TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KỲ MÔN**

**MARCHINE LEARNING**

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

*Người thực hiện*: **NGÔ MỘNG HOÀN – 521H0231**

Lớp **: 21H50301**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

Lời đầu tiên, em xin trân trọng cảm ơn giảng viên ***Lê Anh Cường*** - người đã trực tiếp chỉ bảo, hướng dẫn em trong quá trình hoàn thành bài tiểu luận này.

Em cũng xin được gửi lời cảm ơn đến quý thầy, cô giáo trường ***Đại học Tôn Đức Thắng***, đặc biệt là các thầy, cô khoa ***Công Nghệ Thông Tin***. - những người đã truyền lửa và giảng dạy kiến thức cho em suốt thời gian qua.

Mặc dù đã có những đầu tư nhất định trong quá trình làm bài song cũng khó có thể tránh khỏi những sai sót, em kính mong nhận được ý kiến đóng góp của quý thầy cô để bài tiểu luận được hoàn thiện hơn.

Em xin chân thành cảm ơn!

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi 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 tôi 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 tôi 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)*

*A signature on a white background

Description automatically generated*

*Ngô Mộng Hoàn*

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)

1. **Optimizer**

A graph of a graph with a rainbow colored curve

Description automatically generated with medium confidence

**1.1 Optimizer là gì, tại sao phải dùng nó?**

Optimizer là một thành phần quan trọng trong quá trình huấn luyện mô hình trong machine learning. Nó được dùng để điều chỉnh các tham số của mô hình sao cho mô hình có thể học từ dữ liệu huấn luyện một cách hiệu quả nhất, thông qua việc tối ưu hóa hàm mất mát.

Mục tiêu chính của việc sử dụng optimizer là tìm ra giá trị tối thiểu hoặc gần tối thiểu nhất của hàm mất mát. Điều này cực kỳ quan trọng vì nó giúp mô hình học từ dữ liệu một cách hiệu quả, làm tăng độ chính xác của dự đoán và cải thiện khả năng tổng quát hóa của mô hình.

Chúng ta sử dụng optimizer trong machine learning vì:

***Điều chỉnh hàm mất mát tối ưu*:** Giúp thay đổi tham số mô hình để tìm giá trị tối ưu của hàm mất mát từ dữ liệu huấn luyện.

***Tăng tốc quá trình huấn luyện:*** Optimizer thông minh hóa việc cập nhật tham số, giúp mô hình học nhanh hơn và giảm thời gian cần cho quá trình này.

***Tránh rơi vào điểm cực tiểu cục bộ:*** Một số phương pháp tối ưu hóa có khả năng thoát ra khỏi các điểm cực tiểu cục bộ, tiến gần đến điểm cực tiểu toàn cục của hàm mất mát, nâng cao hiệu suất của mô hình.

Tận dụng optimizer trong machine learning giúp mô hình học hiệu quả từ dữ liệu, tối ưu hiệu suất và cải thiện độ chính xác của việc dự đoán.

**1.2 Các thuật toán tối ưu ?**

**- 1.2.1 Gradient Descent (GD)**

Gradient Descent, một phương pháp thông dụng trong machine learning và các lĩnh vực tương tự, nhằm định vị giá trị nhỏ nhất của một hàm số bằng cách thay đổi các tham số qua các bước lặp.

Thuật toán này hoạt động bằng cách tính toán gradient của hàm số (tức là đạo hàm theo từng tham số), chỉ ra hướng và tỷ lệ thay đổi của hàm số. Tiếp theo, nó di chuyển theo hướng ngược với gradient này để giảm giá trị của hàm số.

Gradient Descent có hai dạng chính:

*Lợi ích:*

Phương pháp cơ bản của Gradient Descent dễ hiểu và đã được áp dụng để tối ưu hóa mạng neural bằng cách cập nhật trọng số sau mỗi chu kỳ.

*Nhược điểm:*

Do đơn giản, Gradient Descent đối mặt với nhiều hạn chế như sự phụ thuộc vào việc khởi tạo ban đầu và tỷ lệ học (learning rate).

Ví dụ, nếu một hàm có hai giá trị tối thiểu toàn cục, kết quả cuối cùng sẽ khác nhau tùy thuộc vào điểm bắt đầu.

Tốc độ học quá nhanh có thể làm thuật toán không hội tụ, lạc vào khu vực gần mục tiêu vì bước nhảy quá lớn; và nếu tốc độ học quá nhỏ, có thể ảnh hưởng đến tốc độ huấn luyện.

* **1.2.2 Momentum**

Gradient with momentum là một phương pháp tối ưu hóa trong huấn luyện mô hình machine learning, đặc biệt trong việc điều chỉnh trọng số của mạng neural. Nó tận dụng việc tích luỹ đà từ các gradient trước để điều chỉnh trọng số, không chỉ dựa vào gradient hiện tại mà còn tính toán sự thay đổi của trọng số dựa trên đà tích luỹ từ các bước trước. Điều này giúp cải thiện tốc độ hội tụ và ngăn chặn những biến động không mong muốn trong quá trình tối ưu hóa.

Công thức cập nhật trọng số theo gradient with momentum thường được biểu diễn như sau:

Δ*w*(*t*)=*α*⋅Δ*w*(*t*−1)−*η*⋅∇*J*(*w*)

Trong đó:

Δ*w*(*t*) là bước cập nhật trọng số tại thời điểm t

Δw(t−1) là đà tích luỹ từ bước trước.

*α* là hệ số đà, thường là một giá trị từ 0 đến 1 để điều chỉnh mức độ ảnh hưởng của đà tích luỹ.

*η* là learning rate.

∇*J*(*w*) là gradient của hàm mất mát *J* theo trọng số *w*.

Phương pháp gradient with momentum hỗ trợ ngăn chặn mô hình rơi vào các cực tiểu cục bộ, cải thiện tốc độ hội tụ và giảm thiểu sự dao động trong quá trình tối ưu hóa. Nó cũng làm cho việc cập nhật trọng số trở nên trơn tru hơn.

* **1.2.3 Adagrad**

Khác với các thuật toán trước đó, không duy trì learning rate như một hằng số trong quá trình huấn luyện. Thay vào đó, nó coi learning rate là một tham số, tức là sẽ điều chỉnh learning rate sau mỗi thời điểm.A math equation with a square root and a square root

Description automatically generated with medium confidence

Trong đó:

n : hằng số

gt : gradient tại thời điểm t

ϵ : hệ số tránh lỗi ( chia cho mẫu bằng 0)

G : là ma trận chéo mà mỗi phần tử trên đường chéo (i,i) là bình phương của đạo hàm vectơ tham số tại thời điểm t.

*Ưu điểm*:

Đáng chú ý của Adagrad là khả năng tự động điều chỉnh tốc độ học, không cần phải thay đổi learning rate thủ công; chỉ cần sử dụng tốc độ học mặc định là 0.01 và thuật toán sẽ tự động điều chỉnh nó.

*Nhược điểm:*

Adagrad là tổng bình phương của gradient sẽ ngày càng tăng theo thời gian, dẫn đến tốc độ học trở nên cực kỳ nhỏ, làm cho quá trình huấn luyện trở nên đóng băng.

1. **Continual Learning và Test Production**

* **2.1 Continual learning**

A diagram of a diagram of a data testing process

Description automatically generated

Continual learning là việc điều chỉnh mô hình khi có dữ liệu mới, đồng bộ hóa với các phân phối dữ liệu hiện tại. Tuy nhiên, sau khi mô hình được cập nhật, không thể áp dụng trực tiếp vào sản xuất mà cần phải trải qua quá trình thử nghiệm (Test Production) để đảm bảo tính an toàn và hiệu quả hơn so với mô hình đang sử dụng trong sản xuất.

*Tại sao cần Continual Learning?*

Dữ liệu ngày nay thường được tạo và cập nhật liên tục từ nhiều nguồn, và vì sự thay đổi của môi trường hoặc xu hướng mới, dữ liệu cũ không còn được coi là đáng tin cậy như trước. Trong thực tế, các mô hình học máy phải linh hoạt và có khả năng học từ dữ liệu mới để duy trì hiệu suất và độ chính xác. Continual Learning giải quyết vấn đề "quên" của mô hình khi chỉ được huấn luyện trên dữ liệu cụ thể, giúp mô hình duy trì khả năng học từ dữ liệu cũ và học thêm thông tin mới mà không làm mất đi kiến thức đã học.

Các bước chính của Continual Learning:

***Thu thập Dữ liệu Liên Tục*:** Đây là bước cơ bản, thu thập dữ liệu mới từ nguồn các nguồn khác nhau liên tục.

***Tiền Xử lý Dữ liệu*:** Dữ liệu mới cần được chuẩn hóa, làm sạch và tiền xử lý để phù hợp với mô hình hoặc để kết hợp với dữ liệu cũ.

***Huấn Luyện Mô Hình Liên Tục:*** Mô hình cần được huấn luyện lại không chỉ với dữ liệu mới mà còn kết hợp với dữ liệu cũ. Quá trình này thường yêu cầu kỹ thuật đặc biệt để tránh hiện tượng "quên" thông tin quan trọng từ dữ liệu cũ.

***Kiểm Tra và Đánh Giá:*** Mô hình sau khi được cập nhật cần được kiểm tra và đánh giá để đảm bảo rằng nó vẫn duy trì hiệu suất và độ chính xác. Điều này có thể bao gồm việc kiểm tra trên dữ liệu cũ và mới để đảm bảo rằng nó vẫn hoạt động hiệu quả.

***Kết Hợp Kiến Thức Cũ và Mới:*** Một phần quan trọng của Continual Learning là khả năng kết hợp thông tin từ dữ liệu cũ và mới mà không làm mất đi kiến thức đã học.

***Tối Ưu Hóa và Cải Tiến:*** Quá trình này liên tục diễn ra, từ việc cải tiến quy trình tiền xử lý, mô hình học tập, đến việc tối ưu hóa cách tích hợp dữ liệu mới vào mô hình.

***Triển Khai Mô Hình:*** Sau khi mô hình đã được cập nhật và kiểm tra, nó có thể được triển khai trong môi trường thực tế để sử dụng với dữ liệu mới.

* **2.2 Test Production**

Test Production trong học máy là bước cần thiết để xác nhận tính đúng đắn và hiệu suất của mô hình trước khi triển khai vào môi trường thực tế. Nó đánh giá chất lượng và khả năng hoạt động của mô hình, thường bắt đầu sau khi đã kiểm tra và huấn luyện mô hình trên dữ liệu huấn luyện và kiểm tra. Tuy nhiên, quá trình này tập trung vào việc thử nghiệm mô hình trong môi trường sản xuất để đảm bảo tính ổn định và hiệu suất, thậm chí khi đối mặt với dữ liệu mới không giống với dữ liệu huấn luyện ban đầu.

Top of Form

Các giai đoạn quan trọng trong quy trình Test Production bao gồm:

Chuẩn bị dữ liệu mới: Việc chuẩn bị và làm sạch dữ liệu mới để kiểm thử mô hình trong môi trường thực tế là yếu tố cần thiết. Dữ liệu này có thể khác biệt so với dữ liệu huấn luyện ban đầu.

Triển khai mô hình: Mô hình được triển khai vào môi trường thực tế, bao gồm việc cấu hình hạ tầng và kết nối với hệ thống để xử lý dữ liệu thời gian thực hoặc các yêu cầu sản xuất.

Kiểm tra hiệu suất: Mô hình được đánh giá về hiệu suất trong môi trường thực tế, đo lường các thông số như độ chính xác, phân loại, hoặc các chỉ số tương ứng với loại bài toán máy học.

Giám sát và cập nhật: Quá trình Test Production không chỉ dừng lại khi mô hình được triển khai. Việc liên tục giám sát hiệu suất và cập nhật mô hình khi cần thiết là quan trọng để duy trì và nâng cao hiệu suất theo thời gian.

Bảo mật và tuân thủ quy định: Bảo vệ dữ liệu và tuân thủ quy định pháp luật là yếu tố quan trọng khi triển khai mô hình xử lý dữ liệu cá nhân hoặc nhạy cảm.

Test Production không chỉ là để đánh giá mô hình, mà còn nhằm cải thiện và điều chỉnh nó để hoạt động tốt nhất trong môi trường thực tế. Hiệu suất và độ chính xác của mô hình khi hoạt động trong sản xuất có thể ảnh hưởng lớn đến kết quả kinh doanh và trải nghiệm người dùng, do đó, quá trình Test Production cần được thực hiện một cách cẩn thận và toàn diện.

Top of Form

Test Production không chỉ là để đánh giá mô hình, mà còn nhằm cải thiện và điều chỉnh nó để hoạt động tốt nhất trong môi trường thực tế. Hiệu suất và độ chính xác của mô hình khi hoạt động trong sản xuất có thể ảnh hưởng lớn đến kết quả kinh doanh và trải nghiệm người dùng, do đó, quá trình Test Production cần được thực hiện một cách cẩn thận và toàn diện.