BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

----------⁂----------

MÔN HỌC HỌC MÁY (MACHINE LEARNING) TRONG PHÂN TÍCH KINH DOANH

Giảng viên hướng dẫn: Tiến sĩ TRẦN DUY THANH

Mã lớp học phần: 251BIM401407

Chủ đề: PHÂN TÍCH VÀ MINH HỌA CÁC PHƯƠNG PHÁP VỀ KỸ THUẬT XẾP HẠNG

SINH VIÊN THỰC HIỆN: LÊ PHƯỚC THỊNH

MÃ SỐ SINH VIÊN: K234161856

Thành phố Hồ Chí Minh, tháng 9 năm 2025

MỤC LỤC

I. TỔNG QUAN LÝ THUYẾT VỀ KỸ THUẬT XẾP HẠNG.

Lý thuyết về Kỹ thuật xếp hạng hình thành nền tảng cho các hệ thống có nhiệm vụ sắp xếp thông tin một cách thông minh và hiệu quả. Khác với việc lọc dữ liệu thông thường, kỹ thuật xếp hạng là một quá trình phức tạp nhằm xác định thứ tự ưu tiên của các đối tượng dựa trên mức độ phù hợp với một tiêu chí cụ thể, thường là một truy vấn từ người dùng. Mục tiêu cuối cùng của nó là tối ưu hóa trải nghiệm người dùng bằng cách đảm bảo rằng những mục quan trọng và liên quan nhất luôn xuất hiện ở các vị trí dễ thấy nhất.

Trong bối cảnh của Học máy, lĩnh vực nghiên cứu này được gọi là Learning to Rank (Máy học để Xếp hạng). Đây là một nhánh quan trọng tập trung vào việc phát triển các mô hình có thể tự động học cách sắp xếp các đối tượng như tài liệu, sản phẩm hoặc tin tức dựa trên dữ liệu lịch sử về mức độ liên quan. Bài toán này không đơn thuần là phân loại nhị phân (liên quan/không liên quan) mà là một bài toán về thứ bậc, nơi thứ tự tương đối giữa các mục quan trọng hơn điểm số tuyệt đối của từng mục.

Lý thuyết về kỹ thuật xếp hạng trong Learning to Rank chủ yếu được phân loại thành ba phương pháp tiếp cận chính dựa trên đơn vị học tập. Phương pháp đầu tiên là Pointwise (theo điểm). Cách tiếp cận này đơn giản hóa bài toán bằng coi mỗi cặp truy vấn-tài liệu là một mẫu độc lập và giải quyết nó như một bài toán hồi quy (dự đoán điểm số) hoặc phân loại (dự đoán nhãn liên quan). Mặc dù dễ hiểu và triển khai, Pointwise có hạn chế là bỏ qua mối quan hệ tương đối giữa các tài liệu.

Phương pháp thứ hai, Pairwise (theo cặp), ra đời để khắc phục hạn chế trên. Thay vì dự đoán điểm số tuyệt đối, Pairwise chuyển bài toán thành việc học cách so sánh. Mô hình được huấn luyện để xác định xem, với một truy vấn cụ thể, tài liệu A

có liên quan hơn tài liệu B hay không. Bằng cách này, mô hình trực tiếp tối ưu hóa thứ tự sắp xếp, tập trung vào mối quan hệ tương đối giữa các cặp tài liệu.

Phương pháp tinh vi và trực tiếp nhất là Listwise (theo danh sách). Listwise xem toàn bộ danh sách các tài liệu ứng với một truy vấn như một đơn vị học duy nhất. Mô hình cố gắng tối ưu hóa trực tiếp chất lượng của cả danh sách kết quả bằng cách sử dụng các chỉ số đánh giá xếp hạng như NDCG hoặc MAP ngay trong hàm mất mát. Mặc dù phức tạp hơn để triển khai, Listwise thường cho kết quả tốt nhất vì nó mô phỏng sát nhất mục tiêu thực tế của bài toán xếp hạng.

Tóm lại, lý thuyết về kỹ thuật xếp hạng không chỉ dừng lại ở các thuật toán sắp xếp đơn giản mà là một hệ thống lý thuyết sâu rộng giải quyết bài toán tối ưu hóa thứ bậc. Sự phát triển từ Pointwise, Pairwise đến Listwise phản ánh nỗ lực không ngừng trong việc tìm kiếm các mô hình ngày càng hiệu quả hơn. Những kỹ thuật này đóng vai trò then chốt trong hiệu suất của các công cụ tìm kiếm, hệ thống đề xuất và vô số ứng dụng trí tuệ nhân tạo khác, định hình cách chúng ta khám phá và tương tác với thông tin trong thế giới số.

II. LÝ THUYẾT VỀ PHƯƠNG PHÁP POINTWISE

Phương pháp Pointwise là một trong ba cách tiếp cận cơ bản nhất trong lĩnh vực Learning to Rank (Máy học để Xếp hạng), bên cạnh Pairwise và Listwise. Điểm khác biệt cốt lõi của Pointwise nằm ở cách nó đơn giản hóa bài toán xếp hạng - một nhiệm vụ vốn phức tạp và liên quan đến thứ bậc - thành một loạt các bài toán dự đoán đơn lẻ và độc lập. Thay vì xem xét mối quan hệ giữa các tài liệu với nhau, Pointwise tập trung vào việc đánh giá từng tài liệu một cách riêng rẽ dựa trên truy vấn của người dùng.

Về bản chất học thuật, Pointwise giải quyết bài toán xếp hạng như một bài toán hồi quy (regression) hoặc phân loại (classification) truyền thống. Trong đó, mỗi cặp (truy vấn, tài liệu) được xem như một điểm dữ liệu độc lập. Mô hình được huấn luyện để dự đoán một "điểm số liên quan" tuyệt đối cho mỗi điểm dữ liệu này. Nếu được áp dụng như một bài toán phân loại, đầu ra sẽ là các nhãn rời rạc như "không liên quan", "ít liên quan", "rất liên quan". Nếu được áp dụng như một bài toán hồi quy, đầu ra sẽ là một điểm số liên tục, chẳng hạn từ 0 đến 5.

Quy trình vận hành của phương pháp này có thể được tóm tắt qua hai giai đoạn chính. Đầu tiên, trong giai đoạn huấn luyện, mô hình học mối quan hệ giữa các đặc trưng (features) của tài liệu (ví dụ: tần suất từ khóa, độ tin cậy của nguồn) và điểm số liên quan do con người gán nhãn. Sau đó, trong giai đoạn dự đoán và xếp hạng, với một truy vấn mới, mô hình sẽ ước tính điểm số liên quan cho từng tài liệu trong danh sách ứng viên một cách độc lập. Thứ hạng cuối cùng được xác định đơn giản bằng cách sắp xếp các tài liệu theo thứ tự điểm số dự đoán giảm dần.

Ưu điểm lớn nhất của Pointwise là tính đơn giản và dễ triển khai. Do quy về các bài toán học máy kinh điển, nó cho phép các nhà phát triển tận dụng hầu hết các thuật toán đã được nghiên cứu kỹ lưỡng như Cây quyết định, SVM hay Hồi quy Logistic mà không cần phải thiết kế các thuật toán phức tạp mới. Điều này làm cho Pointwise trở thành một lựa chọn tuyệt vời để bắt đầu với bài toán Learning to Rank.

Tuy nhiên, điểm hạn chế then chốt của phương pháp này xuất phát từ chính sự đơn giản của nó. Pointwise bỏ qua hoàn toàn ngữ cảnh thứ hạng và mối quan hệ tương đối giữa các tài liệu. Trong thực tế, việc xác định chính xác thứ tự (tài liệu A quan trọng hơn tài liệu B) thường quan trọng hơn là dự đoán một điểm số liên quan tuyệt đối. Việc không xem xét đến mối quan hệ này có thể dẫn đến chất lượng xếp hạng tổng thể chưa thực sự tối ưu so với các phương pháp Pairwise và Listwise.

III. LÝ THUYẾT VỀ PHƯƠNG PHÁP PAIRWISE

Phương pháp Pairwise đại diện cho một bước tiến quan trọng so với phương pháp Pointwise trong lĩnh vực Learning to Rank (Máy học để Xếp hạng). Nếu Pointwise xem xét từng tài liệu một cách cô lập, thì tư tưởng cốt lõi của Pairwise là chuyển bài toán xếp hạng thành bài toán học cách so sánh. Thay vì dự đoán một điểm số liên quan tuyệt đối, mô hình Pairwise tập trung vào việc xác định thứ tự ưu tiên tương đối giữa hai tài liệu bất kỳ khi chúng cùng được đặt trong ngữ cảnh của một truy vấn cụ thể.

Về mặt bản chất học thuật, phương pháp này giải quyết một bài toán phân loại nhị phân cho mỗi cặp tài liệu. Cụ thể, với một truy vấn cho trước và hai tài liệu A và B, mô hình được huấn luyện để trả lời câu hỏi: "Tài liệu A có liên quan hơn tài liệu B hay không?". Đầu ra của mô hình là xác suất hoặc một điểm số thể hiện khả năng tài liệu A được xếp hạng cao hơn tài liệu B. Bằng cách này, Pairwise trực tiếp tối ưu hóa "bậc" (order) của các tài liệu, điều mà Pointwise không làm được.

Quy trình huấn luyện của Pairwise đặc trưng bởi việc tạo ra các cặp dữ liệu huấn luyện. Từ dữ liệu gốc (gồm các truy vấn và danh sách tài liệu đã được gán nhãn liên quan), người ta sẽ tạo ra vô số cặp huấn luyện. Chẳng hạn, nếu với một truy vấn, tài liệu A có nhãn "Rất liên quan" và tài liệu B có nhãn "Ít liên quan", ta sẽ tạo ra một cặp huấn luyện (A, B) với nhãn "A liên quan hơn B". Quá trình này được lặp lại cho tất cả các cặp tài liệu có thể trong danh sách.

Ưu điểm chính của phương pháp Pairwise nằm ở chỗ nó trực tiếp mô hình hóa mối quan hệ thứ hạng, vốn là mục tiêu cốt lõi của bài toán xếp hạng. Mô hình không cần phải dự đoán chính xác một điểm số tuyệt đối khó có thể định nghĩa, mà chỉ cần học được sự ưu tiên tương đối. Điều này thường phù hợp hơn với bản chất của bài toán và trong nhiều trường hợp cho kết quả vượt trội so với Pointwise. Các thuật toán nổi tiếng như RankNet và LambdaMART đều dựa trên nền tảng tư tưởng Pairwise.

Tuy nhiên, phương pháp này cũng tồn tại những hạn chế đáng kể. Việc tạo ra số lượng lớn các cặp huấn luyện có thể dẫn đến sự mất cân bằng dữ liệu nghiêm trọng.

Ví dụ, số cặp mà tài liệu liên quan hơn sẽ chiếm đa số so với các cặp ngược lại. Hơn nữa, số lượng cặp có thể tăng theo hàm bậc hai so với số tài liệu trong một danh sách, dẫn đến chi phí tính toán cao. Một nhược điểm sâu xa hơn là mô hình chỉ tối ưu hóa thứ tự của từng cặp mà không xem xét đến chất lượng của toàn bộ danh sách xếp hạng cuối cùng, đây là điểm mà phương pháp Listwise giải quyết tốt hơn.

IV. LÝ THUYẾT VỀ PHƯƠNG PHÁP LISTWISE.

Phương pháp Listwise được xem là cách tiếp cận trực tiếp và toàn diện nhất đối với bài toán Learning to Rank. Nếu Pointwise xử lý từng tài liệu một cách cô lập và Pairwise tập trung vào mối quan hệ giữa các cặp tài liệu, thì tư tưởng cốt lõi của Listwise là xem toàn bộ danh sách kết quả cho một truy vấn như một đơn vị học tập duy nhất. Phương pháp này thẳng thắn thừa nhận rằng mục tiêu cuối cùng của chúng ta là tạo ra một thứ tự tối ưu cho cả danh sách, và do đó, nó cố gắng tối ưu hóa trực tiếp chất lượng của toàn bộ danh sách đó.

Về bản chất học thuật, Listwise trực tiếp tối ưu hóa các chỉ số đánh giá xếp hạng. Thay vì sử dụng các hàm mất mát thông thường như trong hồi quy hay phân loại, các thuật toán Listwise được thiết kế để tối đa hóa các chỉ số như NDCG (Normalized Discounted Cumulative Gain), MAP (Mean Average Precision), hoặc MRR (Mean Reciprocal Rank). Đây chính là những chỉ số mà người ta thường dùng để đánh giá hiệu năng của một hệ thống xếp hạng trong thực tế. Bằng cách nhúng các chỉ số này vào quá trình tối ưu, Listwise đảm bảo rằng mô hình học được chính xác những gì chúng ta mong đợi ở nó: một danh sách xếp hạng chất lượng cao.

Ưu điểm vượt trội của Listwise nằm ở tính toàn cục và trực tiếp của nó. Phương pháp này tránh được những hạn chế của Pointwise (bỏ qua mối quan hệ tương đối) và Pairwise (chỉ tối ưu từng cặp đôi một, có thể dẫn đến mâu thuẫn và không tối ưu được toàn cục). Listwise xem xét tất cả các tài liệu cùng một lúc, cho phép nó nắm bắt được "bức tranh tổng thể" về thứ hạng và đưa ra quyết định tối ưu nhất cho cả tập hợp. Về lý thuyết, đây là phương pháp gần với mục tiêu thực tế nhất và thường cho chất lượng xếp hạng tốt nhất.

Tuy nhiên, sự tinh vi này đi kèm với độ phức tạp đáng kể. Việc tối ưu hóa trực tiếp các chỉ số như NDCG là rất khó khăn vì chúng thường là các hàm không liên tục (discontinuous), không khả vi (non-differentiable), nghĩa là chúng ta không thể tính đạo hàm một cách thông thường để áp dụng các thuật toán tối ưu như Gradient Descent. Để vượt qua thách thức này, các nhà nghiên cứu đã phát triển những kỹ thuật

ước tính đạo hàm thông minh (như trong ListNet, ListMLE) hoặc xấp xỉ các chỉ số bằng những hàm mất mát liên tục, khả vi có tính chất tương tự.