AI VIET NAM – COURSE 2024

KNN

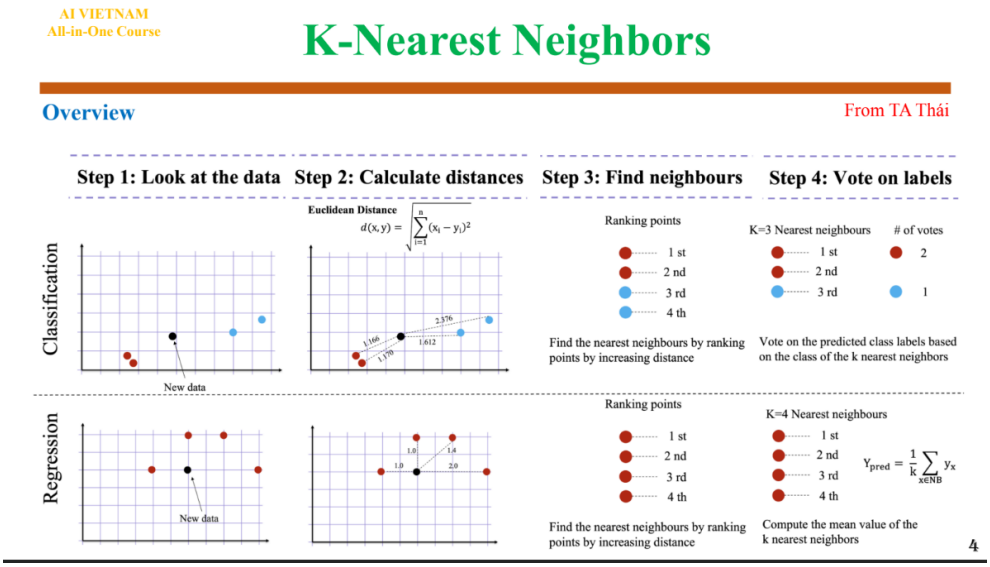
Ngày 19 tháng 2 năm 2024

Ngày tóm tắt:	08/02/2024
Tác giả:	AIO
Nguồn:	Module 4 AIO 2023
Nguồn dữ liệu (nếu có):	
Từ khóa:	KNN, Machine Learning, Entropy
Người tóm tắt:	Nguyễn Mậu Ánh Ngân

1. **Giới thiệu sơ bộ Machine Learning:**

   - Machine Learning là một dạng máy học từ data, không phải lập trình cứng (Lập trình cứng ví dụ như: Lập trình web, phần mềm,...).
   - Data (hay dữ liệu) thường được gọi là label/groundtruth (thông tin có sẵn) có 2 dạng:
     - Đã được xử lý bởi chuyên gia (Gắn nhãn, phân loại,...) -> supervised learning (Ví dụ: tập dữ liệu bệnh án y khoa đã được bác sĩ phân loại).
     - Chưa được xử lý bởi chuyên gia -> unsupervised learning.

2. **K-Nearest Neighbors:**



Hình 1: KNN

- Xem xét hình 1 ta phân tích đối với bài toán phân loại:
  - Chia data thành 2 nhóm theo thuộc tính nào đó là: Đỏ (nhóm 1) va Xanh (nhóm 2). Đưa vào Data màu đen chưa biết hỏi nó thuộc nhóm nào.
  - Tính khoảng cách d từ điểm đen tới các điểm trong nhóm 1 và 2.
  - Sắp xếp các d theo thứ tự từ nhỏ tới lớn.
  - Chọn K = 3 (K là số ứng viên ta sẽ tham khảo d), lọc ra 3 d có giá trị nhỏ nhất (Gần với màu đen nhất).
  - Bầu chọn theo số đông (Vote) có nhiều d thuộc nhóm nào trong 3 d đó thì màu đen thuộc nhóm đó.
- Xem xét hình 1 ta phân tích với bài toán dự đoán (số thực): Ta sẽ thực hiện các bước như trên nhưng sau khi xác định các d trong khoảng K -> ta tính trung bình d của K đối tượng đó -> được số cần tìm.

Suy ra cách thức chung để dùng phương pháp KNN: Xử lý data (chuẩn hóa) -> chọn K -> tính khoảng cách d giữa data test và data đang có -> xắp xếp d (từ nhỏ tới lớn) -> lấy ra k d -> vote thuộc nhóm nào phổ biến thì đáp án là nhóm đó.

- ***CHÚ Ý:***
  - Công thức để tính khoảng cách d hay dùng:

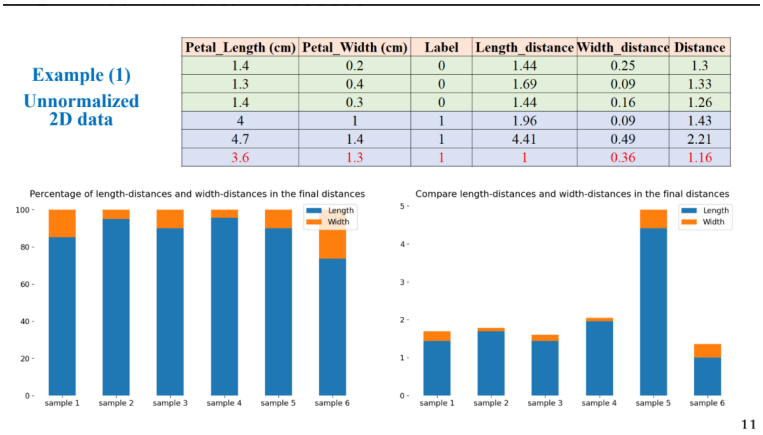
    $\mathrm{d} = \sqrt{\sum_{i}^{x^{train}, x^{test}} \left(x_{\mathrm{i}}^{train} - x_{\mathrm{i}}^{test}\right)^2}$

    Với i là số đặc tính đang xét của data

    Khi i = 1 thì $\mathrm{d} = \left|x^{train} - x^{test}\right|$
  - K là số ứng viên ta sẽ tham khảo để ra quyết định
  - Cố gắng lấy K là số lẻ để tránh tổng số vote bằng nhau
  - Chọn K càng lớn thì tính ổn định của quyết định sẽ cao hơn

3. **Chuẩn hóa data:**

- Ta xét 2 ví dụ data chưa chuẩn hóa



Hình 2: Ví dụ 1

- Nhìn vào 2 cột Length và Width ta thấy: dù cùng đơn vị (Cm) nhưng range (phạm vi) giá trị của 2 cột cách biệt nhau khá lớn. Để hiểu rõ hơn sự chênh lệch ta tính:

– Tính khoảng cách của từng thuộc tính với điểm test (Cho điểm test có Length = 2.6cm và Width = 0.7cm) (CHÚ Ý: 2 CỘT NÀY TRONG HÌNH CHƯA LẤY CĂN)

– Tính tỉ lệ phân trăm giữa khoảng cách vừa tính được của 2 cột:

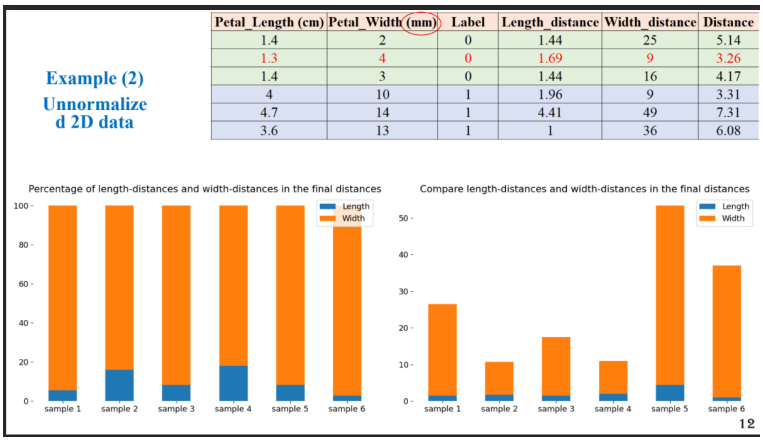
Phần trăm của khoảng cách cột L: $\frac{L}{L+W}$

Phân trăm của khoảng cách cột W: $\frac{W}{L+W}$

– Dùng số liệu tính được ta vẽ được đồ thị thứ nhất của hình 2. Ta thấy rằng: Tổng số ảnh hưởng của Length lên data test mạnh hơn nhiều (Phần màu xanh trội hơn).

– Cộng dồn 2 số khoảng cách ở 2 cột Length Distance và Width Distance ta tiếp tục vẽ đồ thị thứ hai của hình 2. Ta vẫn thấy rằng phần màu xanh cũng trội hơn.

Suy ra: Khi ta đưa data vào để test nó sẽ bị ảnh hưởng nhiều bởi Length (Bias)



Hình 3: Ví dụ 2

- Tiếp tục với tập data trên nhưng lần này ta đổi từ đơn vị cm thành mm. Sau tính toán như trên ta lại thấy phần gây ảnh hưởng nhiều hơn là phần Wedth.

QUA HAI VÍ DỤ TRÊN: Trước khi đưa tập data vào sử dụng train ta phải chuẩn hóa để không gây ra thên vị do độ ảnh hưởng quá cao của một thuộc tính. Công thức chuẩn hóa:

$$\text{d} = \frac{x-\bar{x}}{\sigma}$$

Với x là giá trị data cần chuẩn hóa

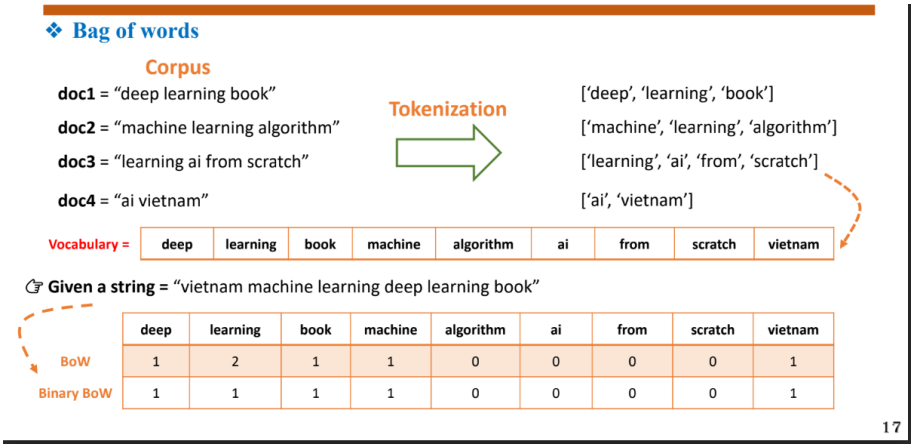
$\bar{x}$ là giá trị trung bình $\bar{x} = \frac{1}{n}\sum_i x_i$ với i là số giá trị của x

σ là standard deviation $\sigma = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$

Hay có thể nói: Chuẩn hóa chính là đưa $\bar{x}$ về 0 và σ về 1. Phù hợp nhất với các dữ liệu thay đổi tuyến tính (Dữ liệu có thể quy đổi đơn vị).

4. **KNN cho dữ liệu Text:**

- Với dữ liệu text, ta phải chuyển thành vector trước khi đưa vào KNN



Hình 4: Phương pháp Bag of Words

- Một trong những phương pháp chuyển đổi đơn giản nhất là BoW (Bag of Words). Ta cùng xét ví dụ như hinh 4:
  - Cho data gồm 4 doc như hình.
  - Tạo dictionary gồm những từ tách ra từ doc.
  - Tạo Volcabulary có độ dài bằng tất cả các từ (mà không trùng nhau). Như hình ta có độ dài là 9 ta lấy đây là độ dài chuẩn của vector chuyển đổi.
  - Tiếp theo, khi đưa bất khì một đoạn text nào để phân loại hay nhận biết ta sẽ gán cho nó một vector có độ dài là 9 với tất cả các phần tử là 0.
  - So sánh vector mới tạo với vector Vocabulary, từ nào có xuất hiện và xuất hiện bao nhiêu lần thì đổi số 0 ở chổ đó thành 1/2/3...
  - Có được vector ta đưa vào KNN.

5. **Entropy:**

- Là một khái niệm được dùng trong Decision Tree.
- Ta cùng xem xét ví dụ để hiểu rõ hơn:
  - Cho một thùng có 9 bóng đỏ và 1 bóng xanh.
  - Sự kiện A là lấy được bóng đỏ -> Xác suất của A (Khả năng A xảy ra) là P(A) = 0.9
  - Sự kiện B là lấy được bóng xanh -> Xác suất của B là P(B) = 0.1
  - Ta quan sát người bốc, ta sẽ thấy mức độ ngạc nhiên khi bốc được bóng xanh là nhiều hơn (Vì khả năng xảy ra thấp hơn) -> Tỉ lệ nghịch với P
  - Gọi S(E) là mức độ ngạc nhiên:
$$S(E) = \frac{1}{P(E))}$$
- Khi P(E) = 1 thì S(E) = 1: có nghĩa là xấc suất bốc được là 1 thì mức độ ngạc nhiên cũng là 1 hay không ngạc nhiên -> Hơi vô lý vì không ngạc nhiên nhưng lại là 1 -> Ta cần chuẩn hóa S(E) để hợp lý hơn. SUy ra:
$$S(E) = log\frac{1}{P(E)} \in [0, \infty]$$
  P(
- Expected Value: Xét ví dụ: Ném đồng xu 100 lần
  - Gọi H là xấc suất xuất hiện mặt ngửa P(H) = 0.5 với 1 là chuyển sang dạng số (Numerize) của H. -> Số lần mà H xảy ra: 0.5*100

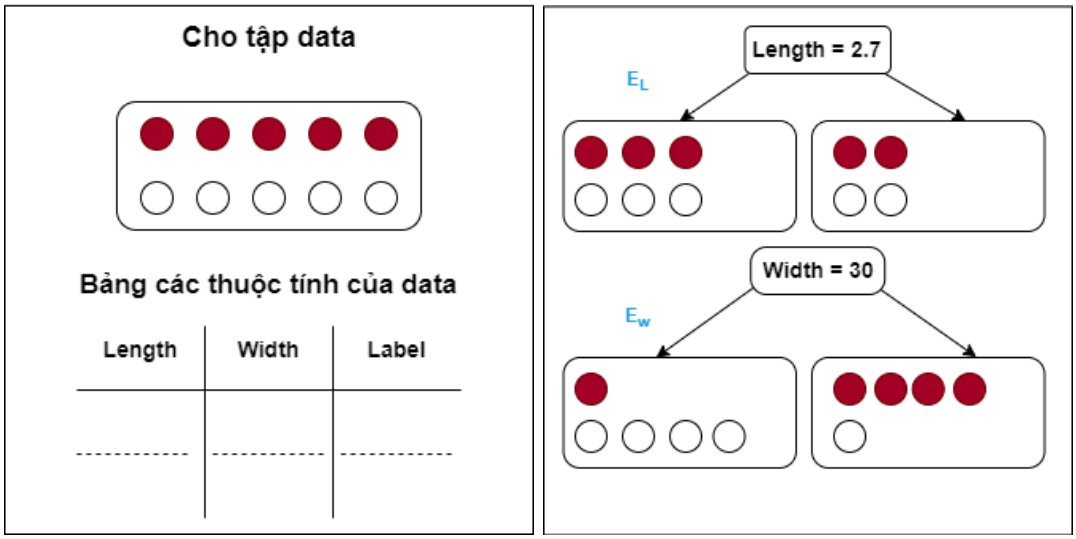
- Gọi T là xắc suất xuất hiện mặt sấp P(T) = 0.5 với 0 là chuyển sang dạng số của T -> Số lần mà T xảy ra: 0.5*100
- Tổng số giá trị = Số lượng H * 1 + Số lượng T * 0 = 0.5*100 + 0.5*100 = 50 (sau 100 lần ném)
- Giá trị trung bình của 1 lần ném: $A = \frac{50}{100} = 0.5$
- Tổng quát: Nếu ném n lần: $A(n) = \frac{0.5*n*1+0.5*n*0}{n} = 0.5 * 1 + 0.5 * 0 = 0.5$
- Gọi X là = 0 hoặc 1, A(n) là E(S) trung bình suprise: $E(S) = X_1 * P(X_1) + X_2 * P(X_2) = \sum_i X_i * P(X_i)$
- Cho rằng X là một hàm g(X). Khi đó g(X) nhận giá trị của S(E): $E(S) = log\frac{1}{P(x)}$

Vậy giá trị trung bình suprise là:

$$E(S) = \sum P(X) * log\frac{1}{P(X)} = -\sum P(X) * logP(X)$$

Nhận định: Công thức Entropy đo mức độ trung bình của Suprise (Hay mức độ không ổn định của tập data).

- Ứng dụng: Xét ví dụ sau



Hình 5: Ứng dụng Entropy

- Chọn điều kiện L = 2.7, ta tách được tập data làm 2 nhánh như hình 6 và tính E_L cho 1 nhánh.
- Chọn điều kiện W = 30, ta tách được tập data làm 2 nhánh như hình 6 và tính E_W cho 1 nhánh.
- Sau khi tính toán theo những gi ta đã thảo luận bên trên, nhin vào hình minh họa ta thấy: $E_L > E_W$ -> mức độ Suprise của điều kiện L nhiều hơn -> Ta phải chọn điều kiện nào có mức độ Suprise thấp để ổn định tập data -> Chọn bên có E nhỏ hơn.

Hay nói cách khác, chọn cách phân loại nào có Entropy bé để dễ lọc dữ liệu và đưa ra quyết định.