**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*: **Ngô Hoàng Khôi – 52000676**

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*: **Ngô Hoàng Khôi – 52000676**

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)*

*Khôi*

*Ngô Hoàng Khôi*

**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)*

*Khôi*

*Ngô Hoàng Khôi*

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.

- 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.

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 optimizer là một yếu tố quan trọng quyết định tốc độ học và hiệu suất của mô hình. Dưới đây là một số phương pháp optimizer phổ biến được sử dụng trong deep learning và so sánh giữa chúng

Gradient descent (GD): là một thuật toán tối ưu hóa được sử dụng rộng rãi trong machine learning để tìm giá trị nhỏ nhất của một hàm mất mát. Mục tiêu của thuật toán này là điều chỉnh các tham số của mô hình (thường được biểu diễn bởi vector trọng số **W**) sao cho giá trị của hàm mất mát là nhỏ nhất.

Stochastic gradient descent (SGD): là một biến thể của thuật toán Gradient Descent (GD) được sử dụng rộng rãi trong quá trình đào tạo mô hình học máy. Trong SGD, thay vì tính toán đạo hàm của toàn bộ tập dữ liệu đào tạo như trong GD, chỉ một mẫu dữ liệu (hoặc một số mẫu nhỏ được gọi là mini-batch) được sử dụng để tính toán đạo hàm ở mỗi bước cập nhật.

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.

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.

- Luôn theo dõi quá trình tối ưu hóa để đảm bảo quá trình mô hình học không bị mắc kẹt ở các điểm cực tiểu. Có thể điều chỉnh các tham số để phù hợp với data

**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 (còn được gọi là Incremental Learning, Life-long Learning) là một khái niệm để học một mô hình cho một số lượng lớn các tác vụ một cách tuần tự mà không quên kiến thức đã thu được từ các tác vụ trước đó, nơi mà dữ liệu trong các tác vụ cũ không còn sẵn sàng nữa trong quá trình đào tạo các tác vụ mới1. Mục tiêu chung của Continual Learning là đảm bảo sự cân đối giữa ổn định và linh hoạt cũng như khả năng tổng quát hóa intra/inter-task phù hợp trong bối cảnh hiệu quả về nguồn lực

Về Test Production trong học máy, việc phát triển một mô hình học máy hoặc học sâu dường như là một nhiệm vụ tương đối đơn giản. Nó thường bao gồm nghiên cứu, thu thập và tiền xử lý dữ liệu, trích xuất đặc trưng, xây dựng và đào tạo mô hình, đánh giá và suy luận. Tuy nhiên, khó khăn xuất hiện khi chúng ta muốn đưa mô hình vào sản xuất trong thế giới thực. Mô hình thường không hoạt động tốt như trong giai đoạn đào tạo và đánh giá. Điều này chủ yếu xảy ra do sự thay đổi khái niệm hoặc dữ liệu và các vấn đề liên quan đến tính toàn vẹn của dữ liệu3. Do đó, việc kiểm tra một mô hình ML trở nên rất quan trọng để chúng ta có thể hiểu được điểm mạnh và điểm yếu của nó và hành động phù hợp

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 2 phương pháp đều là các kỹ thuật giúp cho mô hình học máy hoạt động tốt hơn. Continual Learning giúp mô hình cải thiện hiệu suất theo thời gian. Test Production thu thập các phản hồi từ người dùng để tăng hiệu suất

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.

Ví dụ:

Contnual learning : netflix là một hệ thống điển hình của mô hình này. Hệ thống này đề xuất chương trình ngay sau khi tập cuối cùng kết thúc. Mô hình này cần được đào tạo định kỳ vì có phim mới.

Test production: Google đã đề xuất một quy trình kiểm tra các mô hình học máy như phân loại thư spam.