**Tuần 3: Nguyên lý hoạt động và ưu nhược điểm của KNN**

1. **Nguyên lý hoạt động KNN**

Nguyên lý hoạt động của thuật toán K-nearest neighbors (KNN) là dựa trên ý tưởng rằng các điểm dữ liệu tương tự nhau sẽ nằm gần nhau trong không gian. Để phân loại một điểm dữ liệu mới, KNN sẽ tìm K điểm dữ liệu gần nhất với điểm dữ liệu mới đó, sau đó gán điểm dữ liệu mới cho lớp mà K điểm dữ liệu gần nhất thuộc về.

* **Các bước hoạt động của KNN**

Để phân loại một điểm dữ liệu mới, KNN sẽ thực hiện các bước sau:

* Tính toán khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập huấn luyện.
* Sắp xếp các điểm dữ liệu theo khoảng cách, từ gần nhất đến xa nhất.
* Chọn K điểm dữ liệu gần nhất.
* Gán điểm dữ liệu mới cho lớp mà K điểm dữ liệu gần nhất thuộc về.

**1. Tính toán khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập huấn luyện**

Đầu tiên, KNN sẽ tính toán khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập huấn luyện. Khoảng cách có thể được tính toán bằng nhiều cách khác nhau, chẳng hạn như khoảng cách Euclide, khoảng cách Manhattan, khoảng cách Hamming, hoặc khoảng cách Minkowski.

**2. Sắp xếp các điểm dữ liệu theo khoảng cách, từ gần nhất đến xa nhất**

Sau khi tính toán khoảng cách, KNN sẽ sắp xếp các điểm dữ liệu theo khoảng cách, từ gần nhất đến xa nhất.

**3. Chọn K điểm dữ liệu gần nhất**

Tiếp theo, KNN sẽ chọn K điểm dữ liệu gần nhất. Số lượng điểm dữ liệu gần nhất được chọn là một tham số được gọi là K.

**4. Gán điểm dữ liệu mới cho lớp mà K điểm dữ liệu gần nhất thuộc về**

Cuối cùng, KNN sẽ gán điểm dữ liệu mới cho lớp mà K điểm dữ liệu gần nhất thuộc về.

1. **Ưu Nhược điểm KNN:**

* **Ưu điểm của thuật toán KNN**
* Đơn giản và dễ hiểu: KNN là một thuật toán đơn giản và dễ hiểu. Không cần giả định phân phối dữ liệu hay học một mô hình phức tạp.
* Khả năng xử lý dữ liệu phi cấu trúc: KNN có khả năng xử lý dữ liệu phi cấu trúc, không yêu cầu các giả định về cấu trúc dữ liệu. Điều này cho phép nó áp dụng cho nhiều loại dữ liệu, bao gồm cả dữ liệu văn bản, hình ảnh, và âm thanh.
* Hiệu suất tốt đối với tập dữ liệu nhỏ: KNN hoạt động tốt trên các tập dữ liệu nhỏ với số lượng mẫu ít. Nó không đòi hỏi quá nhiều tính toán trước khi thực hiện dự đoán.
* Dễ dàng tinh chỉnh tham số: K là tham số quan trọng trong thuật toán KNN. Việc điều chỉnh giá trị K có thể ảnh hưởng đến hiệu suất của thuật toán. Tuy nhiên, việc tinh chỉnh K là khá dễ dàng và có thể được thực hiện thông qua quá trình thử và sai.
* **Nhược điểm của thuật toán KNN**
* Phụ thuộc vào kích thước dữ liệu: KNN có hiệu suất giảm khi tập dữ liệu lớn vì việc tính toán khoảng cách giữa các điểm dữ liệu mất nhiều thời gian. KNN không phải là thuật toán phù hợp cho việc xử lý các tập dữ liệu lớn.
* Nhạy cảm với nhiễu và dữ liệu không đồng nhất: KNN dễ bị ảnh hưởng bởi các nhiễu trong dữ liệu và các điểm dữ liệu nằm trong các lớp khác nhau gần nhau. Điều này có thể dẫn đến việc phân loại không chính xác hoặc không ổn định.
* Cần xử lý các biến số và mất cân bằng dữ liệu: KNN không xử lý được các biến số khác nhau và cần sự cân bằng dữ liệu trong các lớp khác nhau. Nếu một lớp có số lượng mẫu nhiều hơn so với lớp khác, KNN có thể dễ dàng bị thiên vị và cho ra kết quả không chính xác.
* Yêu cầu lưu trữ toàn bộ dữ liệu huấn luyện: KNN yêu cầu lưu trữ toàn bộ dữ liệu huấn luyện trong bộ nhớ để tính toán khoảng cách và tìm láng giềng gần nhất. Điều này có thể là một vấn đề khi làm việc với các tập dữ liệu lớn và yêu cầu nhiều tài nguyên.