**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**

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

*Người thực hiện*: **Đỗ Duy Linh – 51702222**

Lớp **: 17050281**

Khoá  **: 21**

**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**

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

*Người thực hiện*: **Đỗ Duy Linh – 51702222**

Lớp **: 17050281**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

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 thầy mà 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)*

*Linh*

*Đỗ Duy Linh*

**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)*

*Linh*

*Đỗ Duy Linh*

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 1](#_Toc154233706)

[TÓM TẮT 3](#_Toc154233707)

[MỤC LỤC 4](#_Toc154233708)

[CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN 5](#_Toc154233709)

[1.1 Giới thiệu môn học 5](#_Toc154233710)

[1.2 Đề tài nghiên cứu 5](#_Toc154233711)

[CHƯƠNG 2: ĐỀ TÀI BÁO CÁO 6](#_Toc154233712)

[**Câu hỏi 1:** 6](#_Toc154233713)

[Câu hỏi 2: 7](#_Toc154233714)

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**

1. Gradient descent (GD):

Gradient descent là một thuật toán tối ưu hóa được sử dụng để giảm thiểu chi phí trong học máy và tối ưu hóa tốc độ tính toán. Thuật toán này hoạt động bằng cách lặp lại việc di chuyển theo hướng giảm nhanh nhất, được xác định bởi âm của đạo hàm của hàm chi phí. Các bước chính bao gồm khởi tạo tham số, tính toán gradient, cập nhật tham số và lặp lại quá trình này cho đến khi hội tụ hoặc đạt đến số lần lặp đã xác định trước. Việc chọn tỷ lệ học (learning rate) là quan trọng để đảm bảo thuật toán hội tụ một cách hiệu quả. Có các biến thể như Batch Gradient Descent, Stochastic Gradient Descent và Mini-Batch Gradient Descent.

1. Stochastic gradient descent (SGD):

Stochastic Gradient Descent (SGD) là một biến thể của thuật toán gradient descent trong máy học. Quá trình tối ưu hóa bao gồm việc lặp lại qua dữ liệu, chọn ngẫu nhiên một điểm dữ liệu ở mỗi bước, tính gradient chỉ dựa trên điểm dữ liệu đó, và cập nhật tham số để giảm giá trị hàm chi phí. Lợi ích của SGD bao gồm khả năng tránh được tối thiểu cục bộ và khả năng tổng quát hóa tốt hơn, đặc biệt khi đối mặt với dữ liệu lớn. Tuy nhiên, sự ngẫu nhiên có thể làm tăng độ dao động và làm giảm tốc độ hội tụ, vì vậy các biến thể như Mini-Batch Gradient Descent thường được sử dụng để kết hợp tính ngẫu nhiên và hiệu suất tính toán.

1. Mini-batch gradient descent:

Mini-batch Gradient Descent là một biến thể của thuật toán gradient descent trong máy học, sử dụng một tập con nhỏ và ngẫu nhiên của dữ liệu (mini-batch) để tính gradient và cập nhật tham số của mô hình. Khác với Batch Gradient Descent (sử dụng toàn bộ dữ liệu) và Stochastic Gradient Descent (sử dụng chỉ một điểm dữ liệu), Mini-batch Gradient Descent kết hợp lợi ích của cả hai, giúp tăng tốc độ hội tụ và giảm độ dao động của hàm chi phí. Kích thước của mini-batch là một siêu tham số quan trọng cần được lựa chọn phù hợp theo từng tình huống.

1. AdaGrad:

Adagrad (Adaptive Gradient Algorithm) là một thuật toán tối ưu hóa được thiết kế để tự động điều chỉ tỷ lệ học cho mỗi tham số của mô hình dựa trên lịch sử của gradient tương ứng. Nó sử dụng biến tích lũy gradient để điều chỉnh tỷ lệ học, giúp thích ứng linh hoạt với độ quan trọng của từng tham số. Tuy nhiên, một nhược điểm là với thời gian, biến tích lũy gradient có thể trở nên lớn, dẫn đến giảm đột ngột của tỷ lệ học. Các biến thể như RMSprop và Adam đã được phát triển để cải thiện vấn đề này.

1. RMSProp:

RMSProp (Root Mean Square Propagation) là một thuật toán tối ưu hóa sử dụng trung bình động bình phương của gradient để điều chỉnh tỷ lệ học cho mỗi tham số của mô hình. Thay vì sử dụng biến tích lũy gradient như Adagrad, RMSprop cập nhật biến tích lũy bằng cách nhân động bình phương của gradient với một hệ số giảm. Điều này giúp giảm đột ngột của tỷ lệ học và làm cho quá trình tối ưu hóa ổn định hơn, đồng thời tránh vấn đề với các tham số quá quan trọng. Công thức cập nhật tính toán tỷ lệ học cũng được điều chỉnh để tránh chia cho 0. RMSprop cải thiện nhược điểm của Adagrad và là một trong những thuật toán quan trọng trong tối ưu hóa gradient.

1. Adam:

Adam (Adaptive Moment Estimation) là một thuật toán tối ưu hóa gradient trong học máy và deep learning. Nó kết hợp ý tưởng từ RMSprop và Momentum để cải thiện quá trình tối ưu hóa. Adam sử dụng hai biến, *m* và *v*, để theo dõi động và trung bình động bình phương của gradient. Các biến này được điều chỉnh để khắc phục bias ban đầu và sau đó được sử dụng để cập nhật tham số của mô hình. Adam thường được coi là một trong những thuật toán tối ưu hóa hiệu quả và linh hoạt, phổ biến trong nhiều ứng dụng deep learning.

1. Nadam:

Nadam là tên kết hợp từ NAG (Nesterov Accelerated Gradient) và Adam

(Adaptive Moment Estimation).

Nadam là một thuật toán tối ưu hóa gradient trong học máy, kết hợp cả Momentum và RMSprop, tương tự như Adam. Nó sử dụng hai biến *m* và *v* để theo dõi gradient và bình phương gradient, với công thức cập nhật đặc biệt liên quan đến Nesterov Momentum. Nadam giúp khắc phục một số vấn đề của Adam trong một số tình huống và được đánh giá là có thể cung cấp hiệu suất tốt hơn trong quá trình tối ưu hóa mô hình trong học máy và deep learning.

Bảng so sánh ưu điểm và nhược điểm của các thuật toán trên:

|  |  |  |
| --- | --- | --- |
|  | Ưu điểm | Nhược điểm |
| Gradient Descent | Dễ triển khai, hiệu quả với dữ liệu lớn | Cần lựa chọn learning rate phù hợp, không hiệu quả trong dữ liệu thưa, nhạy cảm với dữ liệu nhiễu |
| Stochastic Gradient Descent | Dễ triển khai, tốn ít bộ nhớ | Có thể hội tụ chậm, dễ lỗi ở các điểm cực tiểu |
| Mini-batch Gradient Descent | Kết hợp ưu điểm của SGD và BGD | Cần lựa chọn mini-batch phù hợp |
| Adagrad | Thích ứng với từng tham số phù hợp cho dữ liệu (sparse data) | Có thể giảm learning rate quá nhanh |
| RMSProp | Giảm tác động của vấn đề tích tụ bình phương độ dốc của Adagrad | Cần lựa chọn learning rate phù hợp |
| Adam | Kết hợp ưu điểm của RMS prop và Momentum | Cần lựa chọn các siêu tham số như β1, β2 và learning rate |
| Nadam | Kết hợp ưu điểm của Adam và Nesterov Momentum | Cần lựa chọn các siêu tham số như Adam |

**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:**

1. Continual Learning:

Continual Learning là một khái niệm trong lĩnh vực học máy, chú trọng vào khả năng của mô hình học máy học liên tục từ dữ liệu mới mà không quên đi kiến thức đã học trước đó. Trong một số trường hợp, mô hình có thể phải đối mặt với dữ liệu mới, các tác vụ mới, hoặc thậm chí là một môi trường mới mà không được huấn luyện lại từ đầu.

Các thách thức chính của Continual Learning bao gồm:

* Lỗi Quên (Catastrophic Forgetting): Khi mô hình học điều đó mới, nó có thể quên đi kiến thức đã học trước đó.
* Interference (Nhiễu): Các kiến thức mới có thể làm ảnh hưởng đến kiến thức cũ và ngược lại.
* Resource Constraints (Hạn chế tài nguyên): Mô hình cần tối ưu để sử dụng tài nguyên (bộ nhớ, thời gian huấn luyện) một cách hiệu quả khi học liên tục.

Để giải quyết những thách thức này, có nhiều phương pháp Continual Learning, bao gồm sử dụng bộ nhớ (memory), huấn luyện đồng thời trên nhiều tác vụ, và các phương pháp giảm quên.

1. Test Production:

Test Production là quá trình tạo ra bộ dữ liệu kiểm thử để đánh giá hiệu suất của mô hình học máy. Điều này bao gồm việc chia tách dữ liệu thành bộ huấn luyện và bộ kiểm thử, để đảm bảo rằng mô hình được kiểm thử trên dữ liệu mà nó chưa thấy trong quá trình huấn luyện.

Các nguyên tắc quan trọng khi tạo bộ kiểm thử bao gồm:

* Phân phối Tương tự: Bộ kiểm thử nên được chọn sao cho phân phối của nó tương tự với phân phối của dữ liệu thực tế mà mô hình có thể gặp phải trong ứng dụng thực tế.
* Khả năng Đại diện: Bộ kiểm thử cần đại diện cho các trường hợp đặc biệt và biên của dữ liệu một cách đầy đủ.
* Không Trùng Lặp**:** Đảm bảo rằng các mẫu trong bộ kiểm thử không trùng lặp với mẫu trong bộ huấn luyện để đảm bảo sự khách quan khi đánh giá hiệu suất.
* Kiểm Thử Chuẩn Xác: Bộ kiểm thử nên được thiết kế sao cho mô hình có thể đối mặt với các trường hợp khó và đòi hỏi độ chính xác cao.

Quá trình Test Production giúp đánh giá đối với mô hình có hiệu suất tốt trên bộ dữ liệu kiểm thử có tính đại diện và độ đa dạng, cung cấp thông tin quan trọng về khả năng tổng quát hóa của mô hình đối với dữ liệu mới.