ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

SAVADY Sathaphone

Nghiên cứu phương pháp Máy véc tơ hỗ trợ với lề mềm và ứng dụng phân lớp dữ liệu tuyến tính có nhiễu

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2019

Số hóa bởi Trung tâm Học liệu và Công nghệ thông tin – ĐHTN http://lrc.tnu.edu.vn

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

SAVADY Sathaphone

Nghiên cứu phương pháp Máy véc tơ hỗ trợ với lề mềm và ứng dụng phân lớp dữ liệu tuyến tính có nhiễu

Chuyên ngành: Khoa học máy tính
Mã số: 8480101

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. Đàm Thanh Phương

THÁI NGUYÊN - 2019

Số hóa bởi Trung tâm Học liệu và Công nghệ thông tin – ĐHTN http://lrc.tnu.edu.vn

LỜI CAM ĐOAN

Tôi xin cam đoan: Luận văn thạc sỹ chuyên ngành Khoa học máy tính, tên đề tài "Nghiên cứu phương pháp Máy véc tơ hỗ trợ với lề mềm và ứng dụng phân lớp dữ liệu tuyến tính có nhiễu" là công trình nghiên cứu, tìm hiểu và trình bày do tôi thực hiện dưới sự hướng dẫn khoa học của TS. Đàm Thanh Phương, Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên.

Kết quả tìm hiểu, nghiên cứu trong luận văn là hoàn toàn trung thực, không vi phạm bất cứ điều gì trong luật sở hữu trí tuệ và pháp luật Việt Nam. Nếu sai, tôi hoàn toàn chịu trách nhiệm trước pháp luật.

Tất cả các tài liệu, bài báo, khóa luận, công cụ phần mềm của các tác giả khác được sử dụng lại trong luận văn này đều được chỉ dẫn tường minh về tác giả và đều có trong danh mục tài liệu tham khảo.

Thái Nguyên, ngày 18 tháng 7 năm 2019

Tác giả luận văn

SAVADY Sathaphone

LỜI CẢM ƠN

Tác giả xin chân thành cảm ơn TS Đàm Thanh Phương, trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên, là giáo viên hướng dẫn khoa học đã hướng dẫn tác giả hoàn thành luận văn này, xin được cảm ơn các thầy, cô giáo trường Đại học công nghệ thông tin và truyền thông nơi tác giả theo học và hoàn thành chương trình cao học đã nhiệt tình giảng dạy và giúp đỡ.

Xin cảm ơn nơi tác giả công tác đã tạo mọi điều kiện thuận lợi để tác giả hoàn thành chương trình học tập.

Và cuối cùng xin cảm ơn gia đình, bạn bè, đồng nghiệp đã động viên, giúp đỡ tác giả trong suốt thời gian học tập, nghiên cứu và hoàn thành luận văn này.

Xin chân thành cảm ơn.

Thái Nguyên, ngày 18 tháng 7 năm 2019

Tác giả luận văn

SAVADY Sathaphone

DANH SÁCH HÌNH VẼ

2.1 Các mặt phân cách hai lớp tuyến tính 26
2.2 Lề của hai lớp không bằng nhau. 26
2.3 Lề của hai lớp bằng nhau, nhưng chọn phương án có lề rộng hơn. 26
2.4 Phân tích bài toán SVM 27
2.5 Các điểm gần mặt phân cách nhất của hai classes được khoanh tròn.28
2.6 Soft margin SVM khi có nhiễu, SVM sẽ không hoạt động hiệu quả.33
2.7 Soft margin SVM khi dữ liệu gần linearly separable, SVM thuần sẽ không hoạt động hiệu quả. 34
2.8 Mô tả sự được và mất khi chọn lề. 36
2.9 Hàm hinge loss 42

3.1 Minh họa nghiệm tìm đc bởi SVM 47
3.2 Dữ liệu tuyến tính có nhiễu 49
3.3 Đường phân chia tìm được bởi sklearn 52
3.4 Đường phân chia tìm được bởi hinge loss 52
3.5 Đường phân chia tìm được bởi đối ngẫu. Các kết quả là như nhau 52
3.6 Nghiệm tìm bằng sklearn, C =0.1. 53
3.7 Nghiệm tìm bằng sklearn, C=1 53
3.8 Nghiệm khi C=10 53
3.9 Ảnh hưởng của C lên nghiệm, C=100. 54

DANH MỤC KÝ HIỆU, TỪ VIẾT TẮT

$\mathbb{R}$	Tập hợp số thực.
$\mathbb{Z}$	Tập hợp số nguyên.
$\mathbb{C}$	Tập hợp số phức.
$\mathbb{R}^n$	Không gian Euclide n chiều.
$\|\|.\|\|$	Chuẩn Euclide.
SVM	Suport vector machine - máy véc tơ hỗ trợ
Soft SVM	Máy véc tơ hỗ trợ với lề mềm.
Argmin	Bài toán tối ưu tham số.
Duality	Tính chất đối ngẫu, đối ngẫu
Lagrangian	Đối ngẫu Lagrange
Strong duality	Tính chất đối ngẫu mạnh
Weak duality	Tính chất đối ngẫu yếu
KKT	Karush Kuhn Tucker - Điều kiện KKT
$det(A)$	Định thức của ma trận A.
A^T	Chuyển vị của ma trận A.
A^H	Hermitian của ma trận A.
$\|\mathbf{A}\|_F$	Frobenius norm của ma trận A
$diag(A)$	Đường chéo chính của ma trận A
$trace(A)$	trace của ma trận A
$sng(x)$	Hàm xác định dấu.
$\frac{\partial f}{\partial x}$	Đạo hàm của hàm số f theo biến số $x \in R$
$\nabla_{\mathbf{x}} f$	Gradient (đạo hàm) của hàm số f theo véc tơ x.

MỤC LỤC

Lời cam đoan i

Lời cảm ơn ii

Danh sách hình vẽ ii

Danh mục ký hiệu, từ viết tắt v

Mở đầu 1

Chương 1. MỘT SỐ KIẾN THỨC CHUẨN BỊ 4

1.1. Tổng quan về ngôn ngữ Python 4

1.2. Giới thiệu về học máy 7

1.3. Một số kiến thức toán học bổ trợ 11

Chương 2. PHƯƠNG PHÁP MÁY VÉC TƠ HỖ TRỢ SVM ... 25

2.1. Bài toán phân lớp SVM 25

2.2. Bài toán SVM với lề mềm 33

Chương 3. ỨNG DỤNG VÀ LẬP TRÌNH MÔ PHỎNG 46

3.1. Lập trình tìm nghiệm cho SVM 46

3.2. Lập trình tìm nghiệm cho Soft SVM 48

Kết luận chung 55

Tài liệu tham khảo 55

MỞ ĐẦU

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, ..., chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning. [7]

Machine Learning là một tập con của AI. Theo định nghĩa của Wikipedia, Machine learning is the subfield of computer science that "gives computers the ability to learn without being explicitly programmed". Nói đơn giản, Machine Learning là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể. [1]

Phân lớp và hồi quy là hai trong số các thuật toán học máy có giám sát có nhiều ứng dụng trong thực tế. Bài toán phân lớp tuyến tính cơ bản yêu cầu xây dựng một classifier từ dữ liệu training để khi có điểm dữ liệu mới ta có thể dự đoán được nhãn của nó. Giả sử rằng có hai class khác nhau được mô tả bởi các điểm trong không gian nhiều chiều, hai classes này linearly separable, tức tồn tại một siêu phẳng phân chia chính xác hai classes đó. Hãy tìm một siêu mặt phẳng phân chia hai classes đó, tức tất cả các điểm thuộc một class nằm về cùng một phía của siêu mặt phẳng đó và ngược phía với toàn bộ các điểm thuộc class còn lại.

Chúng ta đã biết rằng, thuật toán PLA có thể làm được việc này nhưng nó có thể cho chúng ta vô số nghiệm. Câu hỏi đặt ra là: trong vô số các mặt

phân chia đó, đâu là mặt phân chia tốt nhất theo một tiêu chuẩn nào đó? Chúng ta cần một đường phân chia sao cho khoảng cách từ điểm gần nhất của mỗi class tới đường phân chia là như nhau, như thế thì mới công bằng. Khoảng cách như nhau này được gọi là margin (lề). Ngoài ra, việc margin rộng hơn sẽ mang lại hiệu ứng phân lớp tốt hơn vì sự phân chia giữa hai classes là rạch ròi hơn. Việc này là một điểm khá quan trọng giúp Support Vector Machine mang lại kết quả phân loại tốt hơn so với Neural Network với 1 layer, tức Perceptron Learning Algorithm.[1], [2]

Bài toán tối ưu trong Support Vector Machine (SVM) chính là bài toán đi tìm đường phân chia sao cho margin là lớn nhất. Đây cũng là lý do vì sao SVM còn được gọi là Maximum Margin Classifier. [7]

Tuy nhiên ngay cả SVM với lề cứng cũng sẽ không giải quyết được một số bài toán phân lớp có nhiễu, việc nghiên cứu SVM với lề mềm sẽ giải quyết tốt hơn trường hợp này.

Nội dung của luận văn gồm 3 chương:

Chương 1. Các kiến thức cơ sở

Chương này trình bày các kiến thức chuẩn bị cho việc nghiên cứu.

1.1 Giới thiệu về học máy

1.2 Giới thiệu về ngôn ngữ lập trình Python

1.3 Các kiến thức cơ sở về Đại số tuyến tính

Chương 2. Phương pháp máy véc tơ hỗ trợ SVM

Nội dung chương 2 tập trung vào vấn đề nghiên cứu các phương pháp máy véc tơ hỗ trợ, cụ thể như sau:

2.1 Xây dựng bài toán tối ưu cho SVM

2.2 Bài toán đối ngẫu cho SVM

2.3 Bài toán đối ngẫu Lagrange.

2.4 Bài toán tối ưu không ràng buộc cho Soft SVM.

Chương 3. Ứng dụng và lập trình mô phỏng

Sau khi nắm rõ các nội dung trong chương 2, chương 3 trình bày một số ví