BTVN 1

Bài tập về nhà Học máy - tuần 1

Nộp file word trên OCW

Bài 1

Dựa trên định nghĩa về học máy của Tom Mitchell, hãy xác định các thành phần PET (P: Performance measure, E: Experience, T: Task) cho các bài toán học sau:

a/ Một hệ thống nhận dạng người đi bộ đi qua đường

b/ Hệ thống điều chỉnh nhiệt độ trong một tòa nhà thông minh

c/ Robot hút bụi tự động di chuyển trong nhà

a/ Một hệ thống nhận dạng người đi bộ đi qua đường

- **Task (T):** nhận dạng người đi bộ qua đường
- **Experienca (E):** data bao gồm hình ảnh và video của người đi bộ qua đường, người không qua đường, và phương tiện khác (xe đạp, xe đẩy,...)
- **Performance measure (P):** Tỷ lệ chính xác của việc nhận dạng người đi bộ, bao gồm các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu, và F1-score.

  ▼ **Độ chính xác (Accuracy)**

  Tỉ lệ % các dự đoán đúng trên tổng số dự đoán (bao gồm cả người băng đường và không qua đường)

  Công thức:

  $$\text{Độ chính xác} = \frac{TP + TN}{TP + TN + FP + FN}$$

  Trong đó

  - **TP (True Positive)**: Số trường hợp người đi bộ qua đường được dự đoán đúng.

- **TN (True Negative)**: Số trường hợp không có người đi bộ hoặc người đi bộ không qua đường được dự đoán đúng.
- **FP (False Positive)**: Số trường hợp không có người đi bộ hoặc người đi bộ không qua đường bị dự đoán nhầm là có người đi bộ qua đường.
- **FN (False Negative)**: Số trường hợp có người đi bộ qua đường nhưng bị dự đoán nhầm là không có.

▼ Độ nhạy (Sensitivity, Recall)

Tỷ lệ % các trường hợp có người đi bộ qua đường và được hệ thống nhận dạng đúng.

Công thức

$$\text{Độ nhạy} = \frac{TP}{TP + FN}$$

Ví dụ: Nếu có 100 trường hợp người đi bộ qua đường và hệ thống nhận dạng đúng 90 trường hợp, thì độ nhạy là 90%.

▼ Độ đặc hiệu (Specificity)

Định nghĩa: Độ đặc hiệu là tỷ lệ phần trăm các trường hợp không có người đi bộ hoặc người đi bộ không qua đường được hệ thống nhận dạng đúng.

Công thức:

$$\text{Độ đặc hiệu} = \frac{TN}{TN + FP}$$

Ví dụ: Nếu có 100 trường hợp không có người đi bộ hoặc người đi bộ không qua đường và hệ thống nhận dạng đúng 95 trường hợp, thì độ đặc hiệu là 95%.

▼ F1-score

Là một chỉ số kết hợp giữa độ nhạy và độ chính xác để đánh giá sự cân bằng giữa việc nhận dạng đúng người đi bộ và việc không bỏ sót.

Công thức

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong đó,
Precision là độ chính xác của các dự đoán dương tính, tức là tỷ lệ phần trăm các dự đoán người đi bộ qua đường đúng trong tổng số các dự đoán dương tính (bao gồm cả dự đoán đúng và sai).

$$\text{Precision} = \frac{TP}{TP + FP}$$

Giả sử: Nếu hệ thống có Precision là 80% và Recall là 90%, thì F1-score sẽ là:

$$\text{F1-score} = \frac{2 \times 0.8 \times 0.9}{0.8 + 0.9} \approx 0.85$$

▼ Bài toán tham khảo

Giả sử bạn có một bộ dữ liệu với 1000 hình ảnh/video, trong đó:

- 400 hình ảnh/video có người đi bộ qua đường (positive cases).
- 600 hình ảnh/video không có người đi bộ qua đường hoặc người đi bộ không qua đường (negative cases).

Hệ thống dự đoán:

- 350 hình ảnh/video đúng có người đi bộ qua đường (TP = 350).
- 50 hình ảnh/video có người đi bộ qua đường nhưng bị nhận diện sai (FN = 50).
- 500 hình ảnh/video đúng không có người đi bộ hoặc người đi bộ không qua đường (TN = 500).

- 100 hình ảnh/video không có người đi bộ hoặc người đi bộ không qua đường nhưng bị nhận diện sai là có người đi bộ qua đường (FP = 100).

Từ đó, ta có thể tính các chỉ số như sau:

- Độ chính xác:

$$\text{Accuracy} = \frac{350 + 500}{1000} = \frac{850}{1000} = 85\%$$

- Độ nhạy:

$$\text{Recall} = \frac{350}{350 + 50} = \frac{350}{400} = 87.5\%$$

- Độ đặc hiệu:

$$\text{Specificity} = \frac{500}{500 + 100} = \frac{500}{600} = 83.3\%$$

- F1-score:
  - Precision:

$$\text{Precision} = \frac{350}{350 + 100} = \frac{350}{450} = 77.8\%$$

  - F1-score:

$$\text{F1-score} = \frac{2 \times 0.778 \times 0.875}{0.778 + 0.875} \approx 0.824$$

Vậy, hệ thống có thể được đánh giá dựa trên các chỉ số này để xem xét hiệu quả của việc nhận diện người đi bộ đi qua đường.

b/ Hệ thống điều chỉnh nhiệt độ trong một tòa nhà thông minh

- **Task (T):** Điều chỉnh nhiệt độ trong tòa nhà để duy trì mức nhiệt độ mong muốn.

- **Experience (E):** Dữ liệu lịch sử về nhiệt độ trong tòa nhà, cùng với dữ liệu về các yếu tố ảnh hưởng như thời tiết bên ngoài, số lượng người trong tòa nhà, và các cài đặt nhiệt độ mong muốn trước đó.
- **Performance measure (P):** Mức độ chênh lệch giữa nhiệt độ thực tế và nhiệt độ mong muốn (ví dụ: sai số bình phương trung bình), và mức tiêu thụ năng lượng.

c/ Robot hút bụi tự động di chuyển trong nhà

- **Task (T):** Di chuyển trong nhà và làm sạch bụi bẩn.
- **Experience (E):** Dữ liệu từ các cảm biến về bản đồ nhà, các vị trí có bụi bẩn, và các hoạt động làm sạch trước đó.
- **Performance measure (P):** Diện tích sàn nhà được làm sạch, thời gian để hoàn thành công việc, và mức độ tiêu thụ năng lượng.

Bài 2

Trong một nhà máy sản xuất bánh, trên băng chuyền có ba loại bánh với hành dạng khác nhau lần lượt xuất hiện. Để thực hiện việc gắp bánh và xếp riêng vào ba thùng, một sinh viên được giao nhiệm vụ thiết kế một tay máy có ứng dụng học máy.

Áp dụng cách phân loại khi xem xét đến số lượng và loại giám sát trong quá trình huấn luyện, sinh viên đó sẽ thực hiện nhiệm vụ được giao này dựa trên loại học máy nào? Giải thích lý do.

Lý do chọn loại máy học:

- **Dữ liệu gán nhãn:** Trong bài toán này, cần có dữ liệu gán nhãn rõ ràng về loại bánh, điều này phù hợp với học máy có giám sát.
- **Mục tiêu phân loại:** Nhiệm vụ là phân loại ba loại bánh khác nhau, điều này hoàn toàn phù hợp với học máy có giám sát, nơi mà mô hình học từ các ví dụ đã gán nhãn để phân loại các đối tượng mới.

▼ Tại sao không chọn các loại học máy khác?

a. Đặc điểm của **học không giám sát** sẽ không sử dụng dữ liệu dán nhãn. Trong tình huống này, sinh viên cần phải phân loại bánh theo các nhãn cụ thể.

b. Đặc điểm của **học bán giám sát** là kết hợp lượng nhỏ data có gán nhãn với lượng lớn data không có nhãn. Học bán giám sát thường được sử dụng khi việc thu thập data có gán nhãn khó khăn, hoặc tốn kém. Chỉ cần phân 3 loại bánh dựa trên hình dạng, có khả năng thu thập và chuẩn bị trc data. Vì vậy, không cần sử dụng học bán dám sát khi nhãn đã có đầy đủ.

c. Đặc điểm của **học tăng cường** d là học dựa trên quy tắc thưởng và phạt. Mô hình học cách hành động trong 1 môi trường để tối ưu hóa tổng phần thưởng theo tgian. Đối với tình huống trên, học tăng cường thì phù hợp hơn với các môi trường động và phức tạp, còn nhiệm vụ chỉ xác định và phân loại dựa trên hình dạng thì không cần thiết. Học giám sát sẽ dễ dàng hơn.

Để thiết kế một tay máy:

1. **Xác định nhiệm vụ (Task):** Nhiệm vụ của tay máy là nhận dạng và phân loại ba loại bánh khác nhau dựa trên hình dạng của chúng, sau đó gắp và xếp chúng vào các thùng tương ứng.
2. **Dữ liệu huấn luyện (Training Data):** Để huấn luyện, sinh viên cần chuẩn bị tập dữ liệu đã được gán nhãn, bao gồm nhiều hình ảnh của từng loại bánh với các nhãn tương ứng (loại 1, loại 2, loại 3). Trong đó, mỗi hình ảnh sẽ chứa thông tin về từng loại bánh.
3. **Phân loại có giám sát (Supervised Learning):** Với tập dữ liệu gán nhãn này, sinh viên sẽ sử dụng một thuật toán học máy có giám sát để huấn luyện mô hình. Các bước cơ bản:

    a. Thu thập dữ liệu: bao gồm hình ảnh của các loại bánh và gán nhãn chúng.

    b. Tiền xử lý dữ liệu: Xử lý hình ảnh trước để chuẩn bị cho huấn luyện

    c. Huấn luyện mô hình: Sử dụng các thuật toán học máy để huấn luyện mô hình nhận dạng bánh

d. Đánh giá mô hình: Sử dụng một tập dữ liệu kiểm tra để đánh giá hiệu suất của mô hình.

Bài 3:

Sau bước đầu làm quen với việc xây dựng hệ thống học máy, theo quan điểm của em bước khó khăn và thách thức nhất là bước nào? Vì sao?

Theo em, bước thu thập và chuẩn bị một data tốt là bước khó khăn và thách thức nhất. Ở bước này, ta cần phải thu thập, làm sạch dữ liệu, biến đổi thô để có thể sử dụng được cho các mô hình học máy.

Bởi vì:

- Chất lượng dữ liệu kém, chứa lỗi, thiếu hoặc sai dữ liệu, thiếu sự nhất quán
- Khối lượng dữ liệu lớn
- Định nghĩa và gán nhãn dữ liệu

Bước chuẩn bị và làm sạch data có thể quyết định thành bại của mô hình học máy. Nếu dữ liệu không sạch hoặc không xử lý đúng cách, dù máy có mạnh cỡ nào hay phức tập đến bao nhiêu cũng có thể đưa ra các kết quả không chính xác.

Bài 4:

Phương pháp trích chọn đặc trưng của dữ liệu đóng vai trò quan trọng trong việc xây dựng và giải quyết vấn đề học máy. Phân tích thành phần chính, hay PCA (Principal component analysis), là phương pháp giảm chiều thường được sử dụng để giảm chiều của các tập dữ liệu lớn, bằng cách chuyển đổi một tập hợp biến lớn thành tập hợp nhỏ hơn nhưng vẫn chứa hầu hết thông tin trong tập hợp lớn đó. Thực hiện việc tìm kiếm và tra cứu tài liệu chuyên môn. Sau đó ghi tóm tắt từ 3 đến 5 điểm đáng chú ý của phương pháp này.

Tóm tắt các điểm đáng chú ý của phương pháp PCA:

1. **Giảm chiều dữ liệu**: PCA là kỹ thuật giúp giảm số lượng biến trong dữ liệu mà vẫn giữ được các thông tin quan trọng.

→ Làm giảm độ phức tạp của mô hình và tránh tình trạng overfitting (đọc thêm: https://machinelearningcoban.com/2017/03/04/overfitting/)

▼ *Overfitting là gì?*

Việc này giống như khi bạn ôn thi.

Giả sử bạn không biết đề thi như thế nào nhưng có 10 bộ đề thi từ các năm trước. Để xem trình độ của mình trước khi thi thế nào, có một cách là bỏ riêng một bộ đề ra, không ôn tập gì. Việc ôn tập sẽ được thực hiện dựa trên 9 bộ còn lại. Sau khi ôn tập xong, bạn bỏ bộ đề đã để riêng ra làm thử và kiểm tra kết quả, như thế mới "khách quan", mới giống như thi thật. 10 bộ đề ở các năm trước là "toàn bộ" training set bạn có. Để tránh việc học lệch, học tủ theo chỉ 10 bộ, bạn tách 9 bộ ra làm training set thật, bộ còn lại là validation test. Khi làm như thế thì mới đánh giá được việc bạn học đã tốt thật hay chưa, hay chỉ là *học tủ*.

Vì vậy, *Overfitting* còn có thể so sánh với việc *Học tủ* của con người.

2. **Biến đổi dữ liệu**: PCA thực hiện việc biến đổi dữ liệu gốc thành một tập hợp các thành phần chính (principal components), là các tổ hợp tuyến tính của các biến ban đầu. Các thành phần này được sắp xếp theo thứ tự giảm dần của lượng thông tin mà chúng mang lại (variance).

3. **Giữ thông tin tối đa**: Các thành phần chính được chọn để đảm bảo rằng sự biến thiên của dữ liệu được giữ ở mức cao nhất có thể, do đó thông tin gốc không bị mất đi quá nhiều khi giảm chiều.
4. **Giảm nhiễu**: PCA giúp loại bỏ nhiễu trong dữ liệu bằng cách tập trung vào các thành phần chính có mức độ biến thiên cao, trong khi bỏ qua các thành phần có biến thiên thấp có thể chứa nhiều nhiễu.
5. **Tiền xử lý dữ liệu**: Để PCA hoạt động hiệu quả, dữ liệu cần được chuẩn hóa hoặc chuẩn chỉnh (scaling), đặc biệt là khi các biến có đơn vị đo lường khác nhau.