TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Duy Vũ
Nguyễn Chiêu Bản

HỆ THỐNG GỢI Ý SẢN PHẨM DỰA TRÊN MÔ HÌNH HỌC NHÂN QUẢ

KHÓA LUẬN ĐẠI HỌC KHOA HỌC MÁY TÍNH

Tp. Hồ Chí Minh, tháng 06/2022

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Duy Vũ - 18120264
Nguyễn Chiêu Bản - 18120283

HỆ THỐNG GỢI Ý SẢN PHẨM DỰA TRÊN MÔ HÌNH HỌC NHÂN QUẢ

KHÓA LUẬN ĐẠI HỌC KHOA HỌC MÁY TÍNH
NGƯỜI HƯỚNG DẪN
Th.S. Trần Trung Kiên
TS. Nguyễn Ngọc Thảo

Tp. Hồ Chí Minh, tháng 06/2022

Lời cảm ơn

Tôi xin chân thành cảm ơn ...

Mục lục

Danh sách hình

Danh sách bảng

Chương 1

Giới thiệu

Nhóm em dự định nói rõ phần giới thiệu theo những ý chính sau:

- Phát biểu về bài toán
- Ý nghĩa của bài toán trong bối cảnh bùng nổ thông tin ngày nay
- Thách thức của bài toán trong vấn đề dữ liệu bị bias
- Tác động của vấn đề bias dữ liệu
- Cụ thể hơn về bias dữ liệu trong khóa luận tìm hiểu là selection bias
- Ví dụ về tác động của selection bias trong việc đánh giá các mô hình
- Từ đó dẫn đến bài báo tìm hiểu sử dụng IPS để giải quyết vấn đề selection bias trong việc huấn luyện và đánh giá mô hình

Chương 2

Kiến thức nền tảng

Trong chương này, đầu tiên nhóm chúng em trình bày về thuật toán thuộc nhóm lọc cộng tác (Collaborative filtering) là "Matrix factorization" - thuật toán đề xuất sản phẩm bằng cách phân rã ma trận tương tác. Ngoài ra, nhóm chúng em còn trình bày về hai mô hình học có giám sát cho bài toán phân lớp nhị phân là "Naive Bayes" - mô hình phân lớp dựa trên định lý xác suất Bayes, và "Logistic Regression" - mô hình phân lớp dựa trên hàm sigmoid. Hai mô hình này được sử dụng để ước lượng ma trận xu hướng từ dữ liệu quan sát được. Chương này, đặc biệt là về phần "Matrix Factorization" cung cấp những kiến thức nền tảng để có thể hiểu rõ về những cải tiến mà nhóm em tìm hiểu ở chương kế tiếp.

2.1 "Matrix Factorization"

2.2 "Naive Bayes"

2.3 "Logistic Regression"

Chương 3

Phương pháp tìm hiểu

Chương này nhóm chúng em trình bày về những tìm hiểu của nhóm em thông qua bài báo. Bài báo tập trung nghiên cứu về việc xử lý vấn đề selection bias bằng cách sử dụng độ đo khắc phục bias là Inverse propensity scoring (IPS). Bằng cách sử dụng độ đo này trong quá trình huấn luyện và đánh giá mô hình Matrix Factorization ta có thể thu được mô hình với hiệu suất không bị tác động bởi dữ liệu bị bias. Ngoài ra, trong chương này nhóm chúng em còn đề cập về cách tìm ma trận xu hướng thông qua hai mô hình "Naive Bayes" và "Logistic Regression" đã đề cập ở chương trước.

3.1 Inverse propensity scoring (IPS)

3.2 Ước lượng ma trận xu hướng

3.3 Matrix Factorization từ xu hướng

Chương 4

Các kết quả thí nghiệm

Trong chương này, nhóm chúng em trình bày các kết quả thí nghiệm để đánh giá các đề xuất tìm hiểu được từ bài báo đã được nói ở chương trước. Bộ dữ liệu được dùng để tiến hành thí nghiệm là bộ COAT (bao gồm đánh giá của người dùng cho áo khoác), bộ Yahoo (bao gồm đánh giá của người dùng cho các bài hát), bộ Movielens 100K (bao gồm đánh giá của người dùng cho các bộ phim). Các kết quả thí nghiệm cho thấy khi khi dùng IPS đánh giá mô hình hoàn toàn khớp với hiệu suất thật và tốt hơn nhiều so với các độ đo đánh giá truyền thống. Các kết quả thí nghiệm cũng cho thấy mô hình được huấn luyện dựa trên độ đo IPS cũng cho kết quả tổng quát hóa tốt hơn trên các mức độ selection bias khác nhau.

Chương 5

Tổng kết và hướng phát triển

Tài liệu tham khảo