**Tuần 4: CÁC LOẠI KHOẢNG CÁCH TRONG KNN, CÁCH CHỌN THAM SỐ K VÀ CÁC PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU TRƯỚC KHI SỬ DỤNG THUẬT TOÁN KNN**

* **Các loại khoảng cách được sử dụng trong KNN:**

Khoảng cách là một phép đo mức độ gần gũi giữa hai điểm dữ liệu. Có nhiều loại khoảng cách khác nhau có thể được sử dụng trong KNN, bao gồm:

* Khoảng cách Euclide: Khoảng cách Euclide là khoảng cách tiêu chuẩn giữa hai điểm dữ liệu trong không gian Euclide. Nó được tính bằng công thức sau:

d(x, y) = √(x1 - y1)^2 + (x2 - y2)^2 + ... + (xn - yn)^2

* Khoảng cách Manhattan: Khoảng cách Manhattan là khoảng cách tổng của các khoảng cách theo trục cho hai điểm dữ liệu. Nó được tính bằng công thức sau:

d(x, y) = |x1 - y1| + |x2 - y2| + ... + |xn - yn|

* Khoảng cách Hamming: Khoảng cách Hamming là số lượng bit khác nhau giữa hai điểm dữ liệu. Nó được tính bằng công thức sau:

d(x, y) = |x1 ⊕ y1| + |x2 ⊕ y2| + ... + |xn ⊕ yn|

* Khoảng cách Minkowski: Khoảng cách Minkowski là một tổng có trọng số của các khoảng cách theo trục cho hai điểm dữ liệu. Nó được tính bằng công thức sau:

d(x, y) = (|x1 - y1|^p + |x2 - y2|^p + ... + |xn - yn|^p)^(1/p)

* **Cách chọn tham số K**
* K là một tham số quan trọng của thuật toán KNN, quyết định mức độ ảnh hưởng của các điểm dữ liệu gần nhất đến quyết định phân loại hoặc dự đoán. K càng lớn thì quyết định của KNN sẽ càng dựa trên ý kiến của nhiều điểm dữ liệu.
* Cách chọn tham số K phụ thuộc vào dữ liệu và nhiệm vụ cụ thể. Một số cách chọn tham số K phổ biến bao gồm:

+ Sử dụng phương pháp cross-validation: Phương pháp cross-validation sẽ chia tập dữ liệu thành các tập con, sau đó sử dụng mỗi tập con làm tập kiểm tra để đánh giá hiệu quả của KNN với các giá trị K khác nhau.

+ Sử dụng phương pháp grid search: Phương pháp grid search sẽ thử nghiệm tất cả các giá trị K trong một tập hợp các giá trị K xác định, sau đó chọn giá trị K có hiệu quả nhất.

+ Sử dụng phương pháp dựa trên kiến thức: Phương pháp này sẽ sử dụng kiến thức về dữ liệu và nhiệm vụ cụ thể để chọn tham số K.

* **Trước khi sử dụng thuật toán KNN, cần phải xử lý dữ liệu để đảm bảo rằng dữ liệu phù hợp với thuật toán.**

Một số phương pháp xử lý dữ liệu trước khi sử dụng thuật toán KNN bao gồm:

Loại bỏ các điểm dữ liệu ngoại lệ: Các điểm dữ liệu ngoại lệ là các điểm dữ liệu nằm cách xa các điểm dữ liệu khác trong không gian. Các điểm dữ liệu ngoại lệ có thể làm giảm hiệu quả của KNN.

Tiền xử lý dữ liệu nhiễu: Dữ liệu nhiễu là dữ liệu sai lệch hoặc không chính xác. Dữ liệu nhiễu có thể làm giảm hiệu quả của KNN.

Chuẩn hóa dữ liệu: Chuẩn hóa dữ liệu sẽ đưa tất cả các thuộc tính dữ liệu về cùng một phạm vi. Chuẩn hóa dữ liệu có thể giúp KNN hoạt động hiệu quả hơn.

Tính toán trọng số cho các điểm dữ liệu: Trọng số cho các điểm dữ liệu có thể được sử dụng để điều chỉnh mức độ ảnh hưởng của các điểm dữ liệu gần nhất đến quyết định phân loại hoặc dự đoán.

Dưới đây là mô tả chi tiết về từng phương pháp:

**Loại bỏ các điểm dữ liệu ngoại lệ:**

Xác định các điểm dữ liệu ngoại lệ dựa trên các giá trị thống kê: Các điểm dữ liệu có giá trị nằm ngoài một số ngưỡng thống kê nhất định có thể được coi là các điểm dữ liệu ngoại lệ.

Xác định các điểm dữ liệu ngoại lệ dựa trên các thuộc tính: Các điểm dữ liệu có các thuộc tính nằm ngoài một số phạm vi nhất định có thể được coi là các điểm dữ liệu ngoại lệ.

**Dữ liệu nhiễu có thể được xử lý bằng cách sử dụng các kỹ thuật như:**

Xóa dữ liệu nhiễu: Dữ liệu nhiễu có thể được xóa trực tiếp khỏi tập dữ liệu.

Sửa dữ liệu nhiễu: Dữ liệu nhiễu có thể được sửa chữa bằng cách thay thế các giá trị sai lệch bằng các giá trị chính xác.

Giảm thiểu tác động của dữ liệu nhiễu: Dữ liệu nhiễu có thể được giảm thiểu tác động bằng cách sử dụng các kỹ thuật như bình thường hóa dữ liệu hoặc tính toán trọng số cho các điểm dữ liệu.

**Dữ liệu có thể được chuẩn hóa bằng cách sử dụng các kỹ thuật như:**

Chuẩn hóa min-max: Chuẩn hóa min-max sẽ chuyển đổi tất cả các giá trị dữ liệu về cùng một phạm vi, từ 0 đến 1.

Chuẩn hóa z-score: Chuẩn hóa z-score sẽ chuyển đổi tất cả các giá trị dữ liệu về cùng một giá trị trung bình và độ lệch chuẩn.

**Trọng số cho các điểm dữ liệu có thể được tính toán bằng cách sử dụng các kỹ thuật như:**

Trọng số dựa trên khoảng cách: Trọng số của các điểm dữ liệu gần nhất có thể được tăng lên để tăng cường ảnh hưởng của chúng đến quyết định phân loại hoặc dự đoán.

Trọng số dựa trên độ tin cậy: Trọng số của các điểm dữ liệu có thể được tăng lên nếu chúng có độ tin cậy cao.