**Tuần 2: Sơ lược về KNN**

**K-nearest neighbors (KNN) là một thuật toán học máy có giám sát, được sử dụng để phân loại và hồi quy. KNN hoạt động bằng cách tìm các điểm dữ liệu gần nhất với điểm dữ liệu mới cần phân loại hoặc dự đoán.**

Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại **gần nhau** trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm củng có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp.

* Trong phân loại, KNN sẽ gán điểm dữ liệu mới cho lớp mà các điểm dữ liệu gần nhất thuộc về. Số lượng điểm dữ liệu gần nhất được sử dụng là một tham số được gọi là K.
* Trong hồi quy, KNN sẽ sử dụng các điểm dữ liệu gần nhất để ước tính giá trị của điểm dữ liệu mới.

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

* Khoảng cách Euclide
* Khoảng cách Manhattan
* Khoảng cách Hamming
* Khoảng cách Minkowski

**Các tham số của KNN:**

* K: Số lượng điểm dữ liệu gần nhất được sử dụng để phân loại hoặc dự đoán.
* Khoảng cách: Loại khoảng cách được sử dụng để tính toán khoảng cách giữa các điểm dữ liệu.
* Trọng số: Trọng số được gán cho các điểm dữ liệu gần nhất.

**Cách thức 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 các điểm dữ liệu gần nhất thuộc về.

**Ví dụ:**

Bài toán đặt ra: Bạn có điểm của một môn học nhưng bạn không biết thuộc loại nào (Giỏi, khá, trung bình, yếu). Giả sử bạn không biết bất kì quy tắc nào để phân loại cả.

Có một cách giải quyết là bạn phải đi khảo sát những người xung quanh. Để biết điểm của mình thuộc loại nào thì bạn phải đi hỏi những đứa có điểm gần số điểm mình nhất. Giả sử trong lớp 50 đứa, mình khảo sát 5 đứa gần điểm mình nhất và được dữ liệu như sau:

Điểm của tôi: 7

Điểm của bạn tôi:

* 7.1 => Khá
* 7.2 => Khá
* 6.7 => Khá
* 6.6 => Khá
* 6.4 => Trung bình

Qua kết quả trên thì mình sẽ mạnh dạng đoán là mình loại khá đúng không nào? Với cách này chúng ta có thể phân loại dữ liệu 1 chiều (1 feature) bằng cách làm khá đơn giản. Và các bạn có nhận thấy rằng dữ liệu mình khảo sát càng nhiều, càng rộng thì dự đoán đưa ra càng chính xác (Giả sử lớp bạn không có ai loại khá ngoài bạn thì cho dù bạn lấy bao nhiêu người gần điểm bạn nhất củng sẽ ra kết quả sai).

**Ưu điểm của KNN:**

Đơn giản và dễ hiểu

Không cần tính toán phức tạp

Có thể áp dụng cho nhiều loại dữ liệu

Có thể xử lý dữ liệu nhiễu

**Nhược điểm của KNN:**

Có thể không chính xác cho các tập dữ liệu lớn

Có thể không hiệu quả cho các tập dữ liệu có nhiều chiều

**KNN được sử dụng trong nhiều ứng dụng khác nhau, bao gồm:**

* Sử dụng KNN để phân loại hình ảnh hoa.
* Sử dụng KNN để phân loại văn bản spam.
* Sử dụng KNN để phân loại âm thanh nhạc rock và nhạc pop.
* Sử dụng KNN để phân loại đối tượng trong video.
* Sử dụng KNN để dự đoán giá cổ phiếu.
* Sử dụng KNN để dự đoán doanh số bán hàng.
* Sử dụng KNN để dự đoán thời tiết.