TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KỲ MÔN NHẬP MÔN HỌC MÁY**

**BÁO CÁO CUỐI KỲ**

*Người hướng dẫn*: **TS LÊ ANH CƯỜNG**

*Người thực hiện*: **TRẦN ĐỖ TRUNG KIÊN – 52000679**

**HUỲNH HỒNG SƠN – 52000795**

**NGUYỄN THÀNH HUY - 52100895**

**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 KỲ MÔN NHẬP MÔN HỌC MÁY**

**BÁO CÁO CUỐI KỲ**

*Người hướng dẫn*: **TS LÊ ANH CƯỜNG**

*Người thực hiện*: **TRẦN ĐỖ TRUNG KIÊN – 52000679**

**HUỲNH HỒNG SƠN – 52000795**

**NGUYỄN THÀNH HUY - 52100895**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn đến thầy Lê Anh Cường vì đã mang tới cho chúng em những kiến thức và chỉ dẫn trên những tiết học lý thuyết. Với những kiến thức này chúng em đã có thêm kinh nghiệm để hoàn thành báo cáo cuối kì này. Cuối cùng, chúng em xin kính chúc thầy dồi dào sức khỏe và nhiều niềm vui trong cuộc sống.

**BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng em xin cam đoan đây là báo cáo của chúng em và được sự hướng dẫn của TS. 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 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 chúng em xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng em gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 22 tháng 12 năm 2023*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trần Đỗ Trung Kiên*

*Huỳnh Hồng Sơn*

*Nguyễn Thành Huy*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Trong bài báo cáo này, chúng em sẽ phân tích và đưa ra các giải pháp của chúng em dựa trên các đề tài, yêu cầu thầy giao cho.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc154263115)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc154263116)

[TÓM TẮT iv](#_Toc154263117)

[MỤC LỤC 1](#_Toc154263118)

[CHƯƠNG 1 – TÌM HIỂU KHÁI NIỆM 3](#_Toc154263119)

[1.1 Phương pháp Optimzers (Tối ưu hoá mô hình học máy) 3](#_Toc154263120)

[1.2 Gradient Descent (GD) 3](#_Toc154263121)

[1.3 Stochastic Gradient Descent (SGD): 4](#_Toc154263122)

[1.4 Adaptive Gradient Descent (Adagrad): 4](#_Toc154263123)

[1.5 Adaptive Moment Estimation (Adam): 5](#_Toc154263124)

[2.1 Continual Learning 7](#_Toc154263125)

[2.1.1 Khái niệm 7](#_Toc154263126)

[2.1.2 Các bước triển khai 8](#_Toc154263127)

[2.1.3 Bất cập 8](#_Toc154263128)

[2.1.4 Tổng kết 9](#_Toc154263129)

[3.1 Test Production 10](#_Toc154263130)

[3.1.1 Khái niệm 10](#_Toc154263131)

[3.1.2 Cách thức doanh nghiệp tiếp cận vấn đề 10](#_Toc154263132)

[3.1.3 Tổng kết 12](#_Toc154263133)

[CHƯƠNG 2 – THỰC NGHIỆM MÔ HÌNH 14](#_Toc154263134)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC CHỮ VIẾT TẮT**

GD Gradient Descent

SGD Stochastic Gradient Descent

CHƯƠNG 1 – TÌM HIỂU KHÁI NIỆM

Đây là báo cáo cuối kỳ môn học “Nhập môn học máy”, mục tiêu để sinh viên chúng em có cơ hội tổng hợp, khái quát những kiến thức đã được giảng viên giảng dạy, qua đó tiếp thu và vận dụng những kiến thức đó vào những bài toán khác nhau trong học tập và yêu cầu.

* 1. Phương pháp Optimzers (Tối ưu hoá mô hình học máy)

Trong Deep Learning, các Optimizers chính là các thuật toán thường được sử dụng để đưa ra các tham số phù hợp, làm cho kết quả đầu ra của một mô hình học máy bất kỳ bao gồm ít sai số nhất có thể, từ đó tăng độ chính xác.

Nói chính xác hơn, các Optimizers giúp điều chỉnh các tham số của mô hình trong lúc huấn luyện để giảm thiểu hàm hao hụt (giảm thiểu khoảng cách chênh lệch nhau giữa giá trị huấn luyện và giá trị dự đoán)

Các Optimizers tiếp cận việc này với nhiều phương pháp khác nhau, mỗi thuật toán lại có những quy tắc cập nhật trọng số riêng, nhưng nhìn chung đều có một mục đích là tìm ra được các tham số tối ưu để cải thiện hiệu suất huấn luyện mô hình.

1.2 Gradient Descent (GD)



Có thể xem đây chính là optimizer phổ biến nhất trong tất cả.

Thuật toán này đơn giản là sử dụng đại số để liên tục điều chỉnh giá trị để đạt đến điểm cực tiểu (Local minimum), xuất phát từ một điểm mà chúng ta coi là gần với nghiệm của bài toán, sau đó dùng một phép toán lặp để tiến dần đến điểm cần tìm, cho đến khi đạo hàm gần với 0

+ Điểm mạnh: Có thể sử dụng trong hầu hết các trường hợp, dễ hiểu cho người mới

- Điểm trừ: Khi lượng dữ liệu trở nên khổng lồ, việc tính toán dựa trên thuật toán này rất tốn chi phí vì mỗi lần cập nhập trọng số thì thuật toán sẽ quét qua cả dataset

1.3 Stochastic Gradient Descent (SGD):



Thuật toán này tổng quan có thể xem như là một thuật toán GD cải tiến hơn thuật toán ở trên

Thay vì phải lặp lại hết tất các dataset cho mỗi lần lặp để cập nhật trọng số, SGD chỉ cần lấy ngẫu nhiên một tập hợp các dữ liệu để tính toán trọng số, SGD thì mỗi epoch ứng với N lần cập nhật θ với N là số điểm dữ liệu

Bởi vì lượng dữ liệu giảm đi mỗi lần cập nhật trọng số giảm đi, nên số lần lặp lại tăng lên để duyệt qua hết dữ liệu, từ đó làm tăng thời gian tính toán

+ Điểm cộng: Khắc phục được phần nào chi phí tính toán của thuật toán Gradient Descent bằng một cách tiếp cận dataset khác

- Điểm trừ:

* Để lặp lại đủ số lần sao cho duyệt qua được hết dataset cũng sẽ làm tăng thời gian tính toán của thuật toán này
* Tuy thuật toán này hội tụ tại cực tiểu khá là nhanh, nhưng đường đi của nó lại nhiễu hơn (zigzag) nhiều so với độ mượt mà của Gradient Descent

1.4 Adaptive Gradient Descent (Adagrad):

Thuật toán tối ưu này sử dụng các learning rate (tốc độ học) khác nhau ở mỗi lần lặp

Sự thay đổi của các learning rate phụ thuộc vào sự chênh lệch giữa các tham số trong quá trình huấn luyện, tham số chênh lệch càng nhiều, tốc độ học thay đổi càng ít.



+ Điểm mạnh:

* Phù hợp với thực tế vì các dataset thường sẽ phân vùng mạnh mẽ ở một vài đặc trưng bất kì của dataset đó
* Chúng ta không phải điều chỉnh learning rate một cách thủ công, so với các thuật toán GD nếu ta không tự mình tinh chỉnh learning rate thì có thể nghiệm đi qua đi lại quá xa so với điểm cực tiểu

- Điểm trừ: Vì các learning rate thay đổi quá nhanh và liên tục, sẽ có thời điểm learning rate trở nên cực kì nhỏ, sẽ xảy ra trường hợp mô hình không thể học thêm nữa, từ đó độ chính xác sẽ bị suy giảm

1.5 Adaptive Moment Estimation (Adam):

Thuật toán này được xem như là một phiên bản mở rộng của SGD, được thiết kế để giúp cập nhật trọng số trong các mô hình neural network (là thuật toán SGD nhưng với cấu trúc Momentum kết hợp với thuật toán RMSProp)

Thuật toán này điều chỉnh learning rate thích ứng với từng trọng số của neuron

Tính toán dựa trên trung bình cộng và phương sai của các điểm cực trị, giúp cho thuật toán tối ưu này có thể di chuyển đúng hướng đến cực tiểu và điều chỉnh learning rate với mỗi tham số dựa trên phương sai giữa các điểm cực trị

Trung bình cộng: 

Phương sai: 

Cập nhật trọng số bằng cách: 

+ Điểm mạnh:

* Rất tốt khi áp dụng cho các dữ liệu có độ nhiễu cao hay phân vùng mạnh
* Có thể dự đoán bias và canh chỉnh lại learning rate cho phù hợp

- Điểm trừ:

* Vận hành tốn kém

2.1 Continual Learning

2.1.1 Khái niệm

Các giải pháp Machine Learning truyền thống thường đưa ra các data dự đoán tương lai dựa trên hoặc gần tương tự với các data mà mô hình đã được huấn luyện, thật ra không phải lúc nào cũng là như thế.

Continual Learning từ đó được nghiên cứu để dần giải quyết được vấn đề này. Mô hình sẽ được huấn luyện dựa trên data được cập nhật liên tục.

Mục đích của phương pháp học này là cố gắng bắt chước khả năng tiếp thu kiến thức mới liên tục của con người, giúp mở ra một tương lai mà công nghệ AI có thể tiệm cận được với suy nghĩ con người hơn

A diagram of a process

Description automatically generated with medium confidence

2.1.2 Các bước triển khai

Để dễ hình dung về tiêu chí của Continual Learning, ví dụ một công ty sẽ có các tệp khách hàng mở rộng liên tục, và các dữ liệu sẽ liên tục được thêm mới vào, giả sử là dữ liệu của mỗi cá nhân là không giống nhau, khi đó hệ thống sẽ phải thích ứng liên tục các dữ liệu mới này hòng đưa ra các dự đoán tối ưu nhất.

Một cách tiếp cận mà em tìm hiểu được để triển khai một mô hình Continual Learning đó là quá trình định kỳ tái huấn luyện mô hình

Quá trình này có thể bao gồm các bước sau

1. **Logging**: Thêm vào các dữ liệu mới chuẩn bị cho quá trình tái huấn luyện, tại điểm kết thúc của bước này, ta sẽ có một loạt các data chưa gắn nhãn và có tiềm năng để làm dữ liệu huấn luyện
2. **Curation**: Quy tắc của bước này là phải chọn lọc ra được các data sẽ được ưu tiên gắn nhãn trước và tái huấn luyện. Tại điểm kết thúc của bước này, ta sẽ có một cơ sở dữ liệu đã được gắn nhãn đầy đủ và chuẩn bị cho việc tái huấn luyện
3. **Retraining Trigger**: Chọn ra được một lịch trình tối ưu để tái huấn luyện mô hình sẽ là điều cần lưu ý ở bước này, có thể là 1 lần/tuần hay 1x lần/tháng, …
4. **Dataset Formation**: Từ bộ dữ liệu được huấn luyện, mô hình mới sẽ được hình thành. Điểm cần lưu ý ở bước này là xác định được mô hình nào sẽ tối ưu để làm mô hình mới cho các kỳ huấn luyện khác trong tương lai

2.1.3 Bất cập

Sơ qua ta có thể giả định được ***một số tình huống xấu*** có thể xảy ra trong quá trình triển khai các bước trên:

* **Dữ liệu quá lớn**: Khi luồng dữ liệu mới quá lớn thì không phải lúc nào ta cũng có thể thêm vào huấn luyện hết được mà sẽ phải linh hoạt
* **Gắn nhãn**: Khi các dữ liệu mới phát sinh, việc gắn nhãn có thể trở nên khá phức tạp nếu quy mô mô hình thay đổi dựa trên các hướng đi mới của công ty để thu hút các người dùng mới
* **Chi phí**: Có thể dễ dàng nhận ra rằng chi phí để vận hành mô hình học máy này khá tốn kém khi chúng ta phải duy trì việc học này với các dữ liệu mới liên tục để đưa ra kết quả, nhất là khi dữ liệu ngày càng lớn
* **Thông số**: Khi một mô hình mới được hình thành sau quá trình tái huấn luyện, ta sẽ không biết được ngay liệu mô hình mới với các dữ liệu mới đó có thật sự tối ưu hay không, và các thông số đo lường độ chính xác sẽ khó để áp dụng hơn trong quá trình này, cần thêm nhiều sự giám sát
* **Rủi ro**: Hiển nhiên việc tái huấn luyện mô hình cũng tiềm ẩn nhiều rủi ro nhất định, khi dữ liệu trở nên lớn hơn, tỷ lệ phát sinh các dự đoán thiếu sự chính xác sẽ ngày một cao hơn.

2.1.4 Tổng kết

Tổng kết lại, Continual Learning vẫn còn là một đề tài có khá ít tài liệu nghiên cứu và các tài liệu hiện tại chưa đủ để cho em thấy cái nhìn sâu hơn của phương pháp học máy này. Những gì phương pháp này mang lại bản thân em thấy sẽ rất khả quan trong tương lai, nếu máy tính có thể suy nghĩ và nạp kiến thức mới liên tục, sẽ hỗ trợ con người tốt đến nhường nào.

3.1 Test Production

3.1.1 Khái niệm

Thường khi kiểm tra độ chính xác của một mô hình học máy, mọi người sử dụng các thông số đo lường độ chính xác của các mô hình thông qua các lần học khác nhau và các thuật toán Optimizers. Nhưng trên thực tế, một mô hình học máy hiệu quả còn phụ thuộc và mô hình kinh doanh của doanh nghiệp, lượng dữ liệu được cập nhật, …

3.1.2 Cách thức doanh nghiệp tiếp cận vấn đề

Ở một vài công ty như GreenSteam, MonoHQ, Arkera, họ đều có các cách thức để test một mô hình học máy trên quy mô Production khác nhau

1. **GreenSteam**: Kết hợp giữa tự động hoá testing và thủ công

A close-up of a website

Description automatically generated

Trong quy trình kiểm định này, GreenSteam đã sử dụng Jenkins và smoke tests, thiết lập một môi trường Production Testing, kết quả thu được sẽ đưa qua một pipeline và qua đó, một chuyên gia sẽ xem xét qua các kết quả này, chuyên gia này sẽ phân tích, xác nhận rồi sau đó mới đưa mô hình vào vận hành rộng rãi

1. **MonoHQ**: Phân loại các Transaction statements

A blue background with white text

Description automatically generated

Theo như MonoHQ, bộ phận nghiên cứu trước khi triển khai mô hình trên Production, mô hình sẽ trải qua ba bài kiểm tra;

* Dự đoán phân tán
* Mức độ sai số
* Độ trễ

Các bài test trên để cả nhóm kiểm tra độ tin cậy của các dự đoán mà mô hình cho ra. Nhóm nghiên cứu kiểm tra độ trễ bằng cách tính toán thời gian trung bình của mô hình phản hồi với toàn bộ dữ liệu. Nếu độ trễ nằm ở ngưỡng cho phép, mô hình sẽ sẵn sàng để triển khai trên Production

A/B Testing cũng được cân nhắc khi nhóm cần xác định xem phiên bản nào của mô hình sẽ là tối ưu nhất để triển khai, giúp cả nhóm tránh khỏi việc tạo ra nhiều mô hình khác nhau nhưng các mô hình này lại khó khăn trong việc bảo trì và không giúp ích gì cho nhu cầu người dùng.

1. **Arkera**

A screenshot of a video chat

Description automatically generated

Các dự án Machine Learning của công ty thường được kiểm tra bởi 2 đội nhóm

* Nhóm đảm nhiệm triển khai các Unit và Integration Test (nhóm kỹ thuật)
* Nhóm đảm nhiệm việc validate các mô hình và phân tích các thông số đo lượng (nhóm dữ liệu)

Nhóm kỹ thuật sẽ kiểm tra xem mô hình cuối cùng có tương thích tốt với các dịch vụ khác nhau trong Production hay không bằng cách chạy thử trên một mô hình giả (giống cấu trúc với mô hình thật nhưng chỉ đưa ra các kết quả ngẫu nhiên). Mục tiêu của nhóm kỹ thuật sẽ kiểm tra xem quan hệ input – output của mô hình có đưa ra đúng các giá trị mà họ mong đợi và không sinh ra bất kỳ lỗi nào.

Bên cạnh đó, nhóm dữ liệu sẽ thử huấn luyện, kiểm thử mô hình này dựa trên các data thực tế và phân tích các số liệu đo lường, từ đó xác định xem mô hình này có tối ưu và phù hợp cho các hướng đi của công ty hay không.

3.1.3 Tổng kết

Với các ví dụ trên cho thấy rằng các công ty họ nhận ra các thông số đo lường khi triển khai mô hình còn thiếu thuyết phục, nên họ luôn có các phương án khác nhau để củng cố sự ổn định, tối ưu của một mô hình mang lại, đảm bảo mô hình đó có thể xử lý các vấn đề khác nhau có thể gặp phải.

CHƯƠNG 2 – THỰC NGHIỆM MÔ HÌNH

Chúng em quyết định chọn file diabetes\_012\_health\_indicators\_BRFSS2021.csv để làm dataset cho các yêu cầu ở Phần 2 này: [Link](https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset)

Các yêu cầu của phần 2 đã được chúng em tổng hợp tại github: [Link](https://github.com/ThanhWii/final-machine-learning.git)

**TÀI LIỆU THAM KHẢO**

**Tiếng Anh**

1. Stephan Oladele, ML Model Testing: 4 Teams Share How They Test Their Models, Neptune, [Link](https://neptune.ai/blog/ml-model-testing-teams-share-how-they-test-models), truy cập ngày 18/12/2023
2. Sanket Doshi, Various Optimization Algorithms For Training Neural Network, Towards Data Science, [Link](https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6), truy cập ngày 18/12/2023
3. Debjeet Asitkumar Das, Optimizers in Deep Learning, Medium, [Link](https://medium.com/analytics-vidhya/this-blog-post-aims-at-explaining-the-behavior-of-different-algorithms-for-optimizing-gradient-46159a97a8c1), truy cập ngày 18/12/2023