Thuật toán K-nearest neighbors (KNN)

Mentor: PĐ Hải

Người trình bày: Nguyễn Minh Hiếu

PAYT CLUB PTIT

August 2025

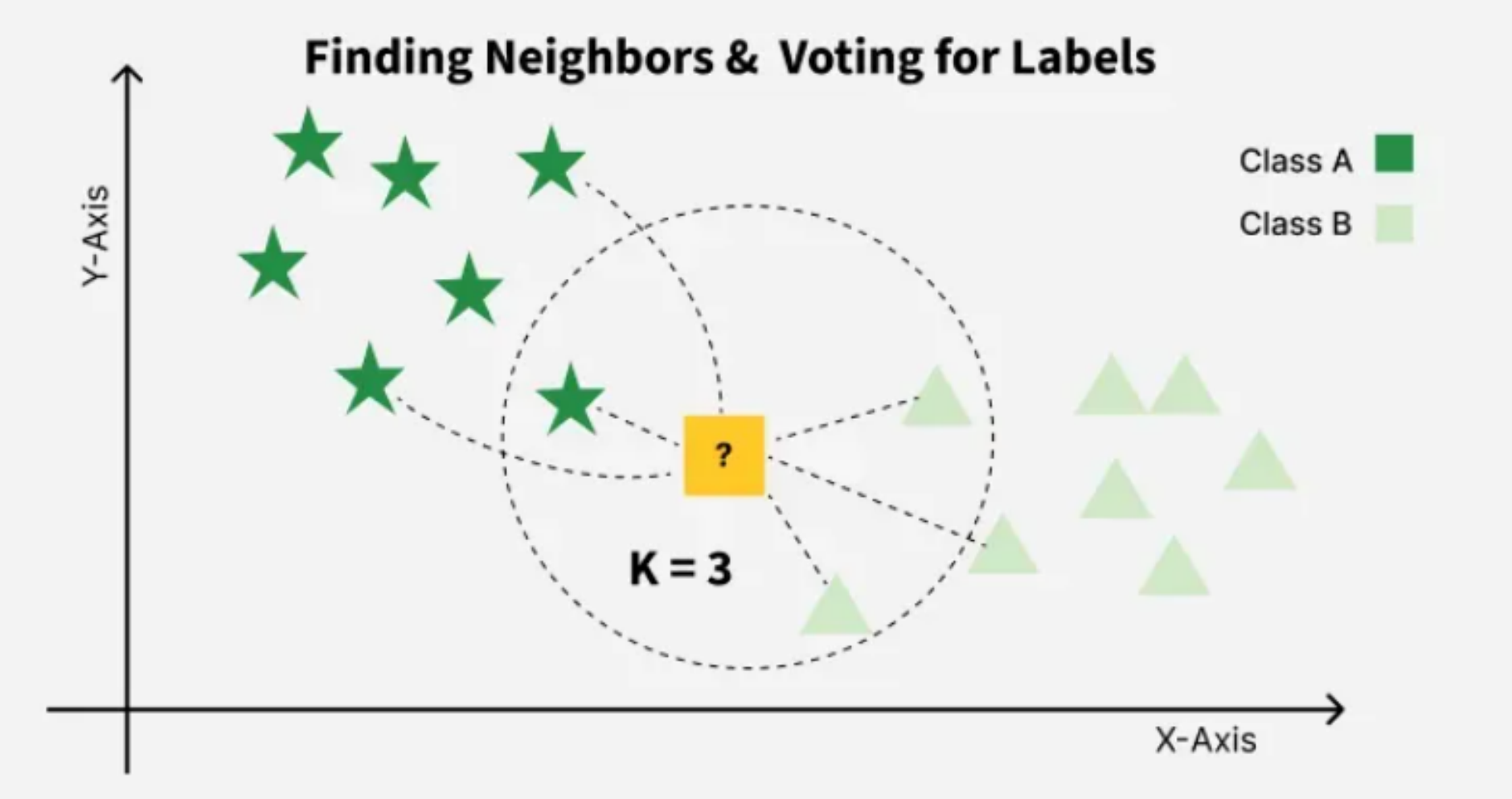
Nội dung

1: Lý thuyết

1: Lý thuyết

- KNN là thuật toán khi training nó **không học** gì, chỉ lưu **toàn bộ** dữ liệu
- Có ứng dụng vào **classification** và **regression** trong Supervised learning
- Khi cần dự đoán nó tìm k điểm **gần nhất** với đầu vào mới, rồi lấy kết quả theo **đa số phiếu** (classification) hoặc **trung bình** (regression) của k điểm này

1: Lý thuyết


Finding Neighbors & Voting for Labels
Class A
Class B
Y-Axis
X-Axis
?
K = 3

2: Thuật toán

2: Thuật toán

Thuật toán KNN:

- Bước 1: Chọn giá trị k tối ưu (k = $\sqrt{|\mathbf{X}_{\text{train}}|}$)
- Bước 2: Tính khoảng cách giữa data points và target point (Sử dụng khoảng cách euclid)

$$D(x_0, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{d} (x_0 - y_i)^2}$$

- Bước 3: Tìm các điểm gần nhất dựa vào khoảng cách đã tính
- Bước 4: Voting đối với bài toán classification (majority voting), lấy average đối với bài toán regression

Ví dụ

name	age	gender	sport
A1	32	0	Football
A2	40	0	Neither
A3	16	1	Chess
A4	34	1	Chess
A5	55	0	Neither

T: 5 age, gender = 1 (female)

$$d = \sqrt{(age_s - age_i)^2 + (gender_s - gender_i)^2}$$

Ví dụ

Khoảng cách giữa T và A1 (tuổi = 32, giới tính = 0):

$$d = \sqrt{(5-32)^2 + (1-0)^2} = \sqrt{729+1} = \sqrt{730} \approx 27.02$$

Khoảng cách giữa T và A2 (tuổi = 40, giới tính = 0):

$$d = \sqrt{(5-40)^2 + (1-0)^2} = \sqrt{1225+1} = \sqrt{1226} \approx 35.01$$

Khoảng cách giữa T và A3 (tuổi = 16, giới tính = 1):

$$d = \sqrt{(5-16)^2 + (1-1)^2} = \sqrt{121+0} = \sqrt{121} = 11.00$$

Khoảng cách giữa T và A4 (tuổi = 34, giới tính = 1):

$$d = \sqrt{(5-34)^2 + (1-1)^2} = \sqrt{841+0} = \sqrt{841} = 29.00$$

Khoảng cách giữa T và A5 (tuổi = 55, giới tính = 0):

$$d = \sqrt{(5-55)^2 + (1-0)^2} = \sqrt{2500+1} = \sqrt{2501} \approx 50.01$$

Ví dụ

Với k = 3:

- 1: A3 (Chess) - 11.00
- 2: A1 (Football) - 27.02
- 3: A4 (Chess) - 29.00

Theo bỏ phiếu => T sẽ chơi Chess

3: Coding

3: Coding

So sánh:

- Code khi không sử dụng thư viện có sẵn
- Code khi sử dụng thư viện có sẵn

4: Thảo luận

4: Thảo luận

4.1: Ứng dụng

- Recommendation Systems: Ví dụ gợi ý phim hoặc sản phẩm bằng cách tìm người dùng có cùng sở thích
- Spam Detection: Nhận dạng bằng cách so sánh email mới với email spam hoặc no-spam đã biết
- Speech Recognition: So khớp các âm thanh giọng nói với từ ngữ đã biết để chuyển đổi thành văn bản chính xác

4: Thảo luận

4.2: Thuận lợi

- Dễ hiểu, dễ sử dụng
- No training step: Không cần train vì data được lưu và sử dụng trong quá trình predict
- Ít tham số: Chỉ cần k và khoảng cách
- Linh hoạt: Ứng dụng cho cả bài toán phân loại và hồi quy

4: Thảo luận

4.3: Bất lợi

- Nhạy cảm với nhiễu khi k nhỏ
- Khi k lớn hoặc dữ liệu lớn thì chạy chậm