Họ và tên: Trần Quang Đãng

MSSV: 52100174

1. Làm sạch dữ liệu

Làm sạch dữ liệu là quá trình loại bỏ lỗi và sự không nhất quán khỏi dữ liệu. Đây là một bước quan trọng trong xử lý dữ liệu, vì ngay cả một lỗi nhỏ cũng có thể ảnh hưởng đáng kể đến kết quả của các thuật toán học máy.

Một số vấn đề phổ biến cần được giải quyết trong quá trình làm sạch dữ liệu bao gồm:

- Dữ liệu bị thiếu: Một số điểm dữ liệu có thể bị thiếu, do lỗi của con người hoặc vấn đề kỹ thuật.
- Dữ liệu ngoại lai: Dữ liệu ngoại lai là các điểm dữ liệu khác biệt đáng kể so với phần còn lại của dữ liệu. Chúng có thể làm sai lệch kết quả của các thuật toán học máy.
- Dữ liệu trùng lặp: Các điểm dữ liệu trùng lặp có thể xảy ra khi dữ liệu được nhập hoặc thu thập nhiều lần.
- Dữ liệu bị hỏng: Dữ liệu bị hỏng là dữ liệu bị hỏng hoặc không đọc được.

Có nhiều kỹ thuật có thể được sử dụng để làm sạch dữ liệu, chẳng hạn như:

- Chèn dữ liệu: Đây bao gồm việc điền các giá trị bị thiếu bằng ước tính.
- Phát hiện và loại bỏ dữ liệu ngoại lai: Đây bao gồm việc xác định và loại bỏ dữ liệu ngoại lai.
- Xóa trùng dữ liệu: Đây bao gồm việc xác định và loại bỏ các điểm dữ liệu trùng lặp.
- Xác thực dữ liệu: Đây bao gồm kiểm tra dữ liệu để tìm lỗi và sự không nhất quán.

2. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là quá trình chuẩn hóa các giá trị của các điểm dữ liệu để chúng có thang đo chung. Điều này rất quan trọng đối với các thuật toán học máy, vì chúng thường giả định rằng dữ liệu được phân phối theo quy luật bình thường.

Có nhiều cách để chuẩn hóa dữ liệu, chẳng hạn như:

- Chuẩn hóa min-max: Đây bao gồm thay đổi tỷ lệ các giá trị của các điểm dữ liệu sao cho chúng nằm giữa giá trị tối thiểu và tối đa.
- Chuẩn hóa Z-score: Đây bao gồm trừ giá trị trung bình khỏi mỗi điểm dữ liệu và sau đó chia cho độ lệch chuẩn.
- Chuẩn hóa log: Đây bao gồm lấy logarit của mỗi điểm dữ liệu.

3. Chuyển đổi dữ liệu

Chuyển đổi dữ liệu là quá trình chuyển đổi dữ liệu thành định dạng phù hợp hơn cho các thuật toán học máy. Điều này có thể liên quan đến việc thay đổi kiểu dữ liệu, thang đo hoặc cấu trúc của dữ liệu.

Một số chuyển đổi dữ liệu phổ biến được sử dụng trong học máy bao gồm:

- Chuyển đổi dữ liệu phân loại thành dữ liệu số: Điều này là cần thiết cho các thuật toán học máy chỉ có thể hoạt động với dữ liệu số.
- Tỷ lệ dữ liệu: Điều này được thực hiện để đảm bảo rằng tất cả các tính năng đều có thang đo tương tự.
- Mã hóa dữ liệu: Điều này được thực hiện để biểu diễn dữ liệu phân loại dưới dạng dữ liệu số.
- Lựa chọn tính năng: Đây bao gồm việc chọn các tính năng quan trọng nhất cho thuật toán học máy.

Dưới đây là một số ví dụ thực tế về các vấn đề xử lý dữ liệu trong học máy:

- Trong hệ thống lọc thư rác, vấn đề làm sạch dữ liệu sẽ liên quan đến việc loại bỏ thư rác khỏi tập dữ liệu đào tạo. Vấn đề chuẩn hóa dữ liệu sẽ liên quan đến việc thay đổi tỷ lệ các giá trị của các tính năng để chúng có thang đo chung. Vấn đề chuyển đổi dữ liệu sẽ liên quan đến việc chuyển đổi dữ liệu phân loại (chẳng hạn như địa chỉ email của người gửi) thành dữ liệu số.
- Trong hệ thống phát hiện gian lận, vấn đề làm sạch dữ liệu sẽ liên quan đến việc loại bỏ các giao dịch gian lận khỏi tập dữ liệu đào tạo. Vấn đề chuẩn hóa dữ liệu sẽ liên quan đến việc thay đổi tỷ lệ các giá trị của các tính năng để chúng có thang đo chung. Vấn đề chuyển đổi dữ liệu sẽ liên quan đến việc mã hóa dữ liệu phân loại (chẳng hạn như loại giao dịch) dưới dạng dữ liệu số.
- Trong hệ thống phân đoạn khách hàng, vấn đề làm sạch dữ liệu sẽ liên quan đến việc loại bỏ dữ liệu không đầy đủ hoặc không chính xác khỏi hồ sơ khách hàng. Vấn đề chuẩn hóa dữ liệu sẽ liên quan đến việc thay đổi tỷ lệ các giá trị của các tính năng để chúng có thang đo chung.