BỘ CÔNG THƯƠNG

ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HCM



***----------------------***

**ĐỒ ÁN MỘT CHỈ**

**ĐỀ TÀI:**

**DỰ ĐOÁN KẾT QUẢ TỐT NGHIỆP DỰA TRÊN**

**MÔ HÌNH HỌC CÓ GIÁM SÁT**

TP.HCM, tháng 11 năm 2018

BỘ CÔNG THƯƠNG

ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HCM



***---------------------***

**ĐỒ ÁN MỘT CHỈ**

**ĐỀ TÀI:**

**DỰ ĐOÁN KẾT QUẢ TỐT NGHIỆP DỰA TRÊN**

**MÔ HÌNH HỌC CÓ GIÁM SÁT**

GVHD: Trần Đức

Sinh viên thực hiện:

2001150075 - Nguyễn Ngọc Thùy Trang

2001150360 - Chang Chia Sheng

2001150110 - Nguyễn Anh Vũ

TP.HCM, tháng 11 năm 2018

LỜI MỞ ĐẦU

Qua từng năm, tỉ lệ sinh viên có điểm những kỳ học của năm 1, 2, 3 đạt thì tốt nghiệp càng cao. Trong những đợt tốt nghiệp gần đây, có rất nhiều biến động về điểm. Vậy các khóa 03DHTH, 04DHTH, 05DHTH khoa Công Nghệ Thông Tin trường DHCNTP. Có nhiều thay đổi như những năm trước hay không? Hãy cùng nhóm em tìm hiểu bài dự đoán dưới đây để có thể hiểu được những môn có kả năng tác động đến việc sinh viên tốt nghiệp.

Trong quá trình thực hiện, nhóm em được học từ giảng viên hướng dẫn là Thầy Trần Đức, nhóm chúng em xin gửi lời cảm ơn sâu sắc đến Thầy. Đồng thời cũng gửi lời cảm ơn đến quý thầy cô Khoa Công Nghệ Thông Tin đã trang bị cho chúng em những kinh nghiệm quý báu, tạo điều kiện cho nhóm em có kiến thức nền tảng để thực hiện môn đồ án một chỉ.

Do lượng kiến thức chưa đầy đủ và thiếu kinh nghiệm chuyên môn vì vậy, trong quá trình thực hiện đề tài nhóm em còn thiếu sót. Mong Thầy góp ý chân thành để giúp chúng em hoàn thành đề tài này một cách tốt nhất có thể.

**NHẬN XÉT CỦA GIẢNG VIÊN**

**………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

**………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

**………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………****………………………………………………………………………………………………………………………………………………………………………………………………**

BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |  |
| --- | --- | --- | --- |
|  | Thuỳ Trang | Chia Sheng | Anh Vũ |
| Tuần 1 | Tìm hiểu và phân tích đề tài,tìm hiểu 3 thuật toán, Bayes, C45, SVM | Tìm hiểu và phân tích đề tài,tìm hiểu 3 thuật toán, Bayes, C45, SVM | Tìm hiểu và phân tích đề tài,tìm hiểu 3 thuật toán, Bayes, C45, SVM |
|
| Tuần 2 | Thu thập dữ liệu và phân tích | Phân tích và tiền xử lý | Phân tích và tiền xử lý |
|
| Tuần 3 | Tìm hiểu về MATLAB, WEKA | Trộn dữ liệu các khóa, tìm hiểu MATLAB WEKA | Tìm hiểu về MATLAB và WEKA |
|
| Tuần 4 | Tìm hiểu về thuật toán Supervised learning, tìm hiểu về MATLAB | Tìm hiểu về thuật toán Supervised learnin | Tìm hiểu về thuật toán Supervised learning, tìm hiểu về MATLAB |
|
| Tuần 5 | Chia dữ liệu thành dữ liệu training | Tổng hợp dữ liệu đưa vào MATLAP | Chia dữ liệu thành dữ liệu testing |
|
| Tuần 6 | Chạy thuật toán Bayes, phân tích kết quả | Chạy thuật toán SVM, phân tích kết quả | Chạy thuật toán C45, phân tích kết quả |
|
| Tuần 7 | so sánh kết quả thuật toán Bayes với mô hình trong WEKA | so sánh kết quả thuật toán C45 với mô hình trong WEKA | so sánh kết quả thuật toán SVM với mô hình trong WEKA |
|
| Tuần 8 | Thiết kế slide power point, chỉnh sửa báo cáo | Tổng hợp tất cả kết quả và chỉnh sửa báo cáo | Viết báo cáo |
|

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Viết tắt** |  | **Tiếng Anh** |  | **Tiếng Việt** |
|  |  | DataMining | 🡪 | Khai thác dữ liệu |
|  |  | Supervised Learning | 🡪 | Học có giám sát |
|  |  | Unsupervised learning | 🡪 | Học không giám sát |
|  |  | Data preprocessing | 🡪 | Tiền xử lý dữ liệu |
|  |  | Classificantion | 🡪 | Phân lớp |
|  |  | Clustering  Association rule  Summerization  Change and deviation detection  Regression | 🡪  🡪  🡪  🡪  🡪 | Gom cụm  Luật kết hợp  Tổng hợp hóa  Phát hiện sự biến đổi  và độ lệch  Hồi quy |
| SVM | 🡪 | Suport Vector Machine |  |  |
| C4.5 | 🡪 | Classes linearly  Information Gain Ratio  Soft margin  Hard Margin | 🡪  🡪  🡪 | Phân lớp tuyến tính  Biên mềm  Biên cứng |

DANH MỤC HÌNH ẢNH

[Hình 3.2 1: Phân lớp tuyến tính 19](#_Toc531328825)

[Hình 3.2 2: Các mặt phân cách hai classes linearly separable 19](#_Toc531328826)

[Hình 3.2 3: Phân lớp phi tuyến 20](#_Toc531328827)

[Hình 3.2 4: Margin của hai classes là bằng nhau và lớn nhất có thể 20](#_Toc531328828)

[Hình 3.2 5: Phân tích bài toán SVM 22](#_Toc531328829)

[Hình 3.2 6: Các điểm gần mặt phân cách nhất của hai classes được khoanh tròn. 24](#_Toc531328830)

[Hình 3.2 7: Soft margin SVM. Khi a) có nhiễu hoặc b) dữ liệu gần linearly separable, SVM thuần sẽ không hoạt động hiệu quả. 24](file:///C:\Users\PC-NAV\Downloads\DOAN1CHI_SO77_VU_SHENG_TRANG%20(1).docx#_Toc531328831)

[Hình 3.2 8: Giới thiệu các biến slack ξn 25](#_Toc531328832)

[Hình 3.2 9: Ví dụ mô hình cây quyết định 26](#_Toc531328833)

DANH MỤC CÁC BẢNG

[Bảng 2. 1: Danh sách các môn để dự đoán tốt nghiệp 12](#_Toc531329589)

[Bảng 4. 1: Confusion Matrix Naive Bayes 29](#_Toc531329686)

[Bảng 4. 2: Confusion Matrix SVM 29](#_Toc531329687)

[Bảng 4. 3: Confusion Matrix C4.5 29](#_Toc531329688)

MỤC LỤC

[CHƯƠNG 1: TỔNG QUAN 1](#_Toc531788281)

[1.1 THÔNG TIN VỀ ĐỀ TÀI 1](#_Toc531788282)

[1.1.1 Tên đề tài 1](#_Toc531788283)

[1.1.2 Giới thiệu chung về đề tài 1](#_Toc531788284)

[1.2. MỤC ĐÍCH VÀ PHẠM VI ĐỀ TÀI 1](#_Toc531788286)

[1.2.1. Mục đích của đề tài 1](#_Toc531788287)

[1.2.2. Yêu cầu của đề tài 1](#_Toc531788288)

[1.2.3. Phạm vi của đề tài 1](#_Toc531788289)

[1.2.4. Môi trường triển khai đề tài 1](#_Toc531788291)

[CHƯƠNG 2: KHẢO SÁT 1](#_Toc531788293)

[2.1. DỮ LIỆU THỰC HIỆN 1](#_Toc531788294)

[2.2. QUI TẮC XÉT TỐT NGHIỆP BẬC ĐẠI HỌC 2](#_Toc531788352)

[CHƯƠNG 3: PHÂN TÍCH 3](#_Toc531788354)

[3.1. MÁY HỌC VÀ DATAMINDING 3](#_Toc531788355)

[3.1.1. Máy học 3](#_Toc531788356)

[3.1.2. Dataminding 3](#_Toc531788359)

[3.2. THUẬT TOÁN PHÂN LỚP 5](#_Toc531788360)

[3.2.1. Khái quát về học có giám sát và không giám sát 5](#_Toc531788361)

[3.2.1.1. Học có giám sát ( Supervised Learning ) 5](#_Toc531788362)

[3.2.1.2. Học không giám sát ( Unsupervised Learning ) 5](#_Toc531788367)

[3.2.2. Naive Bayes 6](#_Toc531788371)

[3.2.2.1. Khái niệm 6](#_Toc531788372)

[3.2.2.2. Định lý 6](#_Toc531788373)

[3.2.3. SVM (Suport Vector Machine) 8](#_Toc531788395)

[3.2.3.1. Khái niệm 8](#_Toc531788396)

[3.2.3.2. Xây dựng bài toán tối ưu cho SVM 11](#_Toc531788397)

[3.2.3.3. Soft Margin SVM 14](#_Toc531788398)

[3.2.3.4. Phân tích toán học 15](#_Toc531788399)

[3.2.4. C4.5 (Information Gain Ratio) 16](#_Toc531788400)

[3.2.4.1. Khái niệm 16](#_Toc531788401)

[3.2.4.2. Diễn giải 17](#_Toc531788403)

[3.2.4.3. Ưu điểm 17](#_Toc531788419)

[3.2.4.4. Nhược điểm 18](#_Toc531788424)

[3.3. QUY TRINH THỰC HIỆN DỰ ĐOÁN 18](#_Toc531788428)

[CHƯƠNG 4: KẾT QUẢ THỬ NGHIỆM 19](#_Toc531788446)

[4.1. Naive Bayes 19](#_Toc531788447)

[4.2. SVM 19](#_Toc531788461)

[4.3. C4.5 19](#_Toc531788475)

[MẪU KHẢO SÁT DỰ ĐOÁN 20](#_Toc531788489)

[TỔNG KẾT 21](#_Toc531788490)

[TÀI LIỆU THAM KHẢO 22](#_Toc531788496)

* 2. TỔNG QUAN

1.1 THÔNG TIN VỀ ĐỀ TÀI

1.1.1 Tên đề tài

DỰ ĐOÁN KẾT QUẢ TỐT NGHIỆP CỦA SINH VIÊN DỰA TRÊN MÔ HÌNH HỌC CÓ GIÁM SÁT.

1.1.2 Giới thiệu chung về đề tài

Nhóm chúng em làm đề tài này để thiết kế một chương trình được huấn luyện trên máy tính để dự đoán một cách hiệu quả và nhanh nhất xem sinh viên có đậu tốt nghiệp hay không.

1.2. MỤC ĐÍCH VÀ PHẠM VI ĐỀ TÀI



1.2.1. Mục đích của đề tài

Dự đoán kết quả tốt nghiệp cho sinh viên khóa 06DHTH dựa trên điểm trung bình từng môn học của 3 năm học.

1.2.2. Yêu cầu của đề tài

Dựa trên điểm trung bình các môn học có sẵn của sinh viên, sử dụng thuật toán phân lớp SVM, C4.5, KNN,và luật kết hợp phân tích kết quả so sánh với mô hình WEKA.

1.2.3. Phạm vi của đề tài

Các khóa 03DHTH, 04DHTH, 05DHTH khoa Công Nghệ Thông Tin trường DHCNTP.

1.2.4. Môi trường triển khai đề tài

MATLAB và WEKA.

* 2. KHẢO SÁT

2.1. DỮ LIỆU THỰC HIỆN

|  |  |  |  |
| --- | --- | --- | --- |
| Cột | Tên môn | Cột | Tên môn |
| 1 | Anh văn sơ cấp | 27 | Lập trình hướng đối tượng |
| 2 | Giáo dục thể chất 1 | 28 | Thiết kế Web |
| 3 | Tin học văn phòng | 29 | Thực hành cơ sở dữ liệu |
| 4 | Anh văn 1 | 30 | Thực hành lập trình hướng đối tượng |
| 5 | Giáo dục quốc phòng - an ninh 1 | 31 | Thực hành thiết kế Web |
| 6 | Giáo dục thể chất 2 | 32 | Tư tưởng Hồ Chí Minh |
| 7 | Kỹ năng học tập hiệu quả | 33 | Xác suất thống kê |
| 8 | Anh văn 2 | 34 | Hàm phức và phép biến đổi Laplace |
| 9 | Giáo dục quốc phòng - an ninh 3 AB | 35 | Đường lối cách mạng của Đảng Cộng sản Việt Nam |
| 10 | Giáo dục thể chất 3 | 36 | Lập trình Windows |
| 11 | Ngôn ngữ lập trình | 37 | Mạng máy tính |
| 12 | Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 2 | 38 | Thực hành lập trình Windows |
| 13 | Thực hành ngôn ngữ lập trình | 39 | Thực hành mạng máy tính |
| 14 | Toán cao cấp A2 | 40 | Truyền thông kỹ thuật số |
| 15 | Toán rời rạc | 41 | Đồ họa máy tính |
| 16 | Kỹ năng giao tiếp | 42 | Thực hành đồ họa máy tính |
| 17 | Cấu trúc dữ liệu và giải thuật | 43 | Hệ quản trị cơ sở dữ liệu |
| 18 | Kiến trúc máy tính | 44 | Phân tích thiết kế hệ thống thông tin |
| 19 | Phương pháp tính | 45 | Phương pháp nghiên cứu khoa học |
| 20 | Thí nghiệm vật lý đại cương | 46 | Thực hành hệ quản trị cơ sở dữ liệu |
| 21 | Thực hành cấu trúc dữ liệu và giải thuật | 47 | Thực hành phân tích thiết kế hệ thống thông tin |
| 22 | Vật lý đại cương 2 | 48 | Thương mại điện tử ngành CNTT |
| 23 | Môi trường và con người | 49 | Trí tuệ nhân tạo |
| 24 | An toàn lao động | 50 | Lý thuyết đồ thị |
| 25 | Cơ sở dữ liệu | 51 | Thực hành lý thuyết đồ thị |
| 26 | Hệ điều hành | 52 | Xếp loại |

*Bảng 2. 1: Danh sách các môn để dự đoán tốt nghiệp*

2.2. QUI TẮC XÉT TỐT NGHIỆP BẬC ĐẠI HỌC

Ngoài các qui tắc xét đủ điểm công tác chính trị và hạnh kiểm thì còn những nguyên tắc như sau:

* Các môn không tích lũy tín chỉ phải >= 5 điểm trở lên (trong bảng 1 thì thể hiện ở cột 1 đến 10),
* Các môn có tích lũy tín chỉ thì phải >= 4 điểm trờ lên (trong bảng 1 là các cột còn lại trừ cột 52).
* Xếp loại gồm những giá trị {0, 1, 2, 3, 4} tương đương với nghĩa {“Không đạt”, “Trung bình”, “Khá”, “Giỏi”, “Xuất Sắc”}.
  1. PHÂN TÍCH

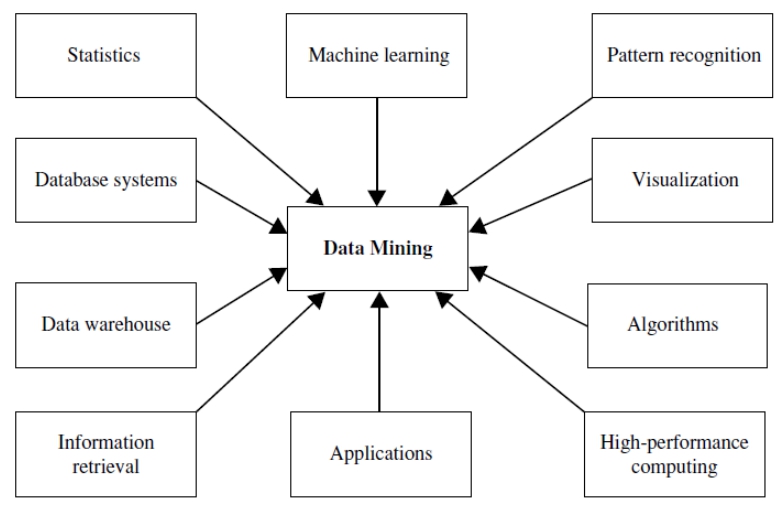
3.1. MÁY HỌC VÀ DATAMINDING

3.1.1. Máy học

Là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

Theo Tom Mitchell, giáo sư nổi tiếng của Đại học Carnegie MellonUniversity – CMU: "Một chương trình máy tính (CT) được xem là học cách thực thi một lớp nhiệm vụ (NV) thông qua trải nghiệm (KN), đối với thang đo năng lực (NL) nếu như dùng (NL) ta đo thấy năng lực thực thi của chương trình có tiến bộ sau khi trải qua (KN)".

3.1.2. Dataminding

Khai thác dữ liệu là sử dụng các kỹ thuật tính toán để phân tích tìm ra các mẫu trong một lượng lớn dữ liệu mà chúng ta khó phát hiện bằng kỹ thuật thông thường.

Hình 1. 1 Các lĩnh vực trong DataMining

*Hình 3. 1 Các lĩnh vực trong DataMining*

Datamining là sự kết hợp của nhiều lĩnh vực gồm:

Một số khái niệm cơ bản cần biết như:

* Data preprocessing (tiền xử lý dữ liệu): nó là quá trình xử lý cái dữ liệu ban đầu mà ta có nhằm cải thiện chất lượng của kết quả khai thác dữ liệu.
* Classificantion (phân lớp): là quá trình gán nhãn cho các mẫu dữ liệu mới với độ chính xác có thể.
* Clustering (gom cụm): quá trình nhóm tích hợp các đối tượng thành những cụm hay nhóm đảm bảo các đối tượng cùng cụm có độ tương tự cao và khác với đối tượng cụm còn lại.
* Association rule (luật kết hợp): là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu.

Các phương pháp khai thác dữ liệu: chia làm 2 nhóm chính

* Kỹ thuật mô tả: mô tả tính chất hoặc đặt tính chung của dữ liệu trong CSDL nó bao gồm:
* Phương pháp phân nhóm (Clustering).
* Phương pháp tổng hợp hóa (Summerization).
* Phương pháp phát hiện sự biến đổi và độ lệch (Change and deviation detection).
* Phương pháp phân tích luật kết hợp (Association Rule)...
* Kỹ thuật dự đoán: đưa ra các dự dón dựa váo các suy diễn trên dữ liệu tạm thời gian các phương pháp
* Phương pháp phân lớp (Classification): phân loại dữ liệu đối tượng mới nếu độ chính xác của bộ phận loại được đánh giá là có thể chấp nhận được.
* Phương pháp hồi quy (Regression): là kỹ thuật thống kê cho phép ước lượng các mối liên kết giữa các biến, mô tả mối liên kết giữa 1 tập các biến dự báo, được chia làm 4 loại: hồi quy tuyến tính và phi tuyến tính, hồi quy đơn biến và da biến, hồi qui có thông số phi thông số và thông số kết hợp, hồi quy đối xứng.

3.2. THUẬT TOÁN PHÂN LỚP

3.2.1. Khái quát về học có giám sát và không giám sát

3.2.1.1. Học có giám sát ( Supervised Learning )

Là một phương pháp của ngành học máy nhằm tìm ra một mô hình phù hợp với các giám sát.

Thuật toán dự đoán đầu ra của một dữ liệu mới dựa trên các bộ đã biết từ trước bộ dữ liệu này còn được gọi là dữ liệu huấn luyện.

Học có giám sát còn được chia nhỏ thành: phân lớp và hồi qui.

Ví dụ: Lọc thư rác, phân loại trang web, dự đoán rủi ro tài chính, dự đoán biến động chỉ số chứng khoán, phát hiện tấn công mạng,…

3.2.1.2. Học không giám sát ( Unsupervised Learning )

Là một phương pháp của ngành học máy nhằm tìm ra một mô hình phù hợp với các giám sát.

Nhãn lớp của tập huấn luyện không biết trước. Trong thuật toán này,chúng ta không biết trước được đầu ra hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán không có giám sát sẽ dựa vào cấu trúc của dữ liệu đề thực hiện một công việc nào đó để thuận tiện trong việc lưu trữ và tính toán.

Học không giám sát còn được chia nhỏ thành: Gom nhóm và kết hợp.

Ví dụ: Phát hiện các cum dữ liệu, cụm tính chất; phát hiện các cộng đồng mạng; phát hiện xu hướng, thị yếu;…

3.2.2. KNN( K-nearest neighbor)

3.2.2.1. Khái niệm

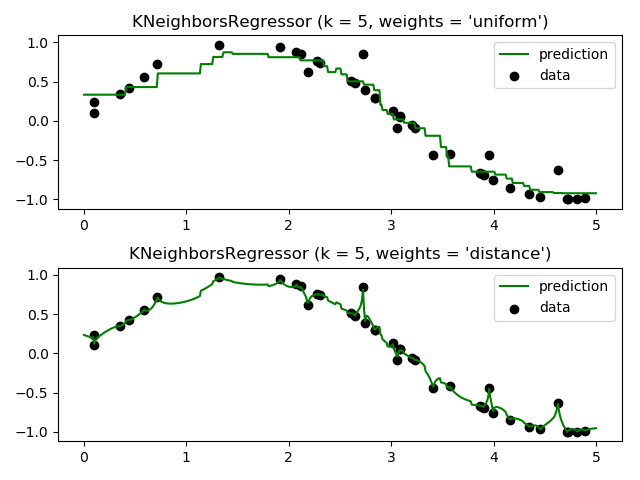
* Là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning
* K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là [Classification](https://machinelearningcoban.com/2016/12/27/categories/#classification-phan-loai) và [Regression](https://machinelearningcoban.com/2016/12/27/categories/#regression-hoi-quy)

3.2.2.2. Khái quát

* Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới (hay kết quả của câu hỏi trong bài thi) được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label. Chi tiết sẽ được nêu trong phần tiếp theo.
* Trong bài toán Regresssion, đầu ra của một điểm dữ liệu sẽ bằng chính đầu ra của điểm dữ liệu đã biết gần nhất (trong trường hợp K=1), hoặc là trung bình có trọng số của đầu ra của những điểm gần nhất, hoặc bằng một mối quan hệ dựa trên khoảng cách tới các điểm gần nhất đó.
* Một cách ngắn gọn, KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu. Hình dưới đây là một ví dụ về KNN trong classification với K = 1.
* Nhận xét:
* Ưu điểm:
* Độ phức tạp tính toán của quá trình training là bằng 0.
* Việc dự đoán kết quả của dữ liệu mới rất đơn giản.
* Không cần giả sử gì về phân phối của các class.
* Khuyết điểm:
* KNN rất nhạy cảm với nhiễu khi K nhỏ.
* Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới *từng* điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

3.2.2.3. KNN cho Regression

Với bài toán Regression, chúng ta cũng hoàn toàn có thể sử dụng phương pháp tương tự: ước lượng đầu ra dựa trên đầu ra và khoảng cách của các điểm trong K-lân cận. Việc ước lượng như thế nào các bạn có thể tự định nghĩa tùy vào từng bài toán.



KNN cho bài toán Regression

3.2.2.4. Chuẩn hoá dữ liệu

Khi có một thuộc tính trong dữ liệu (hay phần tử trong vector) lớn hơn các thuộc tính khác rất nhiều (ví dụ thay vì đo bằng cm thì một kết quả lại tính bằng mm), khoảng cách giữa các điểm sẽ phụ thuộc vào thuộc tính này rất nhiều. Để có được kết quả chính xác hơn, một kỹ thuật thường được dùng là Data Normalization (chuẩn hóa dữ liệu) để đưa các thuộc tính có đơn vị đo khác nhau về cùng một khoảng giá trị, thường là từ 0 đến 1, trước khi thực hiện KNN. Có nhiều kỹ thuật chuẩn hóa khác nhau, các bạn sẽ được thấy khi tiếp tục theo dõi Blog này. Các kỹ thuật chuẩn hóa được áp dụng với không chỉ KNN mà còn với hầu hết các thuật toán khác.

3.2.2.4. Sử dụng các phép đo khoảng cách khác nhau

Ngoài norm 1 và norm 2 tôi giới thiệu trong bài này, còn rất nhiều các khoảng cách khác nhau có thể được dùng. Một ví dụ đơn giản là đếm số lượng thuộc tính khác nhau giữa hai điểm dữ liệu. Số này càng nhỏ thì hai điểm càng gần nhau. Đây chính là [giả chuẩn 0](https://machinelearningcoban.com/math/#norm0) mà tôi đã giới thiệu trong Tab [Math](https://machinelearningcoban.com/math/).

3.2.2.5. Phân loại

* Fine KNN: Phân biệt chi tiết ở các lớp, số lượng hàng xóm được đặt thành 1.
* Medium KNN: Phân. Biệt trung bình giữa các lớp, số lượng hàng xóm được đặt thành 10.
* Coarse KNN: Phân biệt thô giữa các lớp, số lượng hàng xóm được đặt thành 100.
* Cosine KNN: Phân biệt trung bình giữa các lớp, sử dụng thước đo khoảng cách là Csine, số lượng hàng xóm được đặt thành 10.
* Cubic KNN: Phân biệt trung bình cứng giữa các lớp, sử dụng thước đo khoảng cách khối, số lượng hàng xóm được đặt thành 10.
* Weighted KNN: Phân biệt trung bình cứng giữa các lớp, sử dụng trọng lượng khoảng cách , số lượng hàng xóm được đặt thành 10.

3.2.3. SVM (Suport Vector Machine)

**Ưu điểm của SVM là gì?**

Là một kĩ thuật phân lớp khá phổ biến, SVM thể hiện được nhiều ưu điểm trong số đó có việc tính toán hiệu quả trên các tập dữ liệu lớn. Có thể kể thêm một số ưu điểm của phương pháp này như:

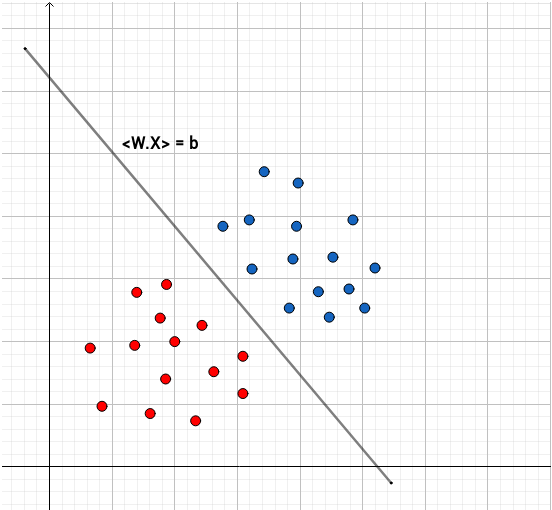
* **Xử lý trên không gian nhiều chiều**: SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn.
* **Tiết kiệm bộ nhớ**: Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.
* **Tính linh hoạt** - phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.

**Nhược điểm của SVM là gì?**

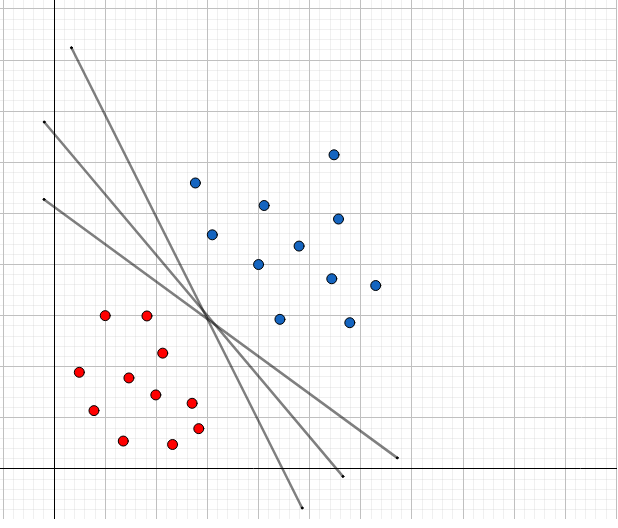
* **Bài toán nhiều chiều**: Trong trường hợp số lượng thuộc tính (**p**) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (**n**) thì SVM cho kết quả không khả quan.
* **Chưa thể hiện rõ tính xác suất**: Việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Tuy nhiên hiệu quả của việc phân lớp có thể được xác định dựa vào khái niệm **margin** từ điểm dữ liệu mới đến siêu phẳng phân lớp mà nhóm chúng em đã bàn luận ở trên.

3.2.3.1. Khái niệm

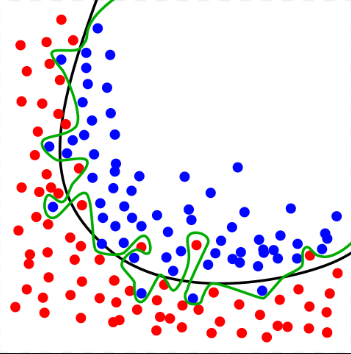
SVM là một thuật toán học có giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng em vẽ đồi thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng dữ liệu có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" phân chia các lớp. Đường bay - nó chỉ hiểu đơn giản là 1 đường thằng có thể phân chia các lớp ra thành hai phần riêng biệt.



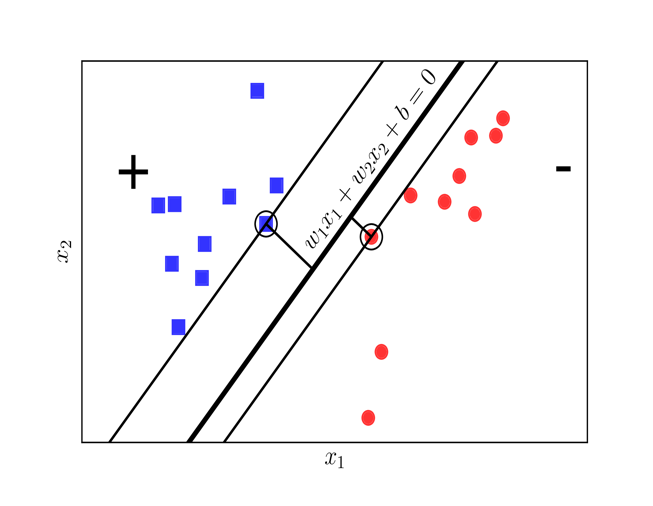
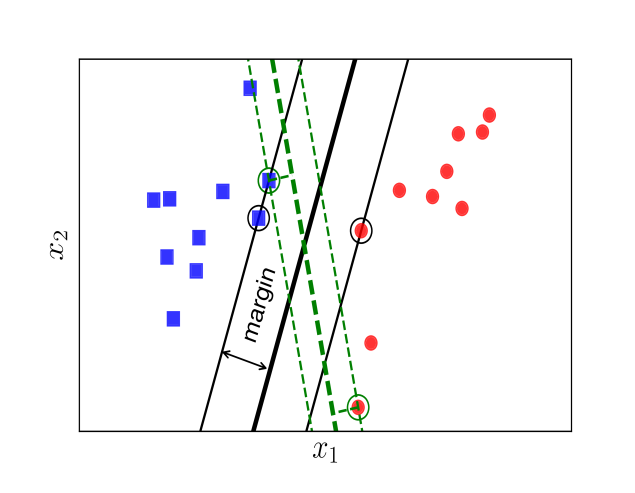
*Hình 3.2 1: Phân lớp tuyến tính*



*Hình 3.2 2: Các mặt phân cách hai classes linearly separable*



*Hình 3.2 3: Phân lớp phi tuyến*



***Hình 3.2 4: Margin của hai classes là bằng nhau và lớn nhất có thể***

Margin là khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng với các phân lớp.

Hình 3.2.4 (bên trái) class tròn đỏ sẽ không được hạnh phúc cho lắm vì đường phân chia gần nó hơn class vuông xanh rất nhiều. Chúng ta cần một đường phân chia sao cho khoảng cách từ điểm gần nhất của mỗi class (các điểm được khoanh tròn) tới đường phân chia là như nhau, như thế thì mới công bằng. Khoảng cách như nhau này được gọi là margin (lề).

Hình 3.2.4 (bên phải) khi khoảng cách từ đường phân chia tới các điểm gần nhất của mỗi class là như nhau. Xét hai cách phân chia bởi đường nét liền màu đen và đường nét đứt màu lục, rõ ràng đường nét liền màu đen nó tạo ra một margin rộng hơn.

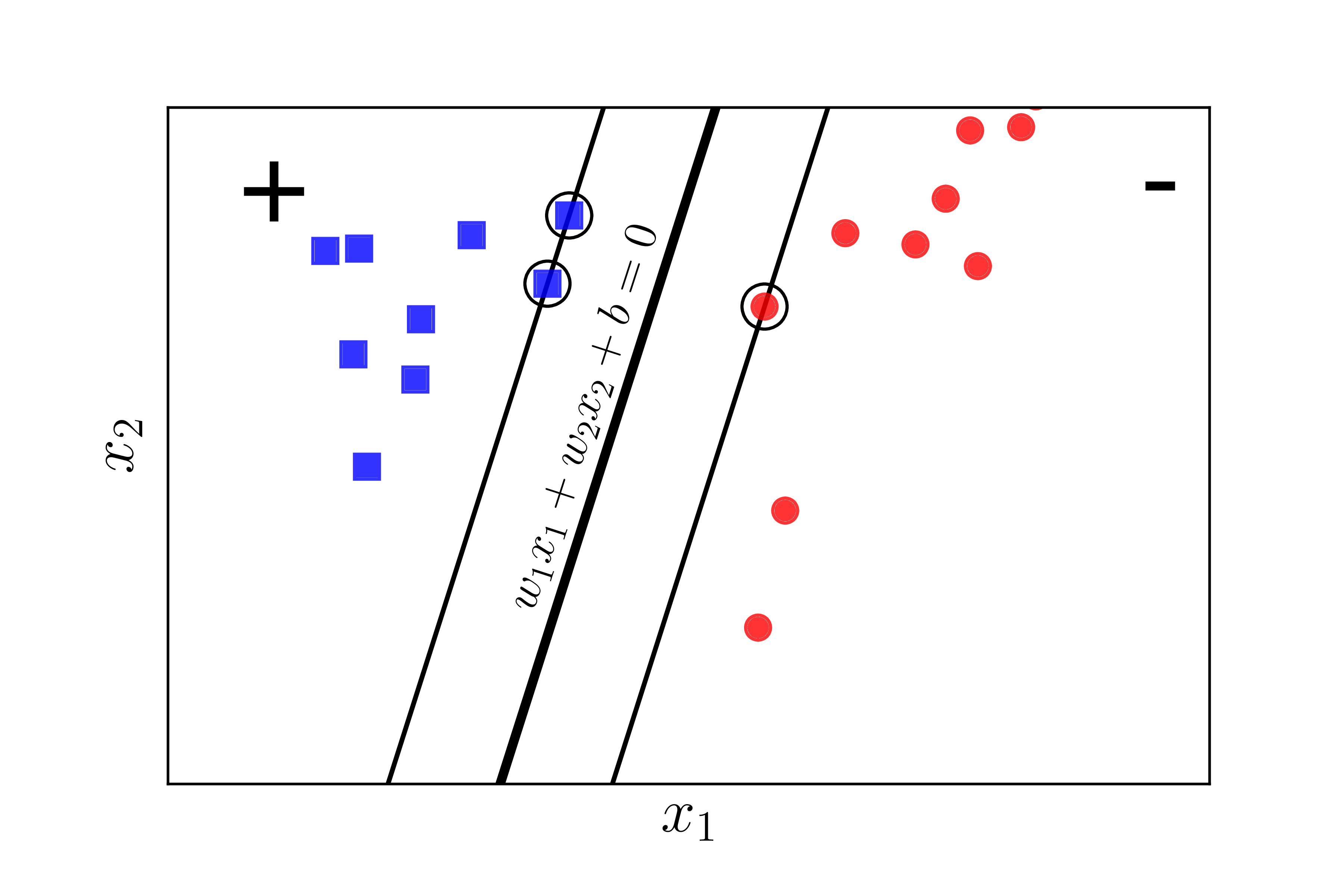
Việc margin rộng hơn sẽ mang lại hiệu ứng phân lớp tốt hơn vì sự phân chia giữa hai classes là rõ ràng hơn.

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.

3.2.3.2. Xây dựng bài toán tối ưu cho SVM

Giả sử rằng các cặp dữ liệu của training set là (x1,y1), (x2,y2), …, (xn,yn) với vector xi∈Rd thể hiện đầu vào của một điểm dữ liệu và yi là nhãn của điểm dữ liệu đó. Trong đó, d là số chiều của dữ liệu và n là số điểm dữ liệu. Giả sử rằng nhãn của mỗi điểm dữ liệu được xác định bởi yi=1 (class 1) hoặc yi=−1 (class 2) giống như trong PLA.

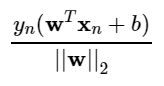
Để giúp các bạn dễ hình dung, chúng ta cùng xét trường hợp trong không gian hai chiều dưới đây. Không gian hai chiều để các bạn dễ hình dung, các phép toán hoàn toàn có thể được tổng quát lên không gian nhiều chiều.



*Hình 3.2 5: Phân tích bài toán SVM*

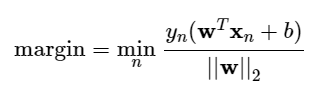
Giả sử rằng các điểm vuông xanh thuộc class 1, các điểm tròn đỏ thuộc class -1 và mặt wTx + b = w1x1 + w2x2 + b = 0 là mặt phân chia giữa hai classes (Hình 3.2.5). Hơn nữa, class 1 nằm về phía dương, class -1 nằm về phía âm của mặt phân chia. Nếu ngược lại, ta chỉ cần đổi dấu của w và b. Chú ý rằng chúng ta cần đi tìm các hệ số w và b.

Ta quan sát thấy một điểm quan trọng sau đây: với cặp dữ liệu (xn,yn) bất kỳ, khoảng cách từ điểm đó tới mặt phân chia là:

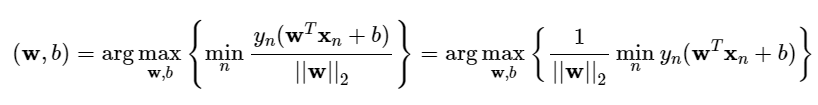


Điều này có thể dễ nhận thấy vì theo giả sử ở trên, yn luôn cùng dấu với phía của xn. Từ đó suy ra yn cùng dấu với wTx + b, và tử số luôn là 1 số không âm.

Với mặt phần chia như trên, margin được tính là khoảng cách gần nhất từ 1 điểm tới mặt đó (bất kể điểm nào trong hai classes):

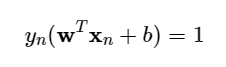


Bài toán tối ưu trong SVM chính là bài toán tìm w và b sao cho margin này đạt giá trị lớn nhất:

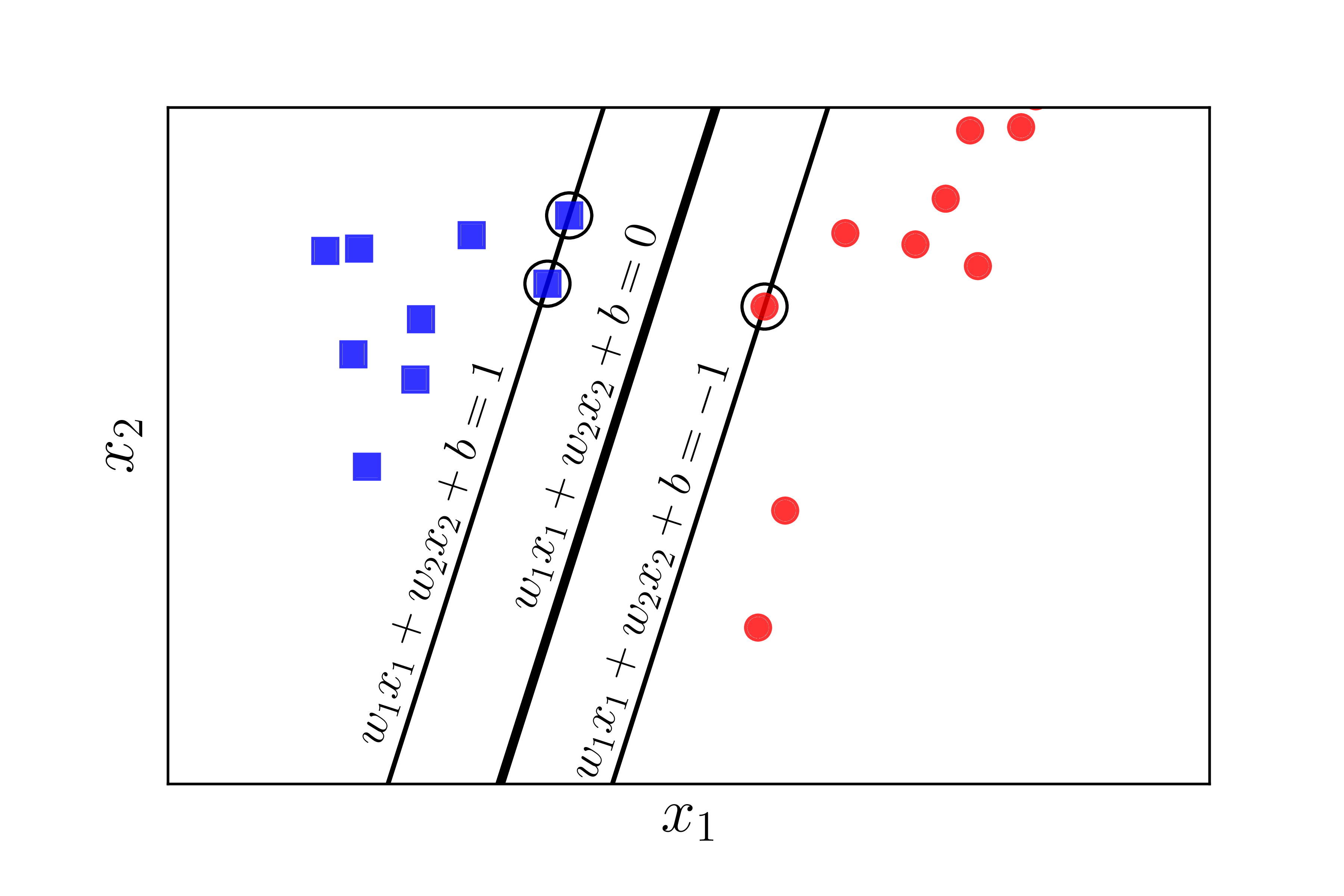


Việc giải trực tiếp bài toán này sẽ rất phức tạp, nhưng các bạn sẽ thấy có cách để đưa nó về bài toán đơn giản hơn.

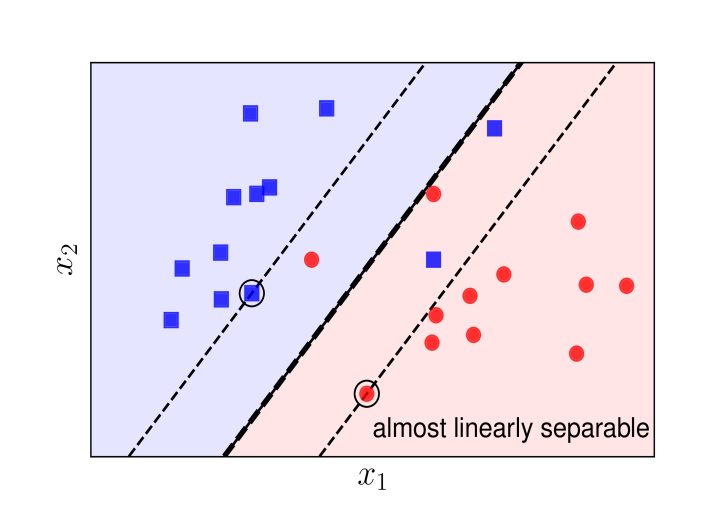
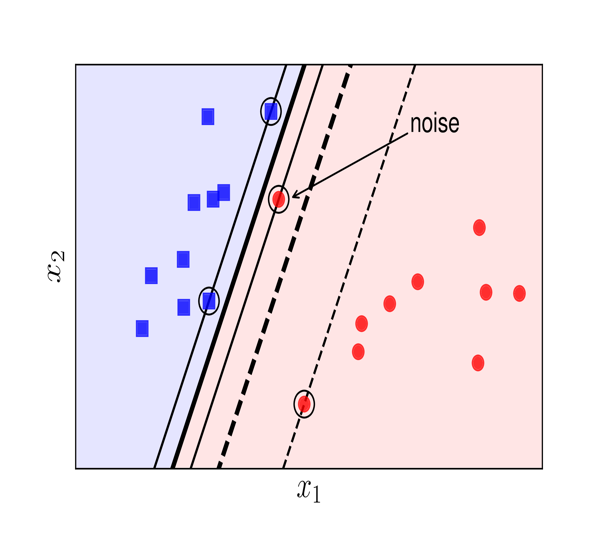
Nhận xét quan trọng nhất là nếu ta thay vectỏ hệ số w bởi kw và b bởi kb trong đó k là một hằng số dương thì mặt phân chia không thay đổi, tức khoảng cách từ từng điểm đến mặt phân chia không đổi, tức margin không đổi. Dựa trên tính chất này, ta có thể giả sử:



**Với những điểm nằm gần mặt phân chia nhất** như Hình 4 dưới đây:



*Hình 3.2 6: Các điểm gần mặt phân cách nhất của hai classes được khoanh tròn.*

3.2.3.3. Soft Margin SVM

***Hình 3.2 7: Soft margin SVM. Khi a) có nhiễu hoặc b) dữ liệu gần linearly separable, SVM thuần sẽ không hoạt động hiệu quả.***

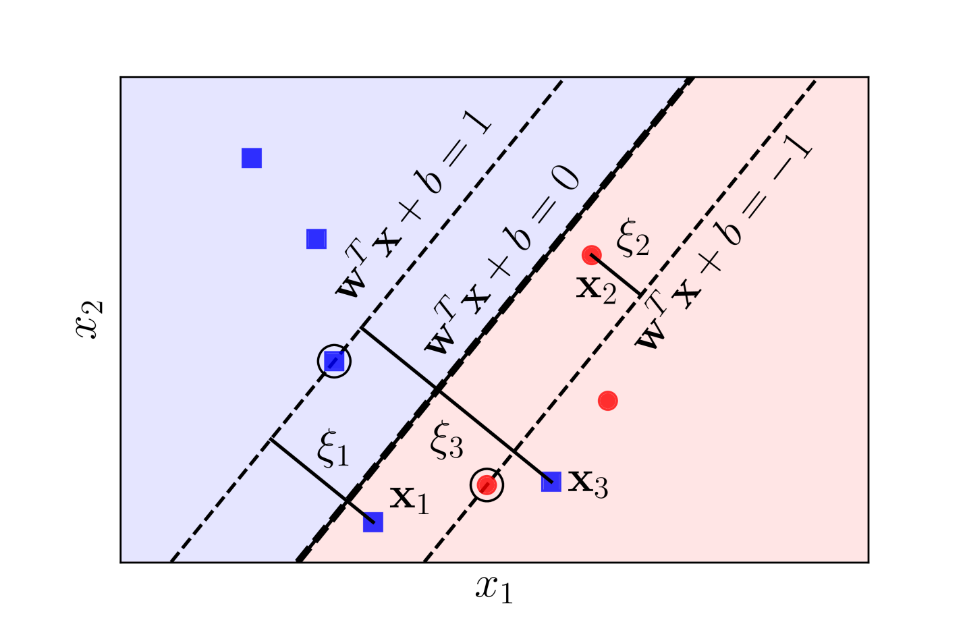
Trong cả hai trường hợp trên, margin tạo bởi đường phân chia và đường nét đứt mảnh còn được gọi là soft margin (biên mềm). Cũng theo cách gọi này, SVM thuần còn được gọi là Hard Margin SVM (SVM biên cứng).

Trong bài này, chúng ta sẽ tiếp tục tìm hiểu một biến thể của Hard Margin SVM có tên gọi là Soft Margin SVM.

3.2.3.4. Phân tích toán học

Như đã đề cập phía trên, để có một margin lớn hơn trong Soft Margin SVM, chúng ta cần hy sinh một vài điểm dữ liệu bằng cách chấp nhận cho chúng rơi vào vùng không an toàn. Tất nhiên, chúng ta phải hạn chế sự hy sinh này, nếu không, chúng ta có thể tạo ra một biên cực lớn bằng cách hy sinh hầu hết các điểm. Vậy hàm mục tiêu nên là một sự kết hợp để tối đa margin và tối thiểu sự hy sinh.

Giống như với Hard Margin SVM, việc tối đa margincó thể đưa về việc tối thiểu (||w||2)2. Để xác định sự hy sinh, chúng ta cùng theo dõi Hình 3.2.8 dưới đây:

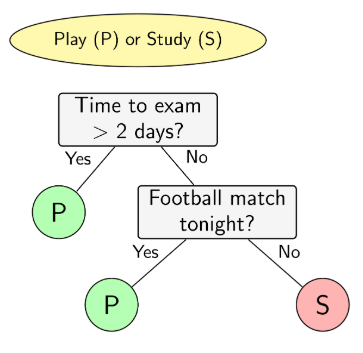


*Hình 3.2 8: Giới thiệu các biến slack ξn*

Với những điểm nằm trong vùng an toàn, ξn=0. Những điểm nằm trong vùng không an toàn nhưng vẫn đúng phía so với đường phân chia tương ứng với các 0<ξn<1, ví dụ x2. Những điểm nằm ngược phía với class của chúng so với đường boundary ứng với các ξn>1, ví dụ như x1 và x3.

3.2.4. C4.5 (Information Gain Ratio)

3.2.4.1. Khái niệm



*Hình 3.2 9: Ví dụ mô hình cây quyết định**.*

Là một thuật toán được sử dụng để tạo ra một [cây quyết định](https://en.wikipedia.org/wiki/Decision_tree) được phát triển bởi [Ross Quinlan](https://en.wikipedia.org/wiki/Ross_Quinlan).

Là sự cải tiến của thuật toán phân lớp ID3.

3.2.4.2. Diễn giải

* Ở ID3:
* Chọn thuộc tính có độ lợi thông tin cao nhất.
* pi là xác suất để một mẫu bất kỳ của D thuộc về lớp Ci được tính bởi |C*i*, D|/|D|.
* Thông tin kỳ vọng để phân lớp một mẫu trong D là:
* Chọn thuộc tính để phân chia tập dữ liệu: xét hết các thuộc tính A của tập dữ liêu huấn luyện ngoại trừ thuộc tính phân lớp theo công thức:

* Độ lợi thông tin dựa trên phân chia theo thuộc tính A theo công thức:

* Ở C4.5:
  + Kết hợp giải thuật phân lớp ID3.
  + Nếu thuộc tính có nhiều giá trị thì độ đo Information Gain tạo ra cây nhiều nhánh cây không tốt dẫn đến cần chuẩn hóa.
  + Chọn thuộc tính có độ đo Gain Ratio lớn nhất làm thuộc tính phân chia.

3.2.4.3. Ưu điểm

Xử lý cả thuộc tính liên tục và rời rạc.

Xử lý dữ liệu đào tạo với các giá trị thuộc tính bị.

Xử lý các thuộc tính với các chi phí khác nhau.

Cắt tỉa cây sau khi tạo.

3.2.4.4. Nhược điểm

Không đảm bảo xây dựng được cây tối ưu.

Có thể gặp overfitting (tạo ra cây quá phức tạp và quá khớp với dữ liệu huấn luyện).

Thường ưu tiên thuộc tính có nhiều giá trị.

3.3. QUY TRINH THỰC HIỆN DỰ ĐOÁN

Bước 1: Thực hiện trên excel, gồm:

* Lấy danh sách các khóa 03DHTH, 04DHTH, 05DHTH gộp thành một danh sách chung.
* Tiền xử lý: bỏ cột môn học có năm thứ 4 (nghĩa là học kì 7,8) và cột điểm trung bình tổng hệ số 4 toàn học kì. Chuyển thành “DSSV\_3K.csv”.

Bước 2: Thực hiện trên matlab, gồm:

* Tải dữ liệu danh sách “DSSV\_3K.scv” lên.
* Tiền xử lý: Chuyển cột 52={0,1,2,3,4} thành 52={0,1}. (trong đó 0 là “không đậu tốt nghiệp” và 1 là “đậu tốt nghiệp” )
* Lớp 0 giữ nguyên còn lớp {1,2,3,4} chuyển thành 1.
* Trộn đều danh sách 10 lần.
* Tách những dòng dữ liệu theo cột 52={0,1} ra làm 2 phần lớp 0 và 1 riêng.
* Phân thành train và test.
* Mỗi phần phân ra thành 70:30 của {0,1}. (trong só train gồm dữ liệu của 70%{0,1} và phần còn lại là test). Lấy cột phân lớp 52 ra.
* Thực hiện cho mô hình học có giám sát gồm những thuật toán: Naïve Bayes, SVM và C4.5.
* Lấy phần dữ liệu train để huấn luyện máy học. Dựa vào đó kiểm tra test từng thuật toán.
* Accuracy, Confusion Matrix từng thuật toán. (\*)
* Lấy phần trăm dự đoán của thuật toán phân lớp tốt nhất dự đoán cho khóa 06. (ở đây là thuật toán svm)
* Đếm số lượng 0 và 1.

Bước 3: Kiểm tra trên weka xem thực hiện có khớp với phần trăm (\*) hay không?

* 1. KẾT QUẢ THỬ NGHIỆM

4.1. Naive Bayes

Accuracy Naive Bayes = 98.742% (time test set in weka: 0.01s)

|  |  |  |  |
| --- | --- | --- | --- |
| "Yes" | "No" | "ClassDistribute" | [] |
| 119 | 0 | "Yes\_test = 1" | 119 |
| 0 | 40 | "No\_test = 0" | 40 |

*Bảng 4.1:* *Confusion Matrix Naive Bayes*

4.2. SVM

**Accuracy SVM =100% (time test set in weka: 0.01s)**

|  |  |  |  |
| --- | --- | --- | --- |
| "Yes" | "No" | "ClassDistribute" | [] |
| 117 | 2 | "Yes\_test = 1" | 119 |
| 0 | 40 | "No\_test = 0" | 40 |

*Bảng 4.2: Confusion Matrix SVM*

4.3. C4.5

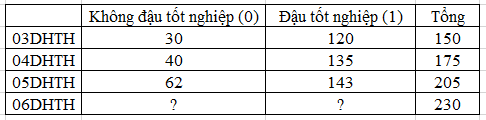
Accuracy C4.5 = 96.226% (time test set in weka: < 0.01s)

|  |  |  |  |
| --- | --- | --- | --- |
| "Yes" | "No" | "ClassDistribute" | [] |
| 117 | 2 | "Yes\_test = 1" | 119 |
| 4 | 36 | "No\_test = 0" | 40 |

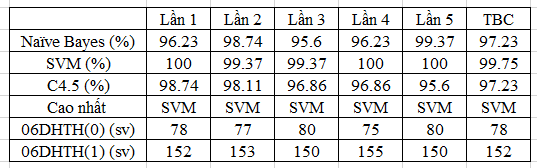
*Bảng 4.3:* *Confusion Matrix C4.5*

MẪU KHẢO SÁT DỰ ĐOÁN

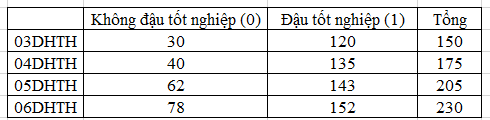
BẢNG THỐNG KÊ DANH SÁCH SINH VIÊN ĐẠI HỌC (dvt: sinh viên)



BẢNG MẪU DỰ ĐOÁN TRONG 5 LẦN



BẢNG KẾT QUẢ (dvt: sinh viên)



**Nhận xét:**

Từ mẫu khảo sát trên, bài toán đã dự đoán trong tổng số 230 sinh viên khóa 06DHTH thì có 78 là sinh viên không đạt yêu cầu để xét tốt nghiệp (chiếm 33.913% trong tổng số) và 152 còn lại sinh viên có khả năng đạt tốt nghiệp sớm (chiếm 66.087% trong tổng số).

Qua dự đoán trên cho thấy tỉ lệ không đạt của sinh viên khóa 06 hiện nay rất cao. Cần phải có những biện pháp hỗ trợ trong việc học tập nhằm giảm bớt tình trạng không đạt này; nếu không khắc phục kịp thời thì tình trạng này sẽ xảy ra và ảnh hưởng đến nhiều thế hệ sau nữa.

TỔNG KẾT

Dựa vào kết quả trên cho thấy thuật toán SVM có dự đoán tốt hơn so với 2 thuật toán Naive Bayes và C4.5.

Tiếp tục trao dồi kiến thức, kỹ thuật lập trình và tìm hiểu sâu hơn các thuật toán trên matlab để giải quyết các bài khó hơn làm trên nhiều phân lớp.

Tiền để chuẩn bị tốt cho luận án, đổ án tốt nghiệp.

Mở rộng phát triển hơn ở mảng phân lớp trên nhiều thuộc tính (gồm không đạt, trung bình, khá, giỏi, xuất sắc) như dữ liệu trên.

TÀI LIỆU THAM KHẢO

* Sách JiaweiHan-Data Mining Concepts and Techniques 3rd Edition-2012.
* Slide bài giảng Data mining trên lớp.
* Dữ liệu rời rạc và dữ liệu liên tục:  
  [1] [https://stats.stackexchange.com/questions/206/what-is-the-difference-between-discrete-data-and-continuous-data](https://stats.stackexchange.com/questions/206/what-is-the-difference-between-discrete-data-and-continuous-data?fbclid=IwAR1LhGLdOFJFTKKY_js5JgySNnFqL6rgLIyxiAxn6nzyanrT-45fNaVWu34)
* Weka:  
  [2] [https://www.slideshare.net/butest/machine-learning-with-weka](https://www.slideshare.net/butest/machine-learning-with-weka?fbclid=IwAR2uXOzcTjtiP6lvth6UGdZZvlpSETqdRDcHlCkuF9C41ZYFottvKKhNgIU)

[3] [http://weka-jp.info/itej/](http://weka-jp.info/itej/?fbclid=IwAR1aKyzjNFF5Rg6jqMcKsdh5m90BTzk5XF99h-2Rwex25Z4ox5UqJ1j7t_E)

* C4.5:  
  [4] [http://ait.edu.vn/Hoc\_thuat/C45.html](http://ait.edu.vn/Hoc_thuat/C45.html?fbclid=IwAR1EAcDAwfQXFunReYUbF-bWt8CzlCG2g9CPTFp7WNKYfAhBa_DLZwiPqpU)

[5] <https://en.wikipedia.org/wiki/C4.5_algorithm>

* SVM:  
  [6] [https://www.youtube.com/watch?v=l7Tr2OyXNU8](https://www.youtube.com/watch?v=l7Tr2OyXNU8&fbclid=IwAR0uME0kPDioDL_-tqFuZAOCK5cg1474dxlLOEWmphxXy7zhM0Bsit1Sq58)

[7] <https://stats.stackexchange.com/questions/82923/mixing-continuous-and-binary-data-with-linear-svm>

[8]<https://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html?fbclid=IwAR2SpPlD_E14iR-wInLd8T68NOsBGsSEqi8pymQSWG8JK6tgJDwDIPq3d4Q#weka>

* Naive Bayes:  
  [9] [https://www.youtube.com/watch?v=XcwH9JGfZOU](https://www.youtube.com/watch?v=XcwH9JGfZOU&fbclid=IwAR03OqKfo4I-uX6_k0p635JEI5wP_0a_WUuqWp-wRjFN48f_5Ux2kN4ELrU)
* Một số các trang web khác liên quan:

[10] <https://machinelearningcoban.com/about/>

[11] [https://www.cs.waikato.ac.nz/ml/weka/courses.html](https://www.cs.waikato.ac.nz/ml/weka/courses.html?fbclid=IwAR18qesLSDRCKYqeJ5xvWdgz8kaZ414MDUTBjg2GPVM8FWdXFWFz07DRiqM)

[12] https://www.slideshare.net/TrngHVit/artificial-intelligence-ai-l9hoc-may