Machine learning

Mục lục

1. Bản chất của Học Máy

Học máy hướng đến việc tạo ra các hệ thống có khả năng tự học từ dữ liệu, cải thiện hiệu suất khi giải quyết một nhiệm vụ cụ thể hoặc một lớp nhiệm vụ tương tự.

Nói cách khác, máy móc được "huấn luyện" từ dữ liệu để tự động rút ra các quy luật, mẫu (patterns), hoặc kiến thức ẩn chứa trong dữ liệu đó, từ đó có thể áp dụng kiến thức đã học vào việc dự đoán, phân loại, hoặc đưa ra quyết định cho các dữ liệu mới mà chưa từng gặp trước đây.

Ví dụ:

Hệ thống lọc thư rác:

Hệ thống được huấn luyện từ tập email đã được gán nhãn "rác" hoặc "không rác". Qua quá trình học, hệ thống tự động rút ra các đặc trưng của email rác, từ đó có thể tự động phân loại các email mới.

Hệ thống nhận diện khuôn mặt:

Hệ thống học từ tập ảnh khuôn mặt, mỗi ảnh được gán nhãn tên người tương ứng. Dựa trên dữ liệu huấn luyện, hệ thống tự động rút ra các đặc trưng của mỗi khuôn mặt, từ đó có thể nhận diện người trong các ảnh mới.

Tầm quan trọng của học máy đóng vai trò then chốt trong việc khai thác và tận dụng lượng dữ liệu khổng lồ đang được tạo ra mỗi ngày. Thay vì lập trình thủ công các quy luật phức tạp, học máy cho phép máy tính tự động học hỏi từ dữ liệu, giúp giải quyết các bài toán khó, phức tạp mà con người khó có thể giải quyết bằng phương pháp truyền thống.

2. Quy trình xây dựng hệ thống học

Để xây dựng một hệ thống học máy hiệu quả, ta cần tuân theo một quy trình có hệ thống, bao gồm các bước chính sau:

Bước 1: Hiểu nhu cầu

● Xác định rõ mục tiêu của việc xây dựng hệ thống.

● Khảo sát lĩnh vực ứng dụng, bài toán cần giải quyết.

● Ví dụ: Xây dựng hệ thống dự đoán rủi ro tín dụng cho khách hàng vay vốn.

Bước 2: Lựa chọn phương pháp phân tích

● Xác định loại bài toán học máy phù hợp (phân loại, hồi quy, phân cụm, ...).

● Nghiên cứu các lớp mô hình và thuật toán học tiềm năng.

● Ví dụ: Lựa chọn mô hình hồi quy logistic (logistic regression) cho bài toán dự đoán rủi ro tín dụng.

Bước 3: Thu thập dữ liệu

● Xác định nguồn dữ liệu, loại dữ liệu, kích thước dữ liệu.

● Thiết kế phương pháp thu thập (khảo sát, thu thập tự động, ...).

● Ví dụ: Thu thập dữ liệu khách hàng từ hệ thống quản lý khách hàng của ngân hàng, bao gồm thông tin cá nhân, lịch sử giao dịch, ...

1 đoạn code đơn giản về thu thập dữ liệu

```
import requests
from bs4 import BeautifulSoup

// URL của trang web cần thu thập dữ liệu

url = "https://example.com"

// Gửi yêu cầu HTTP GET đến URL

response = requests.get(url)

// Phân tích cú pháp HTML của trang web bằng BeautifulSoup

soup = BeautifulSoup(response.content, 'html.parser')

// Lấy tiêu đề của trang web

title = soup.title.text

// Lấy tất cả các liên kết trên trang web

links = [link.get('href') for link in soup.find_all('a')]

// In tiêu đề và các liên kết

print(f"Tiêu đề: {title}")
print(f"Liên kết: {links}")
```

```
Giải thích:
● Thư viện:
○ requests: Gửi yêu cầu HTTP và lấy nội dung trang web.
```

```
○ BeautifulSoup: Phân tích cú pháp HTML, giúp trích xuất dữ liệu dễ dàng.

● Các bước:
○ Gửi yêu cầu GET đến URL và lấy nội dung HTML.
○ Tạo đối tượng BeautifulSoup để phân tích HTML.
○ Sử dụng các phương thức của BeautifulSoup để trích xuất dữ liệu mong muốn (ví
dụ: tiêu đề, liên kết, ...).
```

Bước 4: Tiền xử lý dữ liệu

● Làm sạch dữ liệu: Xử lý dữ liệu thiếu, nhiễu, không nhất quán.

○ Bỏ qua mẫu dữ liệu có quá nhiều giá trị thiếu.

○ Điền giá trị thiếu bằng tay, gán giá trị trung bình, hoặc dự đoán dựa trên các mẫu khác.

○ Chuẩn hóa dữ liệu về cùng một đơn vị đo lường, định dạng.

● Tích hợp dữ liệu: Kết hợp dữ liệu từ nhiều nguồn.

○ Xử lý vấn đề định danh, liên kết dữ liệu từ các nguồn khác nhau.

○ Giải quyết xung đột, không nhất quán giữa các nguồn dữ liệu.

● Biến đổi dữ liệu: Chuyển đổi dữ liệu sang dạng phù hợp cho khai phá.

○ Làm mịn dữ liệu bằng kỹ thuật phân cụm, hồi quy.

○ Xây dựng thuộc tính mới từ các thuộc tính có sẵn.

○ Tổng hợp dữ liệu, giảm số chiều dữ liệu.

● Lưu ý: Bước tiền xử lý có thể chiếm đến 70-90% thời gian và công sức trong toàn bộ quá trình.

● Lợi ích: Giúp cải thiện chất lượng dữ liệu, tăng độ chính xác của mô hình, giảm thời gian huấn luyện.

Một đoạn code đơn giản về tiền xử lý tín hiệu chạy như sau:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler

// Nạp dữ liệu từ file CSV

data = pd.read_csv("data.csv")

// Xử lý dữ liệu thiếu
// Thay thế giá trị thiếu bằng giá trị trung bình của cột

data.fillna(data.mean(), inplace=True)

// Loại bỏ nhiễu
// Loại bỏ các hàng có giá trị ngoại lệ
```

```
data = data[data["Age"] < 100]

// Chuẩn hóa dữ liệu
// Sử dụng StandardScaler để đưa dữ liệu về cùng một phạm vi giá trị

scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
```

```
Giải thích:

● Thư viện:
○ pandas: thao tác dữ liệu dạng bảng.
○ sklearn.preprocessing: cung cấp các công cụ tiền xử lý dữ liệu.

● Các bước:
○ Nạp dữ liệu: pd.read_csv để nạp dữ liệu từ file CSV.
○ Xử lý dữ liệu thiếu: Thay thế giá trị thiếu bằng giá trị trung bình của cột bằng
data.fillna(data.mean()).
○ Loại bỏ nhiễu: Loại bỏ các hàng có giá trị ngoại lệ dựa trên điều kiện
data["Age"] < 100.
○ Chuẩn hóa dữ liệu: Sử dụng StandardScaler để chuẩn hóa dữ liệu, đưa dữ liệu về
cùng một phạm vi giá trị.
```

Bước 5: Mô hình hóa

● Lựa chọn lớp mô hình phù hợp dựa trên bài toán và dữ liệu.

● Lựa chọn thuật toán học phù hợp với lớp mô hình.

● Huấn luyện mô hình trên tập huấn luyện.

● Ví dụ: Huấn luyện mô hình hồi quy logistic bằng thuật toán Gradient Descent.

Bước 6: Đánh giá

● Đánh giá hiệu quả của mô hình trên tập kiểm thử (dữ liệu chưa được sử dụng trong quá trình huấn luyện).

● Lựa chọn tiêu chí đánh giá phù hợp (độ chính xác, ma trận nhầm lẫn, MSE, ...).

● Điều chỉnh siêu tham số của mô hình (tốc độ học, số lượng láng giềng, ...) để tối ưu hiệu quả.

● Ví dụ: Đánh giá mô hình hồi quy logistic bằng độ đo AUC (Area Under the ROC Curve).

Bước 7: Triển khai

● Tích hợp mô hình đã huấn luyện vào hệ thống thực tế.

● Ví dụ: Triển khai mô hình dự đoán rủi ro tín dụng vào hệ thống xét duyệt khoản vay.

Bước 8: Phản hồi

● Theo dõi hiệu quả của mô hình trong thực tế.

● Cập nhật mô hình khi cần thiết (dữ liệu mới, thay đổi nhu cầu, ...).

● Ví dụ: Theo dõi tỷ lệ khách hàng vỡ nợ, cập nhật mô hình dự đoán rủi ro tín dụng định kỳ.

3. Các vấn đề quan trọng trong Machine learning

a) Thuật toán học

● Vai trò: Thuật toán học chịu trách nhiệm tìm ra hàm f thuộc lớp mô hình H sao cho f xấp xỉ tốt hàm mục tiêu y.

● Yêu cầu:

○ Hội tụ: Thuật toán cần đảm bảo tìm ra được nghiệm (hàm f) sau một số hữu hạn bước lặp.

○ Hiệu quả: Thuật toán cần sử dụng tài nguyên tính toán (thời gian, bộ nhớ) một cách hợp lý.

○ Phù hợp: Thuật toán cần phù hợp với lớp mô hình và bài toán đang giải quyết.

● "No Free Lunch Theorem": Không có thuật toán nào tốt nhất cho mọi bài toán.

b) Tập huấn luyện

● Vai trò: Tập huấn luyện cung cấp "kinh nghiệm" cho máy học, quyết định chất lượng của mô hình.

● Yêu cầu:

○ Tính đại diện: Tập huấn luyện cần đại diện cho toàn bộ không gian dữ liệu, tránh bị lệch (bias).

○ Kích thước: Tập huấn luyện cần đủ lớn để mô hình có thể học được các mẫu phức tạp.

○ Ít nhiễu: Tập huấn luyện cần được làm sạch, loại bỏ nhiễu để tránh hướng dẫn sai cho mô hình.

● Lưu ý: Chất lượng của tập huấn luyện ảnh hưởng trực tiếp đến hiệu quả của mô hình.

c) Khả năng tổng quát hoá (Generalization)

● Khái niệm: Khả năng dự đoán chính xác của mô hình trên dữ liệu mới, chưa được sử dụng trong quá trình huấn luyện.

● Mục tiêu: Xây dựng mô hình có khả năng tổng quát hoá cao, tức là dự đoán tốt trên mọi dữ liệu trong tương lai, không chỉ trên tập huấn luyện.

d) Quá khớp (Overfitting)

● Khái niệm: Mô hình học quá tốt trên tập huấn luyện, ghi nhớ cả nhiễu và các đặc trưng ngẫu nhiên, dẫn đến khả năng tổng quát hoá kém.

● Biểu hiện: Lỗi huấn luyện (training error) rất thấp, nhưng lỗi kiểm thử (test error) cao.

● Nguyên nhân:

○ Mô hình quá phức tạp so với dữ liệu.

○ Nhiễu, lỗi trong dữ liệu.

○ Tập huấn luyện quá nhỏ.

e) Kém khớp (Underfitting)

● Khái niệm: Mô hình quá đơn giản, không thể học tốt các mẫu phức tạp trong dữ liệu.

● Biểu hiện: Lỗi huấn luyện và lỗi kiểm thử đều cao.

● Nguyên nhân:

○ Mô hình quá đơn giản.

○ Huấn luyện chưa đủ.

f) Hiệu chỉnh (Regularization)

● Khái niệm: Kỹ thuật giảm quá khớp bằng cách thu hẹp không gian tìm kiếm, thêm ràng buộc vào mô hình, hoặc điều chỉnh quá trình huấn luyện.

● Mục tiêu:

○ Giảm độ phức tạp của mô hình.

○ Cân bằng giữa khả năng xấp xỉ trên tập huấn luyện và khả năng tổng quát hoá.

● Các phương pháp:

○ Thêm hàm phạt: Thêm thành phần phạt vào hàm mất mát, hạn chế độ lớn của các trọng số.

○ Dropout: Ngẫu nhiên "tắt" một số kết nối trong mạng nơ-ron trong quá trình huấn luyện.

○ Early Stopping: Dừng quá trình huấn luyện sớm, trước khi mô hình quá khớp.

○ Data Augmentation: Tạo thêm dữ liệu huấn luyện bằng cách biến đổi các mẫu dữ liệu có sẵn.

Lưu ý: Việc lựa chọn lớp mô hình, thuật toán học, và kỹ thuật hiệu chỉnh phù hợp đóng vai trò quan trọng, quyết định sự thành công của hệ thống học máy. Nên thực hiện phân tích cẩn thận, thử nghiệm nhiều phương án, và đánh giá kỹ lưỡng để lựa chọn giải pháp tối ưu cho bài toán cụ thể.

Nguy hiểm tiềm ẩn khi không phân tích kỹ lưỡng:

Nếu không phân tích kỹ lưỡng các vấn đề nêu trên, hệ thống học máy có thể hoạt động kém hiệu quả, đưa ra dự đoán sai lệch, dẫn đến hậu quả nghiêm trọng, đặc biệt trong các lĩnh vực nhạy cảm như y tế, tài chính, an ninh, ...

Ví dụ:

● Hệ thống chẩn đoán ung thư:

Nếu mô hình bị quá khớp, nó có thể chẩn đoán nhầm một người khỏe mạnh là bị ung thư, dẫn đến lo lắng, tốn kém, thậm chí ảnh hưởng đến tâm lý. Ngược lại, nếu mô hình bị kém khớp, nó có thể bỏ sót các trường hợp ung thư thực sự, gây nguy hiểm đến tính mạng.

● Hệ thống lái xe tự động:

Nếu mô hình không có khả năng tổng quát hoá tốt, nó có thể gặp sự cố khi gặp phải tình huống giao thông mới, chưa từng được huấn luyện, dẫn đến tai nạn.

Vì vậy, cần đặc biệt chú trọng đến việc hiểu rõ các vấn đề trong học máy, phân tích kỹ lưỡng dữ liệu, lựa chọn phương pháp phù hợp, và đánh giá cẩn thận hiệu quả của mô hình trước khi triển khai.

4. Tham khảo

<**https://users.soict.hust.edu.vn/khoattq/ml-dm-course/**>
<**https://users.soict.hust.edu.vn/khoattq/ml-dm-course/IT3190-Tai-lieu-doc.pdf**>
<**https://www.youtube.com/playlist?list=PLaKukjQCR56ZRh2cAkweftiZCF2sTg11_**>

Đến đây kết thúc rồi. Cảm ơn bạn đã lắng nghe!