

4 bước xây dựng hệ thống AI

1. Chuẩn bị dữ liệu

- Là bước quan trọng và dành nhiều thời gian nhất.
- Dữ liệu tốt
 - + Số lượng nhiều
 - + Có sự nhất quán (phải qua được bước tiền xử lý)
- Bước chuẩn bị gồm :
 - + Thu thập dữ liệu
 - + Làm sạch dữ liệu
 - + Gán nhãn
 - + Tăng cường tập dữ liệu

2. Mô hình hóa AI

- Lựa chọn mô hình. + ML : regression, SVM, tree.....
 - + DL : GAN,CNN,LSTM,.....
- Xác định các tham số tối ưu (thông qua vòng lặp kiểm tra , kiểm tra kết quả , từ đó sửa đổi và cải tiến mô hình .
- Để có được mô hình chính xác , cần nhiều thời gian => bổ xung phần cứng => tăng tốc đáng kể thời gian đào tạo mô hình

3. Thiết kế hệ thống .

- Thêm AI vào hệ thống và đảm bảo AI phải tồn tại với các thành phần khác 1 cách mượt mà
- Mô phỏng : để xác minh hệ thống hoạt động đúng trong tất cả các trường hợp.

4. Triển khai

Vai trò của cơ sở dữ liệu

- Dataset giúp ta huấn luyện mô hình 1 cách chính xác vì ML phụ thuộc rất nhiều vào dữ liệu
- Dataset phải được định dạng có trật tự để bổ xung thông tin có ích cho việc huấn luyện
- Dataset dùng để đánh giá mô hình

4 Hướng tiếp cận AI

1. Suy nghĩ như con người

- Cần xác định cách con người suy nghĩ , cần đi vào bên trong hoạt động thực tế của trí óc con người

- Thông qua việc xem xét nội tâm : có nắm bắt suy nghĩ của chính chúng ta
 - Thông qua các thí nghiệm tâm lý
2. Hành động như con người
 - Được đề xuất bởi Alan Turing (1950)
 - Máy tính được thăm vấn bởi con người từ xa và vượt qua bài kiểm tra khi người thăm vấn không biết liệu có máy tính hoặc con người ở đầu bên kia
 3. Suy nghĩ đúng đắn và hợp lý
Cơ sở đúng + cấu trúc lập luận đúng => kết luận đúng
 4. Hành động đúng đắn hợp lý

FEATURE ENGINEERING

1. ONE HIT ENCODING

- Biến đổi các giá trị thành các đặc trưng nhị phân chỉ chứa giá trị 1 hoặc 0. Mỗi mẫu trong đặc trưng phân loại sẽ được biến đổi thành 1 vector chỉ với một trong các giá trị là 1 tương ứng với vị trí của đặc trưng đó
Vd : màu sắc có 3 giá trị : “đỏ”, “vàng”, “Xanh” tương ứng 1 2 3
 $\Rightarrow \text{Đỏ} = [100]$
 $\Rightarrow \text{Vàng} = [010]$
 $\Rightarrow \text{Xanh} = [001]$

2. BUNNING

- Được sử dụng để biến đổi các đặc trưng số liên tục thành dạng các đặc trưng phân loại
Phân biệt . Các giá trị hoặc số rời rạc này được gọi là danh mục (bin). Trong đó các giá trị số thô , liên tục sẽ được nhóm lại
 - Vd: đặc trưng “tuổi”
 - Có thể nhóm lại thành các bin
 - Bin1: 0 -> 5 tuổi
 - Bin2 : 6 -> 10 tuổi
 - Bin 3 : 11-15 tuổi

3. NORMALIZATION – CHUẨN HÓA

- Chuyển đổi phạm vi giá trị thực của 1 đặc trưng số sang 1 dải giá trị tiêu chuẩn , thường trong khoảng $[-1,1]$ hoặc $[0,1]$
- Tính bằng công thức $y = (x - \min) / (\max - \min)$
- Y là biến sau khi normalization , x là biến trước
- Max , min là giá trị LN và NN của feature trong tập dữ liệu

VD: ta cần normalize giá trị 350 trong khoản [350,1450]

Ta có : $x=350$

$$\Rightarrow y = \frac{(350-350)}{(1450-350)} = 0 \text{ nằm trong khoảng } [0,1]$$

Hoặc $x = 1100$

$$\Rightarrow y = \frac{(1100-350)}{(1450-350)} = 0.68 \text{ nằm trong khoảng } [0,1]$$

Nếu 1 điểm dữ liệu nào đó nằm ngoài khoảng [min , max] , ta có thể bỏ khỏi tập dữ liệu.

4. STANDARDLUZATION

Thay đổi tỷ lệ các giá trị đặc trưng để của phân phối chuẩn với $u=0$ và $O = 1$ (lần lượt là kỳ vọng 1 phương sai)

Teaplate mactching : so khớp mẫu

Tính $s1 \rightarrow s7$

Vd : chứng minh ảnh 1 là chữ K

ảnh 2 là chữ S

\Rightarrow ảnh 5 là chữ gì ??

Ta tính : $d(5,1) = d1$ (khoảng cách từ ảnh 1 \rightarrow ảnh 5)

$d(5,2) = d2$ (khoảng cách từ ảnh 2 \rightarrow ảnh 5)

nếu $d1 < d2$ thì ảnh 5 giống chữ K và ngược lại

cách tính :

$d(5,1) =$

$$\sqrt{(S_{1 \text{ Ảnh 1}} - S_{1 \text{ Ảnh 5}})^2 + (S_{2 \text{ Ảnh 1}} - S_{2 \text{ Ảnh 5}})^2 + \dots + (S_{7 \text{ Ảnh 5}} - S_{7 \text{ Ảnh 5}})^2}$$

tương tự $d(5,2)$

$d2 < d1 \Rightarrow$ ảnh 5 là chữ S

Các công thức tính Hu Moment

1. Diện tích $S = M = \sum_X \sum_y S(x, y) = \text{moo}$

2. Tìm trọng tâm $x = \frac{\sum_X \sum_y X.S(x,y)}{\sum_X \sum_y S(x,y)}$, $y = \frac{\sum_X \sum_y Y.S(x,y)}{\sum_X \sum_y S(x,y)}$

3. Central normalized moments $M_{pq} = \frac{m_{pq}}{\text{moo}^{\frac{p+q}{2}+1}}$

4. Central moments

$$m_{pq} = \sum_X \sum_y S(x - x_{ngang})^p \cdot (y - y_{ngang})^q \cdot S(x, y), \text{ pq} = 0, 1, 2, 3, \dots$$

5. Hu moment feature

- $S1 = M_{20} + M_{02}$
- $S2 = (M_{20} - M_{02}) \cdot (M_{20} + M_{02}) + 4 M_{11} M_{11}$
- $S3 = (M_{30} - 3M_{12})^2 + (M_{30} - 3M_{21})^2$
- $S4 = (M_{30} + M_{12})^2 + (M_{03} + M_{21})^2$
- $S5 =$
- $S6 =$
- $S7 =$

1. AUC là gì ?

Là 1 số liệu phổ biến được sử dụng trong ML cung cấp 1 thước đo tổng hợp cho hiệu suất phân loại.

2. Bạn đánh giá thế nào về việc AI thay thế bsĩ để chuẩn đoán bao nhiêu phần trăm , có thích hợp với nền y tế VN không ?

- AI chỉ là công cụ hỗ trợ các bác sĩ chuẩn đoán ung thư vú chứ không phải thay thế
- Vn, hay các nước đang phát triển đều đang đẩy mạnh xu hướng VN 4.0 , ứng dụng CN vào các lĩnh vực là rất cần thiết . Tuy nhiên , khả năng ứng dụng AI chỉ đáp ứng 1 phần nhu cầu không thể thay thế các bác sĩ .

3. Nhược điểm của AI trong chuẩn đoán ung thư phổi ?

- Nếu kích thước nhỏ hơn các kích thước mẫu trong nghiên cứu thì khó có thể phát hiện ra được . Để AI có thể phát hiện và chuẩn đoán tốt thì cần phải có bộ dữ liệu phong phú , tuy nhiên dữ liệu hiện nay đang bị hạn chế.

4. AI phát hiện ung thư phổi có thể phát hiện ung thư khác không ?

- Nếu có phong phú về dữ liệu và các đặc trưng thì AI có thể sử dụng để chuẩn đoán các loại ung thư khác.

5. Chi phí có phù hợp để ứng dụng đại trà ?

- Hiện nay việc chuẩn đoán ung thư bằng AI ở nước ta vẫn còn khá mới , chỉ dừng lại ở mức độ nghiên cứu . Để có thể ứng dụng đại trà thì chúng ta cần thời gian , Về giai đoạn đầu thì có thể chi phí cao hơn so với chuẩn đoán bình thường.

6. Ở Việt Nam , việc đưa vào ứng dụng vào trong thực tế, hệ thống phần mềm AI đã đạt kết quả như thế nào ?

- Nghiên cứu và ứng dụng AI ở VN ở giai đoạn đầu . 1 số ứng dụng AI hỗ trợ chuẩn đoán do cá công ty quốc tế phát triển đã được triển khai tại 1 số bệnh viện . Tuy nhiên, dữ liệu người Việt có tính đặc thù, điều này làm giảm hiệu quả của các thuật toán AI vốn được huấn luyện trên dữ liệu.