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 đợi 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? Hãy cùng nhóm em tìm hiểu bài dự đoán dưới đây để có thể ước chừng, dự đoán được phần nào 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

	Thùy 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

MỤC LỤC

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	→	Gom cụm
		Association rule	→	Luật kết hợp
		Summerization	→	Tổng hợp hóa
		Change and deviation detection	→	Phát hiện sự biến đổi và độ lệch
		Regression	→	Hồi quy
SVM	→	Suport Vector Machine		
		Classes linearly	→	Phân lớp tuyến tính
C4.5	→	Information Gain Ratio		
		Soft margin	→	Biên mềm
		Hard Margin	→	Biên cứng

DANH MỤC HÌNH ẢNH

DANH MỤC CÁC BẢNG

CHƯƠNG 1: 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, Naive Bayes, 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.

CHƯƠNG 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).

CHƯƠNG 3: PHÂN TÍCH

3.1. MÁY HỌC VÀ DATAMINDING

3.1.1. Máy học

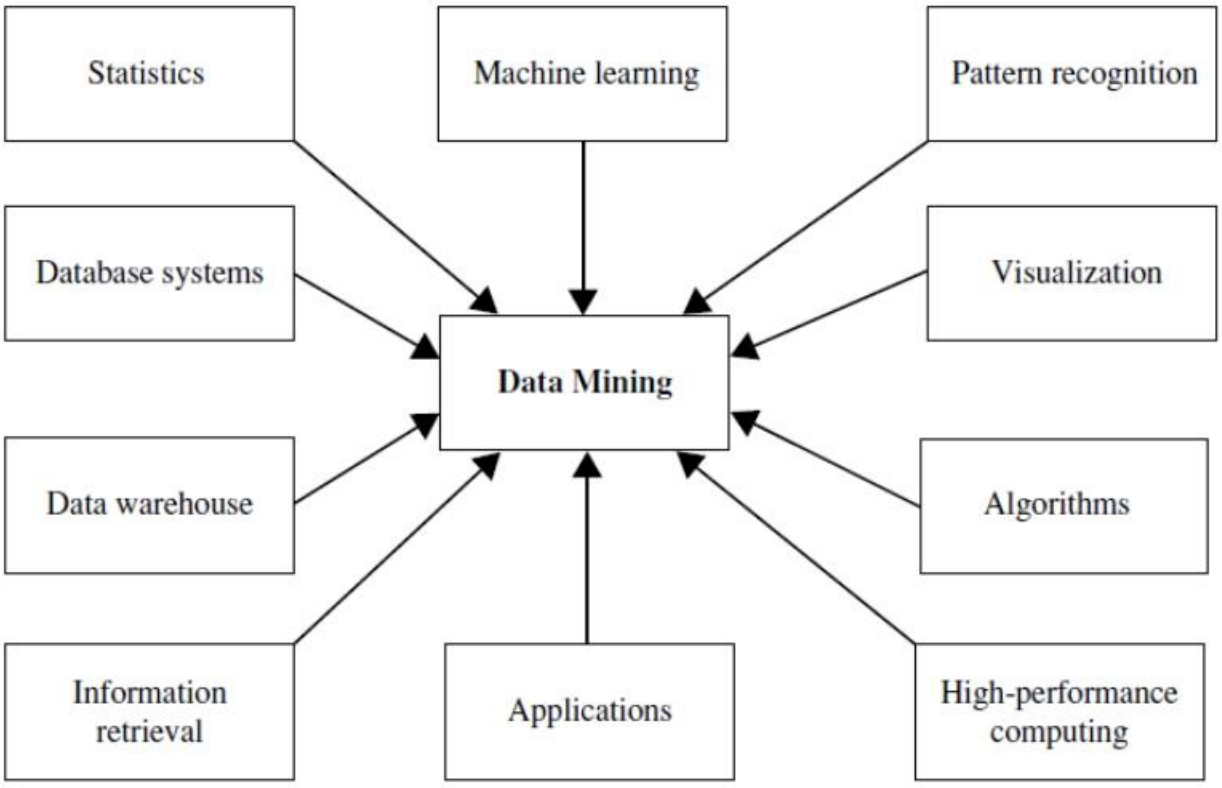
Là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

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.

Datamining là sự kết hợp của nhiều lĩnh vực gồm:



Hình 3. 1 Các lĩnh vực trong DataMining

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. Naive Bayes

3.2.2.1. Khái niệm

Là phân lớp dựa trên thông kê thực hiện việc dự đoán xác suất của một mẫu thuộc về lớp nào dựa trên giá trị của thuộc tính biết trước.

Có khả năng thực thi hiệu quả so với cây quyết định hay mạng neuron.

Chạy nhanh trên cơ sở dữ liệu với độ chính xác cao.

Xem các thuộc tính xảy ra độc lập với nhau.

3.2.2.2. Định lý

Gọi X là mẫu dữ liệu chưa biết nhãn.

C_i là giả thuyết X thuộc về phân lớp C_i.

Việc phân lớp là quá trình xác định P(Ci|X), xác suất mà giả thuyết đúng với mẫu dữ liệu X cho trước.

$P(C_i)$ là xác suất có thể ước lượng từ dữ liệu huấn luyện.

P(X) là xác suất mẫu dữ liệu được quan sát.

$P(X|C_i)$ là khả năng quan sát mẫu X khi cho trước giả thuyết về phân lớp.

- Định lý Bayes:

$$\boldsymbol{P(C_i|X)} = \frac{\boldsymbol{P(X|C_i).P(C_i)P(X)}}{\boldsymbol{P(X)}}$$

- Dự đoán X thuộc về lớp C_i khi và chỉ khi $P(C_i|X)$ là cao nhất trong số $P(C_m|X)$ của tất cả m lớp.
- Do P(X) là hằng số cho mọi lớp nên chỉ cần tìm cực đại của:

$$\boldsymbol{P(C_i|X)} = \frac{\boldsymbol{P(X|C_i).P(C_i)}}{\boldsymbol{P(X)}}$$

- Naive Bayes:

$$\boldsymbol{P(X|C_i)} = \prod_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{P(x_k|C_i)} = \boldsymbol{P(x_1|C_i)} * \boldsymbol{P(x_2|C_i)} * \ldots * \boldsymbol{P(x_n|C_i)}$$

- Naive Bayes giả sử giá trị của mọi thuộc tính đều độc lập nên:
    - Một tập học D_train, trong đó mỗi ví dụ học x được biểu diễn là một vectơ n chiều: (x_1, x_2, ..., x_n).
    - Một tập xác định các nhãn lớp: C={c_1, c_2, ..., c_m}.
    - Với một mẫu mới X□? X sẽ được phân vào lớp nào.

➢ Nhận xét:

- Ưu điểm:
    - Dễ dàng thực thi.
    - Đạt được kết quả khá tốt trong hầu hết các trường hợp.
- Khuyết điểm: Việc giả sử các thuộc tính độc lập có thể sẽ làm giảm độ chính xác vì thực tế có thể tồn tại sự phụ thuộc giữa chúng.

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 số chiều cao**: 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 dịnh.
- **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.

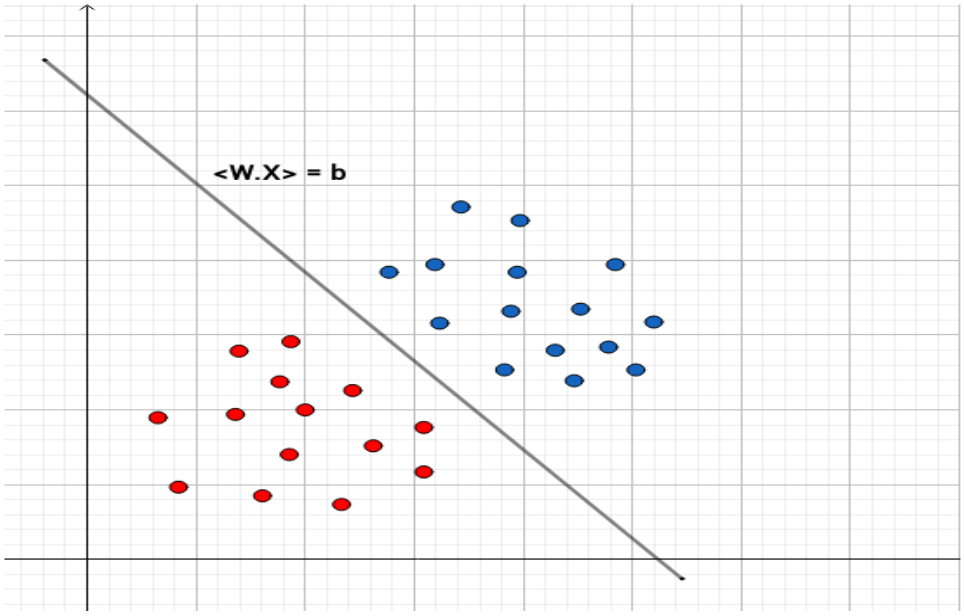
Nhược điểm của SVM là gì?

- **Bài toán số chiều cao**: 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 khá tồi..
- **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 chúng ta đã bàn luận ở trên.

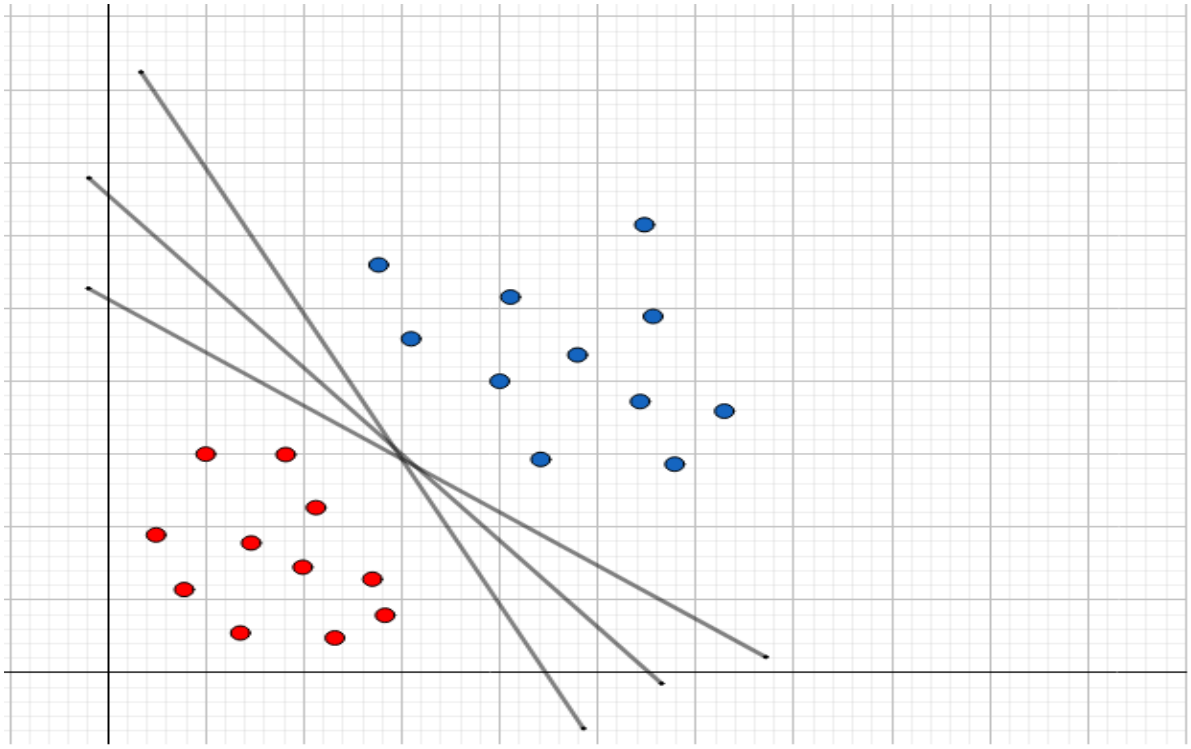
3.2.3.1. Khái niệm

SVM là một thuật toán 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 ta vẽ đồi thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay"

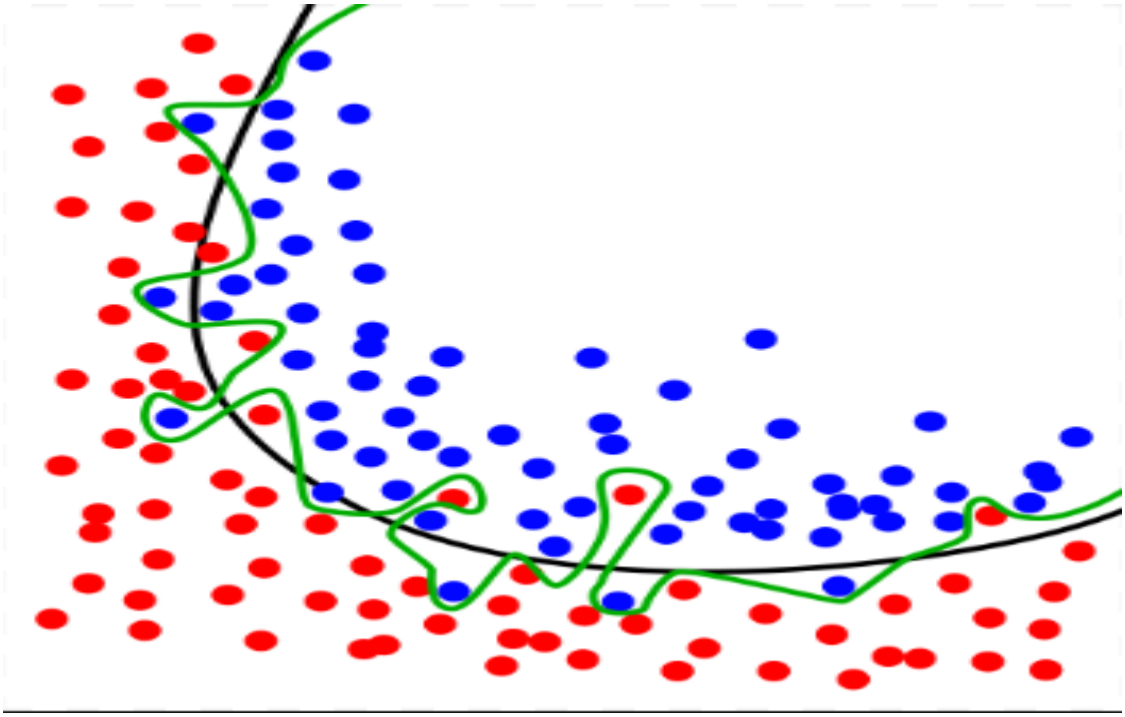
phân chia các lớp. Đường bay - nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.



Hình 3.2 1: Phân lớp tuyến tính

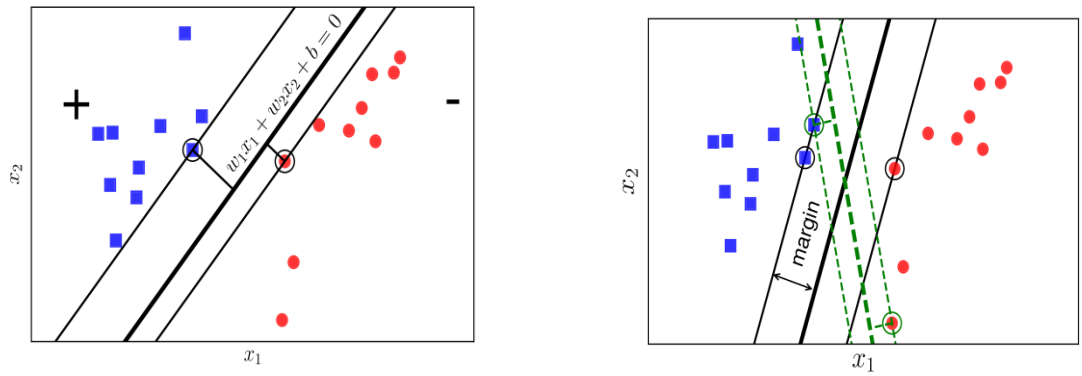


Hình 3.2 2: Các mặt phân cách hai classes linearly separable



Hình 3.2 3: Phân lớp phi tuyến

Margin trong SVM là gì?



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.

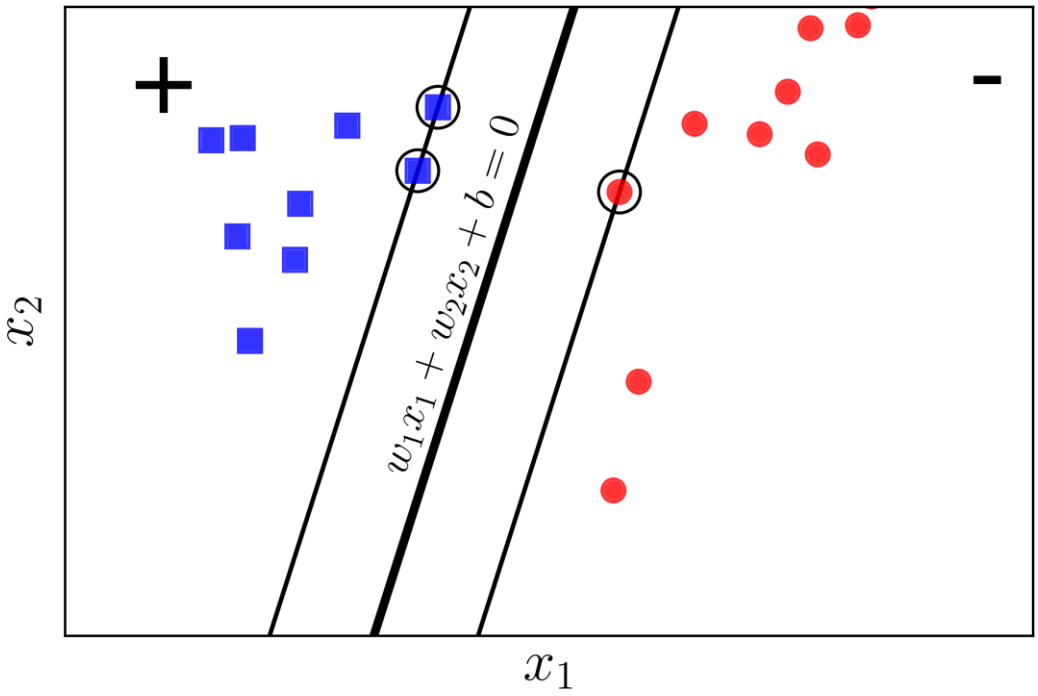
Việc margin rộng hơn sẽ mang lại hiệu ứng phân lớp tốt hơn vì sự phân chia giữa hai classes là rạch ròi hơn.

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. Nguồn gốc của tên gọi Support Vector Machine sẽ sớm được làm sáng tỏ.

3.2.3.2. Xây dựng bài toán tối ưu cho SVM

Giả sử rằng các cặp dữ liệu của training set là $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ với vector $x_i \in R_d$ thể hiện đầu vào của một điểm dữ liệu và y_i là nhãn của điểm dữ liệu 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 $y_i=1$ (class 1) hoặc $y_i=-1$ (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 $w^Tx + b = w_1x_1 + w_2x_2 + b = 0$ 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 (x_n,y_n) bất kỳ, khoảng cách từ điểm đó tới mặt phân chia là:

$$\frac{y_n(\mathbf{w}^T\mathbf{x}_n + b)}{||\mathbf{w}||_2}$$

Điều này có thể dễ nhận thấy vì theo giả sử ở trên, ynyn luôn cùng dấu với phía của x_n. Từ đó suy ra y_n cùng dấu với $w^Tx + b$, 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):

$$\text{margin} = \min_n \frac{y_n(\mathbf{w}^T\mathbf{x}_n + b)}{||\mathbf{w}||_2}$$

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

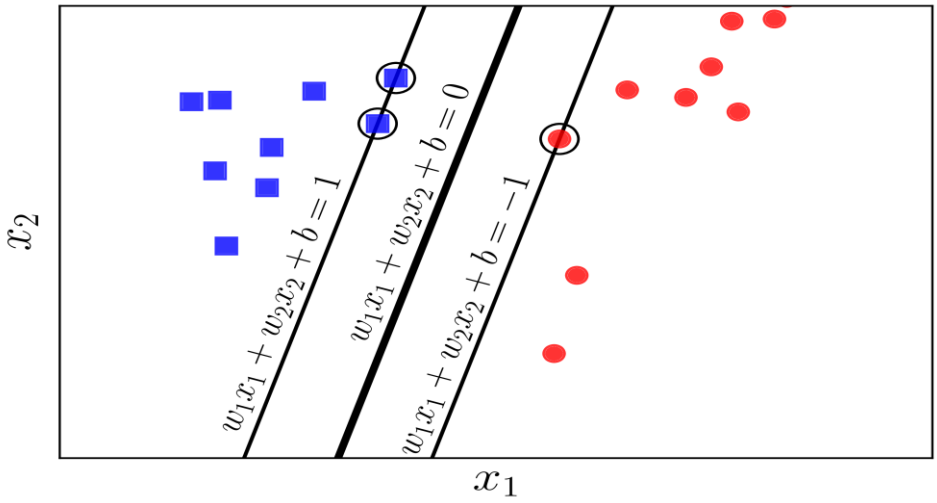
$$(\mathbf{w}, b) = \arg\max_{\mathbf{w},b} \left\{ \min_n \frac{y_n(\mathbf{w}^T\mathbf{x}_n + b)}{||\mathbf{w}||_2} \right\} = \arg\max_{\mathbf{w},b} \left\{ \frac{1}{||\mathbf{w}||_2} \min_n y_n(\mathbf{w}^T\mathbf{x}_n + b) \right\}$$

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 k_w và b bởi k_b 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:

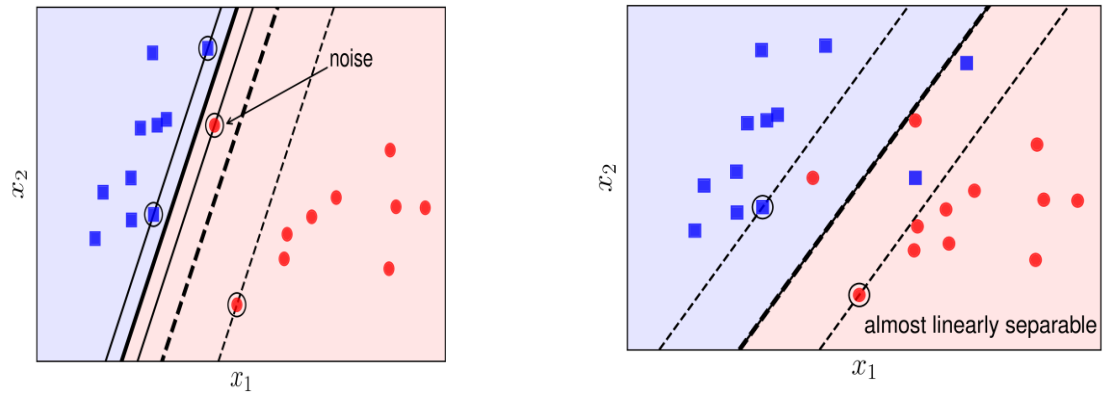
$$y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$

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.

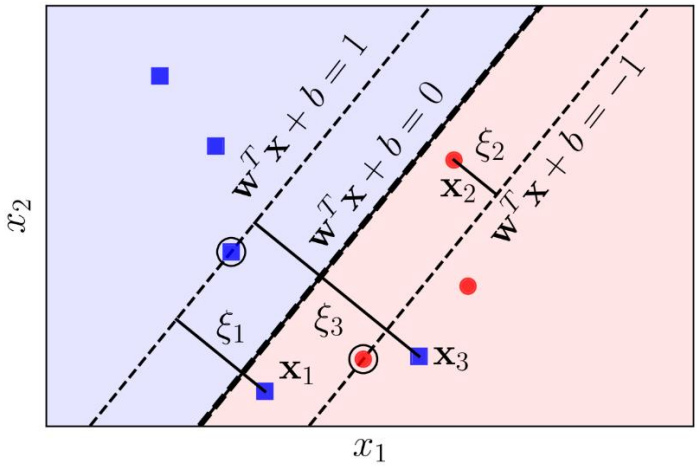
Trong cả hai trường hợp trên, margin tạo bởi đường phân chia và đường nét đứt mảnh còn được gọi là soft margin (biên mềm). Cũng theo cách gọi này, SVM thuần còn được gọi là Hard Margin SVM (SVM biên cứng).

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 margin có 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:

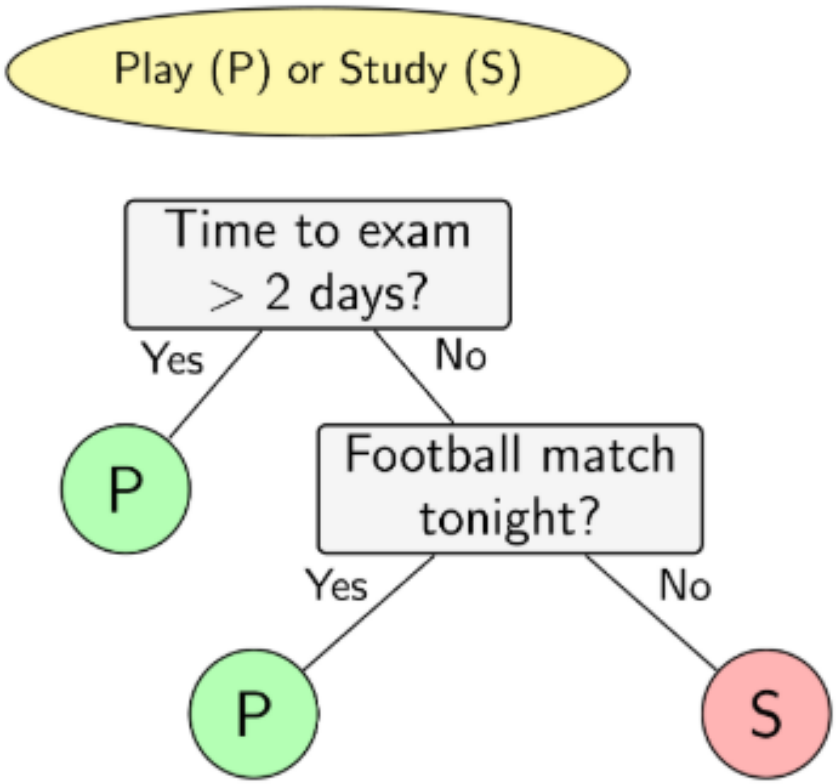


Hình 3.2 8: Giới thiệu các biến slack ξ_n

Với những điểm nằm trong vùng an toàn, $\xi_n=0$. Những điểm nằm trong vùng không an toàn nhưng vẫn đúng phía so với đường phân chia tương ứng với các $0<\xi_n<1$, ví dụ x_2. Những điểm nằm ngược phía với class của chúng so với đường boundary ứng với các $\xi_n>1$, ví dụ như x_1 và x_3.

C4.5 (Information Gain Ratio)

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 được phát triển bởi Ross Quinlan.

Là sự cải tiến của thuật toán phân lớp ID3.

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 C_i được tính bởi $|C_i, D|/|D|$.
    - Thông tin kỳ vọng để phân lớp một mẫu trong D là:
        - $$\boldsymbol{Info(D) = -\sum_{i=1}^{m} p_i * log_2(p_i)}$$
    - 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:

$$\boldsymbol{Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j)}$$

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

$$\boldsymbol{Gain(A) = Info(D) * Info_A(D)}$$

- 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.

$$\boldsymbol{GainRatio(A) = Gain(A)/SplitInfo(A)}$$

$$\boldsymbol{SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2(\frac{|D_j|}{|D|})}$$

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.

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?

CHƯƠNG 4: KẾT QUẢ

4.1. Naive Bayes

Accuracy Naive Bayes = 98.742%

"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%

"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%

"Yes"	"No"	"ClassDistribute"	[]
117	2	"Yes_test = 1"	119
4	36	"No_test = 0"	40

Bảng 4. 3: Confusion Matrix C4.5

CHƯƠNG 5: TỔNG KẾT

5.1. NHẬN XÉT

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

5.2. HƯỚNG PHÁT TRIỂN

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

Phát triển nhiều hơn trong trí tuệ nhân tạo để chuẩn bị tốt cho luận án tốt nghiệp.

Có thể phát triển phân lớp trên nhiều thuộc tính.

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
- Weka:
  [2] https://www.slideshare.net/butest/machine-learning-with-weka
  [3] http://weka-jp.info/itej/
- C4.5:
  [4] http://ait.edu.vn/Hoc_thuat/C45.html
  [5] https://en.wikipedia.org/wiki/C4.5_algorithm
- SVM:
  [6] https://www.youtube.com/watch?v=l7Tr2OyXNU8
  [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
- Một số các trang web khác liên quan:
  [10] https://machinelearningcoban.com/about/
  [11] https://www.cs.waikato.ac.nz/ml/weka/courses.html
  [12] https://www.slideshare.net/TrngHVit/artificial-intelligence-ai-l9hoc-may