**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**



**BÁO CÁO BÀI TẬP ỨNG DỤNG THUẬT TOÁN**

**NHÓM 7**

**GVHD: ThS. Nguyễn Thị Anh Thư**

**1.**Nguyễn Thế Tiến Đạt **MSSV:** 22520225

**2.**Vũ Quang Huy **MSSV:** 22520587

**3.**Võ Nhất Phương **MSSV:** 22521172

**4.**Lê Nguyễn Thùy Dương **MSSV:** 22520298

**5.**Đoàn Ngọc Thanh Sơn **MSSV:** 21521385

**Tp. Hồ Chí Minh, 04/2024**

**NỘI DUNG THẢO LUẬN**:

**Bài tập: Ứng dụng của các thuật toán khai phá dữ liệu:**

* **Phân loại (Classification)**
* **Phân cụm (Clustering)**
* **Luật kết hợp (Association Rule)**
* **Hồi quy (Regression)**
* **Chuỗi thời gian (Time series)**

**Trả lời:**

**PHÂN LOẠI (CLASSIFICATION)**

|  | **Hệ thống phân loại khách hàng tiềm năng trong marketing** | **Hệ thống phân loại email spam** |
| --- | --- | --- |
| **Mục đích** | Giúp tối ưu hóa chiến lược marketing, tăng tỷ lệ chuyển đổi và tối đa hóa hiệu suất chiến dịch. | Giúp cải thiện trải nghiệm người dùng bằng cách loại bỏ thông tin không mong muốn, giảm bớt nguy cơ về việc mở các email gây hại và bảo vệ người dùng khỏi các mối đe dọa an ninh mạng. |
| **Ứng dụng** | Phân loại các khách hàng dựa trên đặc điểm và hành vi của họ để tập trung các nỗ lực marketing vào nhóm đối tượng có tiềm năng cao nhất. | Là nhận diện và loại bỏ các email không mong muốn hoặc có nguy cơ là spam khỏi hộp thư đến của người dùng |
| **Phạm vi triển khai** | Có thể linh hoạt, từ doanh nghiệp nhỏ sử dụng các giải pháp SaaS đơn giản đến doanh nghiệp lớn triển khai hệ thống phức tạp tích hợp với các hệ thống sẵn có | Thường là trên các máy chủ email hoặc dịch vụ email như Gmail, Outlook, hay Yahoo Mail. Hệ thống này thường hoạt động tự động và không yêu cầu sự can thiệp đặc biệt từ người dùng.. |
| **Kỹ thuật** | Dựa trên nhiều yếu tố như hành vi mua hàng trước đây, thông tin cá nhân, tương tác với thương hiệu, và các chỉ số khác để xác định mức độ quan tâm và tiềm năng mua hàng của mỗi khách hàng. | Dựa trên các yếu tố như từ khóa, địa chỉ email nguồn, đường dẫn URL, và cấu trúc của email được sử dụng để xác định xem một email có phải là spam hay không |
| **Quy trình** | -Thu thập dữ liệu: Thu thập dữ liệu từ nhiều nguồn khác nhau như hệ thống CRM, trang web, email marketing, mạng xã hội, và các nguồn dữ liệu bên ngoài khác.  -Tiền xử lý dữ liệu: Tiền xử lý dữ liệu để loại bỏ dữ liệu không chính xác, thiếu sót và chuẩn hóa dữ liệu.  -Rút trích đặc trưng: Rút trích các đặc trưng quan trọng từ dữ liệu như hành vi trực tuyến, thông tin cá nhân, và lịch sử mua hàng.  -Phân loại khách hàng: Sử dụng các thuật toán phân loại như cây quyết định, máy vector hỗ trợ (SVM), hoặc mạng nơ-ron để phân loại khách hàng thành các nhóm dựa trên đặc trưng.  -Đào tạo mô hình: Sử dụng dữ liệu đã được gán nhãn để đào tạo mô hình phân loại khách hàng.  -Kiểm định và điều chỉnh: Kiểm định mô hình trên tập dữ liệu kiểm tra và điều chỉnh các siêu tham số nếu cần thiết để cải thiện hiệu suất.  -Triển khai và tích hợp: Triển khai mô hình vào hệ thống CRM hoặc các công cụ marketing và tích hợp với quy trình marketing hiện có. | -Thu thập dữ liệu: Thu thập dữ liệu từ các email đã gửi và nhận, bao gồm nội dung, tiêu đề, địa chỉ email nguồn, và các đặc điểm khác.  -Tiền xử lý dữ liệu: Tiền xử lý dữ liệu để chuẩn hóa và rút trích đặc trưng quan trọng như từ khóa, địa chỉ IP, đường dẫn URL.  -Xây dựng mô hình: Sử dụng các thuật toán máy học như Naive Bayes, Random Forest, hoặc Support Vector Machines để xây dựng mô hình phân loại spam và không spam.  -Đào tạo mô hình: Sử dụng dữ liệu đã được gán nhãn để đào tạo mô hình máy học.  -Kiểm định và điều chỉnh: Kiểm định mô hình trên tập dữ liệu kiểm tra và điều chỉnh các siêu tham số nếu cần thiết để cải thiện hiệu suất.  -Triển khai và tích hợp: Triển khai mô hình vào hệ thống email và tích hợp với quy trình xử lý email hiện có. |
| **Ngôn ngữ** | Python, Java, Scala,R,... | Python, Java, Scala,R,... |
| **Xu hướng** | Kết hợp các công nghệ trí tuệ nhân tạo và học máy để tạo ra các hệ thống phân loại khách hàng thông minh hơn và tối ưu hóa chiến lược marketing. | Phát triển các kỹ thuật phát hiện spam mới như phát hiện tấn công dựa trên AI hoặc phát hiện mối đe dọa mới từ các chiến lược spam tiến bộ. |

**PHÂN CỤM (CLUSTERING)**

|  | **Hệ khuyến nghị (Recommendation System)** | **Hệ thống phân loại và lưu trữ tài liệu** |
| --- | --- | --- |
| **Mục đích** | - Cung cấp các đề xuất về các hạng mục/nội dung đề xuất cho người dùng  - Dự đoán và khuyến nghị những sản phẩm/dịch vụ/nội dung mà người dùng có thể quan tâm | - Gán các lớp hoặc danh mục khác nhau cho các tài liệu theo yêu cầu, giúp lưu trữ, quản lý và phân tích tài liệu  - Có thể thực hiện phân loại một cách tự động hóa |
| **Ứng dụng** | Hệ thống khuyến nghị của Youtube, Netflix,... | - Quản lý tài liệu trong các công ty, doanh nghiệp  - Lưu trữ thư viện  - Quản lý hồ sơ y tế  … |
| **Phạm vi triển khai** | - Nền tảng phát trực tuyến  - Mạng xã hội và truyền thông  - Thương mại điện tử | Trong doanh nghiệp, y tế, nghiên cứu, pháp lý, giáo dục… |
| **Kỹ thuật** | - Content-based  - Collaborative filtering  - Demographic  - Knowledge-based  - Hybrid recommender system | - Phân loại Naive Bayes  - Cây quyết định (Decision Tree)  - K-nearest neighbor  - Xử lý ngôn ngữ tự nhiên (NLP)  - Mạng nơ ron (Neural Network) |
| **Quy trình** | - Thu thập dữ liệu từ những hành vi của người dùng  - Phân tích nội dung sản phẩm và so sánh với các dữ liệu đã thu thập từ người dùng  - Tính toán các tương đồng và đưa ra những sản phẩm/dịch vụ có tương thích cao | - Xác định định dạng của các file lưu trữ  - Phân loại cấu trúc tài liệu: có cấu trúc, phi cấu trúc, bán cấu trúc dựa trên các thuật toán phân loại  - Gắn thẻ và ghi chú cho mỗi tài liệu với từ khóa, mô tả, và metadata khác để tăng cường khả năng tìm kiếm và truy xuất  - Tìm kiếm và truy xuất thông tin tài liệu lưu trữ dựa trên các từ khóa hoặc thuộc tính |
| **Ngôn ngữ** | Phổ biến nhất là Python | Nhiều ngôn ngữ có thể hỗ trợ như: Python, Java, C#,... |
| **Xu hướng** | - Hệ khuyến nghị là hệ thống khá cần thiết hiện nay vì có thể đáp ứng khách hàng giúp họ tìm ra những sản phẩm/dịch vụ mới  - Giúp người dùng có thể tìm kiếm nhanh chóng dựa vào những đề xuất hệ thống đưa ra | - Việc quản lý và lưu trữ một lượng lớn dữ liệu là điều cần thiết và sử dụng hệ thống này sẽ giúp tăng khả năng quản lý  - Ngoài ra, còn có thể tăng cường hiệu suất truy xuất, không gian lưu trữ dữ liệu và khả năng mở rộng |

**HỒI QUY (REGRESSION)**

|  | **Hệ thống nghiên cứu  hành vi tiêu dùng** | **Hệ thống dự báo giá cổ phiếu và thị trường tài chính** |
| --- | --- | --- |
| **Mục đích** | * Nghiên cứu hành vi tiêu dùng nhằm hiểu rõ hơn về cách khách hàng tìm kiếm, lựa chọn, sử dụng và đánh giá sản phẩm hoặc dịch vụ. * Xác định các yếu tố ảnh hưởng đến quyết định mua sắm của khách hàng. | * Dự báo giá cổ phiếu và thị trường tài chính giúp các nhà đầu tư, doanh nghiệp, và chuyên gia tài chính đưa ra quyết định đầu tư thông minh. * Hiểu rõ hơn về yếu tố ảnh hưởng đến biến động giá cổ phiếu. |
| **Ứng dụng** | * Thiết kế chiến lược marketing phù hợp. * Tối ưu hóa quy trình chăm sóc khách hàng. * Tăng doanh số bán hàng. | * Dự báo giá cổ phiếu để đưa ra quyết định mua hoặc bán. * Đánh giá rủi ro và lợi nhuận trong đầu tư. |
| **Phạm vi triển khai** | * Thời trang: Nghiên cứu về việc mua sắm thời trang và ảnh hưởng của quảng cáo. * Thực phẩm: Xem xét hành vi mua sắm thực phẩm và sự ảnh hưởng của đánh giá sản phẩm. | * Thị trường tài chính: Dự báo giá cổ phiếu, chỉ số thị trường, tỷ giá hối đoái, v.v. * Doanh nghiệp: Dự báo giá cổ phiếu của một công ty cụ thể. |
| **Kỹ thuật** | * Hồi quy tuyến tính | * Hồi quy tuyến tính đa biến * Machine learning |
| **Quy trình** | * Thu thập dữ liệu về hành vi tiêu dùng. * Phân tích dữ liệu để hiểu rõ hơn về quyết định mua sắm của khách hàng. * Áp dụng kỹ thuật hồi quy để dự đoán và tối ưu hóa quyết định đầu tư. | * Thu thập dữ liệu. * Xây dựng mô hình hồi quy. * Đánh giá và tinh chỉnh mô hình. |
| **Ngôn ngữ** | * Sử dụng ngôn ngữ thống nhất và rõ ràng để trình bày kết quả nghiên cứu. | * Python |
| **Xu hướng** | * Nghiên cứu hành vi tiêu dùng ngày càng tập trung vào việc cá nhân hóa thương hiệu và tối ưu hóa trải nghiệm khách hàng. * Sự phát triển của công nghệ và dữ liệu giúp nghiên cứu hành vi tiêu dùng trở nên hiệu quả hơn. | * Sử dụng dữ liệu lớn và học máy để cải thiện dự báo. * Tích hợp dữ liệu phi tài chính (ví dụ: dữ liệu xã hội) vào mô hình dự báo. |

**LUẬT KẾT HỢP (ASSOCIATION RULE)**

|  | **Hệ thống quản lý bán lẻ** | **Hệ thống phân tích giỏ hàng** |
| --- | --- | --- |
| **Mục đích** | * Tìm kiếm các luật kết hợp đơn giản, ít thuộc tính(phù hợp với dữ liệu nhỏ) | * Tìm kiếm các luật kết hợp phức tạp, nhiều thuộc tính(phù hợp với dữ liệu lớp) * Giải quyết vấn đề vụ nổ kết hợp(combination explosion) |
| **Ứng dụng** | * Phân tích giỏ hàng * Khuyến mại * Phân tích thị trường | * Phân tích giỏ hàng * Khuyến mại * Phân tích thị trường |
| **Phạm vi triển khai** | * Bán lẻ * Chăm sóc sức khỏe * Viễn thông * Sản xuất   (Lưu ý Apriori phù hợp với dữ liệu nhỏ như bán hàng tại cửa hàng nhỏ) | * Bán lẻ * Ngân hàng * Sản xuất   (FP-growth phù hợp với dữ liệu lớn và phức tạp như dữ liệu của ngân hàng) |
| **Kỹ thuật** | * Sử dụng thuật toán Apriori để tìm kiếm các tập mục thường xuyên (frequent itemset) và sau đó tạo ra các luật kết hợp từ các tập mục này. | * Sử dụng cấu trúc dữ liệu "FP-tree" để lưu trữ thông tin về các tập mục thường xuyên. * Sử dụng thuật toán FP-growth để tìm kiếm các luật kết hợp từ "FP-tree". |
| **Quy trình** | * Thu thập dữ liệu * Chuẩn bị dữ liệu * Khai phá luật kết hợp bằng thuật toán Apriori * Đánh giá luật kết hợp * Ứng dụng luật kết hợp | * Thu thập dữ liệu * Chuẩn bị dữ liệu * Khai phá luật kết hợp bằng thuật toán FP-growth * Đánh giá luật kết hợp * Ứng dụng luật kết hợp |
| **Áp dụng ngôn ngữ** | Python, Java,... | Python, Java,... |
| **Xu hướng** | * Khai phá kết hợp trong thời gian thực * Khai phá kết hợp trong dữ liệu đồ thị * Khai phá kết hợp trong dữ liệu đa phương tiện | * Khai phá kết hợp trong thời gian thực * Khai phá kết hợp trong dữ liệu đồ thị * Khai phá kết hợp trong dữ liệu đa phương tiện |

**CHUỖI THỜI GIAN (TIME SERIES)**

|  | **Hệ thống nghiên cứu cổ phiếu** | **Hệ thống phân tích dự báo thời tiết** |
| --- | --- | --- |
| **Mục đích** | Dự đoán giá cổ phiếu trong tương lai để ra quyết định đầu tư.  Phân tích xu hướng thị trường và tìm ra các cơ hội giao dịch.  Đánh giá và quản lý rủi ro trong các quỹ đầu tư hoặc các quỹ hưu trí.  Xây dựng chiến lược giao dịch dựa trên thông tin từ phân tích chuỗi thời gian | Dự đoán các biến động trong thời tiết như nhiệt độ, độ ẩm, áp suất không khí, và lượng mưa trong tương lai.  Phân tích xu hướng thời tiết để đưa ra cảnh báo và quản lý rủi ro cho các lĩnh vực như nông nghiệp, hàng không, du lịch và năng lượng.  Hỗ trợ trong việc ra quyết định liên quan đến quản lý tài nguyên tự nhiên và địa lý. |
| **Ứng dụng** | Các công ty quản lý quỹ đầu tư và quỹ hưu trí sử dụng hệ thống này để tối ưu hóa lợi nhuận và quản lý rủi ro.  Các nhà đầu tư cá nhân sử dụng để hỗ trợ quyết định đầu tư của họ.  Các công ty môi giới chứng khoán sử dụng để cung cấp dịch vụ tư vấn đầu tư cho khách hàng của họ. | Các dự báo thời tiết dùng để cung cấp thông tin cho các ứng dụng di động, trang web và truyền thông để cảnh báo người dùng về điều kiện thời tiết dự kiến.  Trong nông nghiệp, dự đoán thời tiết giúp các nhà nông lên kế hoạch gieo trồng, chăm sóc và thu hoạch một cách hiệu quả.  Trong ngành năng lượng, dự báo thời tiết hỗ trợ quản lý việc sản xuất và phân phối năng lượng, đặc biệt là trong các loại năng lượng như điện gió và điện mặt trời. |
| **Phạm vi triển khai** | Có thể triển khai ở bất kỳ thị trường chứng khoán nào trên thế giới, miễn là có dữ liệu lịch sử.  Có thể áp dụng cho các loại cổ phiếu, từ cổ phiếu công nghệ đến cổ phiếu dầu khí hoặc cổ phiếu ngân hàng. | Mô hình Weather data có thể triển khai ở bất kỳ vùng địa lý nào trên thế giới, miễn là có dữ liệu thời tiết lịch sử và hiện tại.  Các tổ chức như dự báo thời tiết quốc gia, công ty dịch vụ thời tiết, và các tổ chức nghiên cứu thời tiết có thể triển khai mô hình này. |
| **Kĩ thuật** | Sử dụng các mô hình phân tích chuỗi thời gian như ARIMA, SARIMA, GARCH, hoặc mạng nơ-ron LSTM.  Tiền xử lý dữ liệu để loại bỏ nhiễu và điền vào các giá trị bị thiếu.  Đánh giá và tinh chỉnh mô hình dựa trên kết quả kiểm định và đánh giá. | Sử dụng các mô hình phân tích chuỗi thời gian như ARIMA, SARIMA, GARCH, hoặc mạng nơ-ron LSTM.  Tiền xử lý dữ liệu để loại bỏ nhiễu và điền vào các giá trị bị thiếu.  Sử dụng dữ liệu lịch sử và hiện tại về thời tiết để huấn luyện mô hình và dự đoán tương lai. |
| **Quy trình** | 1. Thu thập dữ liệu giá cổ phiếu lịch sử. 2. Tiền xử lý và phân tích dữ liệu. 3. Chọn mô hình phù hợp và huấn luyện mô hình. 4. Đánh giá hiệu suất của mô hình. 5. Tối ưu hóa và điều chỉnh mô hình. 6. Triển khai mô hình và theo dõi hiệu suất trong thời gian thực. | 1. Thu thập dữ liệu thời tiết lịch sử từ các nguồn đáng tin cậy như cơ quan dự báo thời tiết quốc gia hoặc các trang web thời tiết. 2. Tiền xử lý và phân tích dữ liệu thời tiết để chuẩn bị cho việc xây dựng mô hình. 3. Chọn mô hình phù hợp và huấn luyện mô hình trên dữ liệu lịch sử. 4. Đánh giá hiệu suất của mô hình bằng cách sử dụng các phương pháp đánh giá và kiểm định. 5. Tối ưu hóa và điều chỉnh mô hình dựa trên kết quả đánh giá. 6. Triển khai mô hình để dự đoán và cung cấp thông tin về thời tiết cho người dùng cuối. |
| **Áp dụng ngôn ngữ** | Ngôn ngữ lập trình phổ biến như Python và R thường được sử dụng.  Các thư viện và framework như Pandas, NumPy, scikit-learn (Python), và ggplot2 (R) thường được sử dụng để phân tích và mô hình hóa dữ liệu. | Các ngôn ngữ lập trình như Python sử dụng NumPy để xử lý và mô hình hóa dữ liệu và biểu diễn dữ liệu. |
| **Xư hướng** | Xu hướng hiện tại trong phân tích chuỗi thời gian bao gồm sự phát triển của mô hình học sâu như LSTM và kết hợp các phương pháp học máy với phân tích chuỗi thời gian truyền thống để cải thiện hiệu suất. | Trong lĩnh vực này đang chuyển dịch về sự kết hợp giữa phương pháp học sâu và các kỹ thuật chuỗi thời gian truyền thống. Các mô hình học sâu như LSTM đang ngày càng được ứng dụng rộng rãi để cải thiện khả năng dự báo và phân tích. |