TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**MAI XUÂN HÀ - 52100885**

**TÌM HIỂU LÝ THUYẾT HỌC MÁY**

**BÁO CÁO CUỐI KỲ**

**NHẬP MÔN HỌC MÁY**

**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**



**MAI XUÂN HÀ - 52100885**

**TÌM HIỂU LÝ THUYẾT HỌC MÁY**

**BÁO CÁO CUỐI KỲ**

**NHẬP MÔN HỌC MÁY**

Người hướng dẫn

**TS. LÊ ANH CƯỜNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

**LỜI CẢM ƠN**

Đầu tiên, em xin cảm ơn Trường Đại học Tôn Đức Thắng vì đã đưa môn nhập môn Học máy vào chương trình giảng dạy. Đặc biệt, em xin cám ơn sâu sắc đến giảng viên của bộ môn Thầy Lê Anh Cường đã truyền đạt những kiến thức quý báu cho em trong quá trình học tập. Đây sẽ là những kiến thức quý báu trong hành trang để chúng em có thể đi vươn xa trong tương lai.

Đây là một môn học vô cùng thú vị, bổ ích và có tính thực tế. Đảm bảo cung cấp đầy đủ kiến thức và gắn liền với nhu cầu thực tiễn của sinh viên Công Nghệ Thông Tin. Mặc dù em đã cố gắng hết sức những dự án sẽ có thể có những thiếu sót và những sai sót trong dự án, kính mong thầy cô xem xét và góp ý để em được hoàn thiện hơn.

Em xin chân thành cảm ơn.

*TP. Hồ Chí Minh, ngày 15 tháng 12 năm 2023*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

|  |
| --- |
| *Hà*  *Mai Xuân Hà* |

**CÔNG TRÌNH 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 khóa 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 Dự 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 Dự 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 15 tháng 12 năm 2023*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

|  |
| --- |
| *Hà*  *Mai Xuân Hà* |

**TÌM HIỂU LÝ THUYẾT HỌC MÁY**

**TÓM TẮT**

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

* Giới thiệu vấn đề
* Các phương pháp cơ bản, công thức, ưu nhược điểm của nó
* So sánh

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

* Tìm hiểu về Continual Learning
* Tìm hiểu về Test Production
* Ứng dụng

**MỤC LỤC**

[DANH MỤC HÌNH VẼ v](#_Toc154131221)

[DANH MỤC BẢNG BIỂU vi](#_Toc154131222)

[DANH MỤC CÁC CHỮ VIẾT TẮT vii](#_Toc154131223)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 8](#_Toc154131224)

[1.1 Lý do chọn đề tài 8](#_Toc154131225)

[1.2 Mục tiêu thực hiện đề tài 8](#_Toc154131226)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 8](#_Toc154131227)

[2.1 Bài toán tối ưu hóa 8](#_Toc154131228)

[2.1.1 Bài toán tối ưu hóa NP-hard với trường hợp của TSP 9](#_Toc154131229)

[2.2 Bài toán Traveling salesman problem 9](#_Toc154131230)

[2.2.1 Công thức toán học 10](#_Toc154131231)

# DANH MỤC HÌNH VẼ

**No table of figures entries found.**

# DANH MỤC BẢNG BIỂU

**No table of figures entries found.**

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| GD | Gradient Descent |
| SCD | Stochastic Gradient Descent |
| NAG | Nesterov accelerated gradient |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# CÁC PHƯƠNG PHÁP OPTIMIZER

## Giới thiệu

Trong quá trình huấn luyện một mô hình học máy, việc điều chỉnh trọng số của mô hình rất quan trọng để đảm bảo các độ chính xác và tối ưu hóa của dự đoán tốt nhất có thể.

Các phương pháp tối ưu hóa sẽ liên kết hàm mất mát và các tham số của mô hình bằng cách cập nhật mô hình dựa trên kết quả của hàm mất mát và tối ưu hóa mô hình tốt nhất bằng cách chỉnh sửa các trọng số.

Các phương pháp đó ảnh hưởng đến tốc độ hội tụ của mô hình, tổng quát hóa và hiệu suất của mô hình trên dữ liệu mới.

## Các phương pháp cơ bản

### Gradient Descent và các biến thể

#### Gradient Descent (GD)

GD là một thuật toán tối ưu hóa được sử dụng trong nhiều lĩnh vực, bao gồm học máy, học sâu và khoa học dữ liệu. Mục đích của GD là tìm giá trị tối ưu của một hàm bằng cách di chuyển dần dần trong không gian tìm kiếm theo hướng có độ dốc lớn nhất (tốc dộ ảnh hưởng bởi learning rate).

Công thức:

* là giá trị của trọng số mới
* lá giá trị của trọng số hiện tại
* là learning rate
* là đạo hàm của hàm số tại trọng số đó

Ưu điểm:

* Đơn giản và dễ hiểu
* Có thể được áp dụng cho nhiều loại hàm mục tiêu
* Có thể được sử dụng để tối ưu hóa nhiều tham số cùng lúc

Nhược điểm:

* Có thể bị mắc kẹt tại cực đại hoặc cực tiểu cục bộ
* Có thể chậm hội tụ
* Có thể nhạy cảm với learning rate α
* Không phù hợp với bộ dữ liệu được cập nhật liên tục

#### Stochastic Gradient Descent (SGD)

Đối với tập dữ liệu bình thường GD hoạt động rất tốt đến việc tìm ra điểm cực tiểu và ít ảnh hưởng bởi độ nhiễu và độ ngẫu nhiên hơn. Nhưng đối với tập dữ liệu siêu lớn, việc sử dụng GD là không khả thi vì nó lượng tài nguyên sử dụng để chạy qua toàn bộ dữ liệu là siêu lớn. Để giải quyết vấn đề này, SGD được áp dụng với bài toán có dữ liệu lớn. Phương pháp này sẽ chọn ra một mẫu dữ liệu ngẫu nhiên để thực hiện GD trên nó trong mỗi lần lặp, từ đó cải thiện thời gian huấn luyện và tốc độ đạt tới điểm cực tiểu.

Ưu điểm:

* Thuật toán giải quyết được đối với cơ sở dữ liệu lớn mà GD không làm được.

Nhược điểm:

* Đường đi tới điểm cực tiểu bị ảnh hưởng nhiều hơn bởi dữ liệu nhiễu
* Vẫn chịu ảnh hưởng trực tiếp bởi learning rate

### Momentum và các biến thể

#### Momentum

Phương pháp Momentum khắc phục các hạn chế của GD bằng cách thêm vào giá trị vận tốc (đà) nhằm giúp thuật toán GD vượt qua các điểm tối thiểu cục bộ nhanh hơn.

Công thức:

* là giá trị của trọng số mới
* lá giá trị của trọng số hiện tại
* là learning rate
* là đạo hàm của hàm số tại trọng số đó
* là tham số điều chỉnh vận tốc (đà)
* là giá trị vận tốc trước đó của trọng số w trước đó

Ưu điểm:

* Giúp GD vượt qua các điểm local minium nhanh hơn và tiến tới global minium dễ hơn

Nhược điểm:

* Bởi vì có thêm vận tốc (đà) nên sẽ dao động quanh global minium làm mất thời gian để triệt tiêu vận tốc và dừng lại.

#### Nesterov accelerated gradient (NAG) (Nesterov Momentum)

Thuật toán momentum tuy giúp GD vượt qua các local minium dễ hơn, tuy nhiên do có vận tốc nên việc dừng lại ở global minium tốn rất nhiều thời gian. Thuật toán

# CƠ SỞ LÝ THUYẾT

## Bài toán tối ưu hóa

Trong khoa học máy tính và toán học, bài toán tối ưu hóa là bài toán tìm kiếm lời giải tốt nhất trong tất cả các lời giải khả thi.

Một bài toán tối ưu hóa là một bài toán yêu cầu tìm một giải pháp tối ưu cho một hàm mục tiêu, trong khi thỏa mãn một số ràng buộc.

Bài toán tối ưu hóa có thể được chia thành hai loại tùy thuộc vào việc các biến là liên tục hay rời rạc.

### Bài toán tối ưu hóa NP-hard với trường hợp của TSP

Một bài toán NP là một bài toán mà giải pháp của nó có thể được xác minh nhanh chóng. Điều này có nghĩa là có một thuật toán có thể xác minh tính đúng đắn của một giải pháp trong thời gian polynomial.

Một bài toán NP là một bài toán mà giải pháp của nó có thể được xác minh nhanh chóng. Trong trường hợp của TSP, một giải pháp có thể được xác minh nhanh chóng bằng cách tính tổng độ dài của tất cả các cạnh trong chu trình.

Bài toán người bán hàng là một bài toán NP-hard có nghĩa là không có thuật toán nào có thể giải quyết nó một cách hiệu quả trong thời gian polynomial.

Điều này có nghĩa là nếu có một thuật toán NP-đúng để giải quyết TSP, thì thuật toán đó cũng có thể được sử dụng để giải quyết tất cả các bài toán NP khác.

## Bài toán Traveling salesman problem

Là bài toán tối ưu hóa tổ hợp, trong đó cần có những tiêu chí sau:​

* Nhân viên bán hàng (Salesman) nhắm đến việc đi thăm toàn bộ thành phố và quay trở lại thành phố ban đầu của mình.​
* Salesman bắt buộc phải đi qua mỗi thành phố một lần với chi phí tối thiểu.​

Thông thường độ dài của chuyến tham quan đại diện cho chi phí của giải pháp (tổng khoảng cách mà salesman phải đi qua)

Bài toán người bán hàng du lịch (TSP), là một bài toán thuộc nhóm bài toán NP-hard.​

TSP có thể được giải quyết như một vấn đề, đồng thời cũng là một thử nghiệm so sánh để đo lường năng lực của các giải thuật.

### Công thức toán học

Công thức toán học của TSP: