**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN HỌC MÔN**

**NHẬP MÔN HỌC MÁY**

**MACHINE LEARNING**

**FINAL ESSAY**

*Người hướng dẫn*: **Thầy Lê Anh Cường**

*Người thực hiện*: **Nguyễn Tiến Dũng – 52000883**

Lớp **: 503044**

Khoá  **: 24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN HỌC MÔN**

**NHẬP MÔN HỌC MÁY**

**MACHINE LEARNING**

**FINAL ESSAY**

*Người hướng dẫn*: **Thầy Lê Anh Cường**

*Người thực hiện*: **Nguyễn Tiến Dũng – 52000883**

Lớp **: 503044**

Khoá  **: 24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

Chúng em xin chân thành gửi lời cảm ơn này đến Thầy **Lê Anh Cường** giảng viên phụ trách giảng dạy bộ môn nhập môn học máy. Nhờ có sự tận tình giảng dạy, truyền đạt kiến thức của quý cô mà chúng em mới đủ kiến thức để hoàn thành báo cáo cuối kỳ này.

Song song với đó, chúng em cũng xin gửi lời cảm ơn đến Khoa Công Nghệ Thông Tin, trường Đại học Tôn Đức Thắng vì đã tạo điều kiện cho chúng em học tập, nghiên cứu trong suốt quá trình học tập môn học này nói riêng và cả quá trình học tại môi trường Đại học nói chung. Một lần nữa chúng em xin gửi lời cảm ơn chân thành đến mọi người và chúc tất cả thật nhiều sức khỏe.

*TP. Hồ Chí Minh, ngày 23 tháng 12 năm 2023*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Tiến Dũng*

**ĐỒ ÁN / BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Thầy **Lê Anh Cường**. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án môn học còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án môn học của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. HCM, ngày 23 tháng 12 năm 2023*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Tiến Dũng*

TÓM TẮT

Cùng với sự phát triển của các lĩnh vực kinh tế, xã hội, nhu cầu ứng dụng công nghệ thông tin ngày càng cao và không ngừng biến đổi. Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán cho phép máy tính học từ dữ liệu mà không cần được lập trình rõ ràng. Học máy có thể được sử dụng để giải quyết nhiều loại vấn đề, bao gồm phân loại, dự đoán, phân cụm và xử lý ngôn ngữ tự nhiên.

Môn học nhập môn học máy cung cấp cho sinh viên kiến thức cơ bản về học máy. Môn học này bao gồm các chủ đề như các khái niệm cơ bản về học máy, tiền xử lý dữ liệu, xây dựng mô hình học máy và ứng dụng học máy.

Kiến thức và kỹ năng cần có để theo học môn học này bao gồm kiến thức cơ bản về toán và thống kê, và khả năng lập trình bằng Python hoặc một ngôn ngữ lập trình khác.

Mục tiêu của môn học này là giúp sinh viên hiểu được các khái niệm cơ bản về học máy và cách áp dụng học máy để giải quyết các vấn đề thực tế.

Học máy là một lĩnh vực đang phát triển nhanh chóng và có nhiều ứng dụng tiềm năng trong nhiều lĩnh vực khác nhau. Môn học nhập môn học máy cung cấp cho sinh viên nền tảng vững chắc để theo đuổi lĩnh vực này.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc121002887)i

[TÓM TẮT](#_Toc121002887) 3

[CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN 5](#_Toc121002888)

[1.1 Giới thiệu môn học 5](#_Toc121002889)

[1.2 Đề tài nghiên cứu 5](#_Toc121002890)

[PHẦN 2: LỢI THẾ VÀ CHIẾN LƯỢC CẠNH TRANH 7](#_Toc121002891)

[2.1 Câu hỏi 1 7](#_Toc121002892)

[2.2 Câu hỏi 2 9](#_Toc121002892)

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN

1.1 Giới thiệu môn học

Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán cho phép máy tính học từ dữ liệu mà không cần được lập trình rõ ràng. Học máy có thể được sử dụng để giải quyết nhiều loại vấn đề, bao gồm phân loại, dự đoán, phân cụm và xử lý ngôn ngữ tự nhiên.

Môn nhập môn học máy cung cấp cho sinh viên kiến thức cơ bản về học máy. Môn học này bao gồm các chủ đề sau:

- Các khái niệm cơ bản về học máy: Định nghĩa học máy, các loại học máy, các thuật toán học máy phổ biến.

- Tiền xử lý dữ liệu: Làm sạch dữ liệu, chuyển đổi dữ liệu, chuẩn hóa dữ liệu.

- Xây dựng mô hình học máy: Chọn mô hình, đào tạo mô hình, đánh giá mô hình.

- Ứng dụng học máy: Một số ứng dụng của học máy trong thực tế.

Học máy có thể được ứng dụng trong nhiều lĩnh vực khác nhau, bao gồm:

- Phân loại: Phân loại hình ảnh, phân loại văn bản, phân loại âm thanh, v.v.

- Dự đoán: Dự đoán giá cổ phiếu, dự đoán doanh thu, dự đoán điểm thi, v.v.

- Phân cụm: Phân cụm khách hàng, phân cụm sản phẩm, phân cụm dữ liệu, v.v.

- Xử lý ngôn ngữ tự nhiên: Phân tích cảm xúc, dịch tự động, nhận dạng ngôn ngữ, v.v.

Môn nhập môn học máy là một môn học quan trọng đối với sinh viên theo học các ngành liên quan đến trí tuệ nhân tạo. Môn học này cung cấp cho sinh viên kiến thức và kỹ năng cần thiết để áp dụng học máy để giải quyết các vấn đề thực tế.

1.2 Đề tài nghiên cứu

- Tìm hiểu, so sánh các phương pháp Optimizer trong huấn luyện mô hình học máy;

- Tìm hiểu về Continual Learning và Test Production khi xây dựng một giải pháp học máy để giải quyết một bài toán nào đó.

CHƯƠNG 2: ĐỀ TÀI BÁO CÁO

**Câu hỏi 1:** Tìm hiểu, so sánh các phương pháp Optimizer trong huấn luyện mô hình học máy

**Trả lời:**

Trong huấn luyện mô hình học máy, phương pháp tối ưu hóa (Optimizer) là một thuật toán giúp tìm ra các tham số tối ưu của mô hình, sao cho mô hình có thể dự đoán chính xác nhất. Các tham số tối ưu của mô hình là các giá trị của các trọng số và bias trong mô hình, được xác định sao cho hàm mất mát của mô hình đạt giá trị thấp nhất. Có nhiều phương pháp tối ưu hóa khác nhau, mỗi phương pháp có những ưu và nhược điểm riêng. Đây là một số phương pháp tối ưu hóa phổ biến trong huấn luyện mô hình học máy:

Gradient descent (GD): Đây là phương pháp tối ưu hóa đơn giản nhất, được sử dụng rộng rãi trong nhiều bài toán học máy. Phương pháp này sử dụng đạo hàm của hàm mất mát để cập nhật các tham số của mô hình.

Stochastic gradient descent (SGD): Đây là một biến thể của GD, được sử dụng khi dữ liệu huấn luyện rất lớn. SGD chỉ cập nhật các tham số của mô hình dựa trên một mẫu dữ liệu ngẫu nhiên trong mỗi lần lặp.

Momentum: Phương pháp này sử dụng đạo hàm của hàm mất mát và tốc độ của quá trình cập nhật tham số để cập nhật các tham số của mô hình. Momentum giúp quá trình tối ưu hóa ổn định hơn và tránh bị mắc kẹt ở các điểm cục bộ tối ưu.

AdaGrad: Phương pháp này sử dụng đạo hàm của hàm mất mát và độ lớn của các tham số để cập nhật các tham số của mô hình. AdaGrad giúp quá trình tối ưu hóa nhanh hơn và hiệu quả hơn ở các tham số có độ lớn lớn.

RMSProp: Phương pháp này cũng sử dụng đạo hàm của hàm mất mát và độ lớn của các tham số để cập nhật các tham số của mô hình. RMSProp tương tự như AdaGrad, nhưng có một số cải tiến giúp quá trình tối ưu hóa ổn định hơn.

Adam: Phương pháp này là một biến thể của RMSProp, được kết hợp với momentum để tạo ra một phương pháp tối ưu hóa hiệu quả và ổn định.

|  |  |  |
| --- | --- | --- |
|  | Ưu điểm | Nhược điểm |
| Gradient descent | Đơn giản, dễ hiểu | Có thể bị mắc kẹt ở các điểm cục bộ tối ưu |
| Stochastic gradient descent | Hiệu quả hơn GD khi dữ liệu huấn luyện lớn | Có thể bị dao động |
| Momentum | Ổn định hơn GD và SGD | Có thể bị chậm chạp khi dữ liệu huấn luyện lớn |
| AdaGrad | Nhanh hơn GD và SGD ở các tham số có độ lớn lớn | Có thể bị chậm chạp khi dữ liệu huấn luyện lớn |
| RMSProp | Tương tự như AdaGrad, nhưng ổn định hơn | Có thể bị chậm chạp khi dữ liệu huấn luyện lớn |
| Adam | Hiệu quả và ổn định | Có thể phức tạp hơn các phương pháp khác |

Lựa chọn phương pháp Optimizer phù hợp phụ thuộc vào nhiều yếu tố, bao gồm:

- Kích thước của dữ liệu huấn luyện: Nếu dữ liệu huấn luyện lớn, thì các phương pháp như SGD, AdaGrad, RMSProp hoặc Adam sẽ hiệu quả hơn GD.

- Kích thước của mô hình: Nếu mô hình có nhiều tham số, thì các phương pháp như AdaGrad, RMSProp hoặc Adam sẽ giúp quá trình tối ưu hóa ổn định hơn.

- Hiệu suất tính toán: Một số phương pháp optimizer, như Adam, có thể yêu cầu nhiều tài nguyên tính toán hơn các phương pháp khác.

Khi sử dụng optimizer, cần lưu ý một số vấn đề sau:

- Chọn giá trị của các tham số: Các tham số của optimizer, chẳng hạn như learning rate, có thể ảnh hưởng đến hiệu quả của quá trình tối ưu hóa. Cần lựa chọn giá trị của các tham số phù hợp với bài toán cụ thể.

- Theo dõi quá trình tối ưu hóa: Cần theo dõi quá trình tối ưu hóa để đảm bảo quá trình này đang diễn ra đúng hướng. Nếu quá trình tối ưu hóa bị mắc kẹt ở các điểm cục bộ tối ưu, có thể cần điều chỉnh các tham số của optimizer hoặc thay đổi kiến trúc của mô hình.

**Câu hỏi 2:** Tìm hiểu về Continual Learning và Test Production khi xây dựng một giải pháp học máy để giải quyết một bài toán nào đó.

**Trả lời:**

Continual Learning là một kỹ thuật học máy cho phép mô hình học hỏi và cải thiện hiệu suất của mình theo thời gian, ngay cả khi dữ liệu huấn luyện không thay đổi. Kỹ thuật này có thể được sử dụng để giải quyết các bài toán học máy trong các môi trường thay đổi liên tục, chẳng hạn như thị trường chứng khoán hoặc các hệ thống phát hiện gian lận.

Có nhiều cách tiếp cận khác nhau để thực hiện Continual Learning. Một cách tiếp cận phổ biến là sử dụng Incremental Learning. Trong phương pháp này, mô hình được cập nhật dần dần với dữ liệu mới. Các cập nhật này được thực hiện theo cách sao cho mô hình không quên những gì nó đã học được trước đây.

Một cách tiếp cận khác là sử dụng Lifelong Learning. Trong phương pháp này, mô hình được huấn luyện trên nhiều tập dữ liệu khác nhau, mỗi tập dữ liệu đại diện cho một giai đoạn khác nhau của môi trường. Mô hình học cách chuyển đổi kiến thức từ các tập dữ liệu này sang các tập dữ liệu khác.

Còn đối với Test Production là một kỹ thuật kiểm tra cho phép mô hình học máy được triển khai trong môi trường sản xuất sớm hơn, ngay cả khi mô hình vẫn đang trong quá trình phát triển. Kỹ thuật này có thể được sử dụng để thu thập phản hồi từ người dùng sớm hơn và để cải thiện hiệu suất của mô hình.

Có nhiều cách tiếp cận khác nhau để thực hiện Test Production. Một cách tiếp cận phổ biến là sử dụng A/B testing. Trong phương pháp này, mô hình được triển khai trong hai môi trường khác nhau, một môi trường sử dụng mô hình mới và một môi trường sử dụng mô hình cũ. Sự khác biệt về hiệu suất của hai mô hình được sử dụng để đánh giá mô hình mới.

Một cách tiếp cận khác là sử dụng Shadow testing. Trong phương pháp này, mô hình mới được triển khai trong một môi trường riêng biệt với môi trường sản xuất. Mô hình mới được theo dõi và so sánh với mô hình cũ để đánh giá hiệu suất của nó.

Cả Continual Learning và Test Production đều là những kỹ thuật quan trọng trong xây dựng một giải pháp học máy hiệu quả. Continual Learning giúp mô hình học hỏi và cải thiện hiệu suất của mình theo thời gian, trong khi Test Production giúp thu thập phản hồi từ người dùng sớm hơn và để cải thiện hiệu suất của mô hình.

Một số ví dụ cụ thể về việc sử dụng Continual Learning và Test Production:

Continual Learning:

- Mô hình nhận dạng khuôn mặt được sử dụng để xác thực người dùng trong một hệ thống ngân hàng. Mô hình này có thể được cập nhật với dữ liệu mới về khuôn mặt của người dùng để cải thiện độ chính xác của nó.

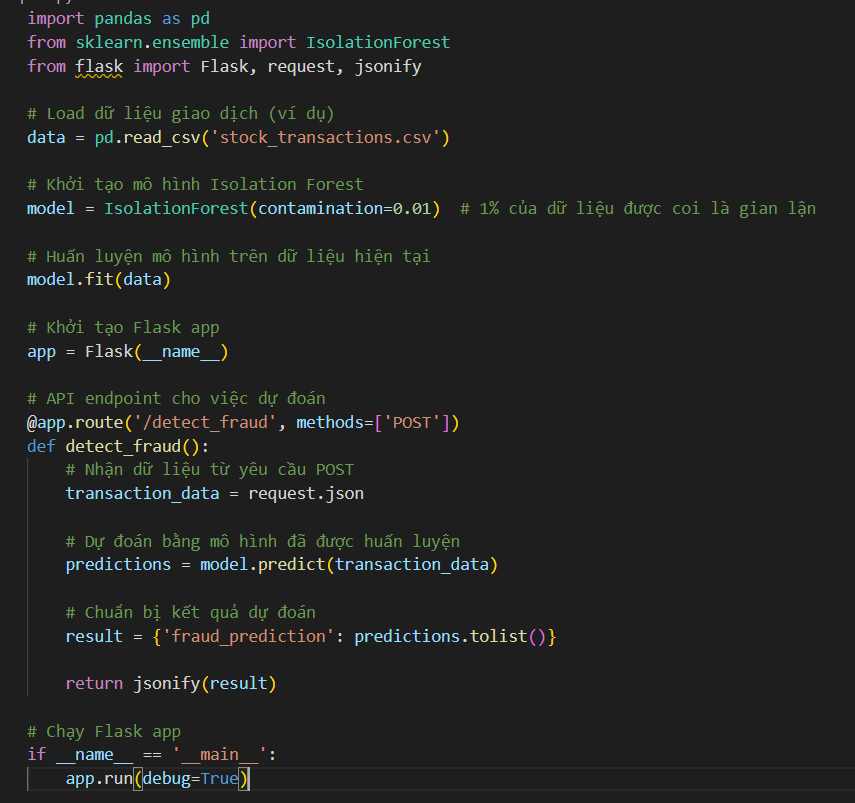
- Mô hình phát hiện gian lận được sử dụng để phát hiện các giao dịch gian lận trên thị trường chứng khoán. Mô hình này có thể được cập nhật với dữ liệu mới về các giao dịch gian lận để cải thiện hiệu quả của nó.

Test Production:

- Một công ty đang phát triển một mô hình dịch tự động. Mô hình này có thể được triển khai trong một môi trường sản xuất sớm để thu thập phản hồi từ người dùng về độ chính xác của bản dịch.

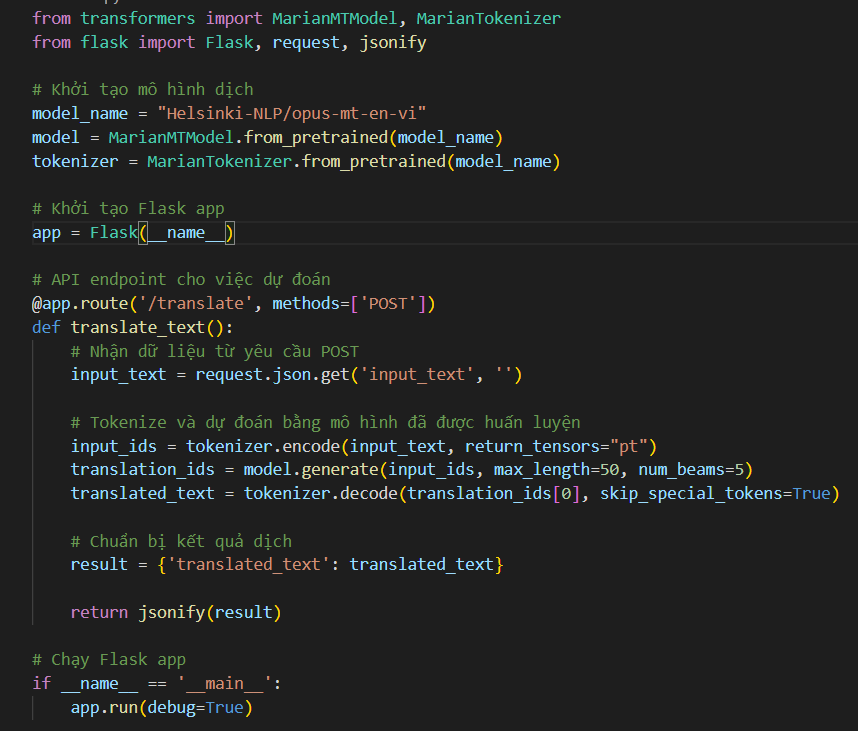
- Một công ty đang phát triển một mô hình phân tích thị trường. Mô hình này có thể được triển khai trong một môi trường sản xuất sớm để thu thập phản hồi từ các nhà đầu tư về độ chính xác của các dự đoán của nó.

Dưới đây là một ví dụ đơn giản khi sử dụng phương pháp Continual Learning để phát hiện gian lận được sử dụng để phát hiện các giao dịch gian lận trên thị trường chứng khoán. Thực tế cần có sự tinh chỉnh và mở rộng tùy thuộc vào đặc điểm của dữ liệu và các yêu cầu cụ thể.



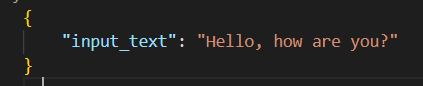
Hình 1: Mô hình Continual Learning

Dưới đây là một ví dụ đơn giản khi sử dụng mô hình Test Production để phát triển một mô hình dịch tự động. Để triển khai một mô hình dịch tự động Machine Translation trong môi trường sản xuất, bạn có thể sử dụng thư viện transformers của Hugging Face, cùng với Flask để xây dựng một API đơn giản.



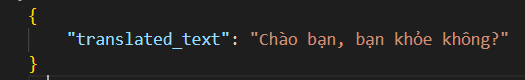
Hình 2: Mô hình Test Production

Gửi yêu cầu POST đến “http://127.0.0.1:5000/translate” với dữ liệu JSON:



Hình 3: Dữ liệu JSON

Kết quả thu được sẽ trả về một JSON chứa văn bản được dịch:



Hình 4: Kết quả văn bản