụng kỹ thuật đánh trọng số: Gán các trọng số khác nhau cho các láng giềng để cải thiện độ chính xác trong trường hợp dữ liệu bị nhiễu hoặc phân lớp không đồng đều.

Kết luận

Thuật toán KNN là một thuật toán học máy đơn giản và hiệu quả. Thuật toán này có thể được sử dụng để hồi quy các dữ liệu có số lượng đặc trưng lớn và không tuyến tính. Để cải thiện hiệu quả của KNN trong bài toán hồi quy, có thể áp dụng các phương pháp như thử nghiệm các giá trị K khác nhau, sử dụng các thước đo khoảng cách khác nhau và sử dụng kỹ thuật đánh trọng số.