BỘ CÔNG THƯƠNG

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------------------------

**ĐỒ ÁN TỐT NGHIỆP**

**“DỰ ĐOÁN XẾP HẠNG TỐT NGHIỆP DỰA TRÊN MÔ HÌNH HỌC CÓ GIÁM SÁT KẾT HỢP ASSOCIATION RULE”**

**Giáo viên hướng dẫn: TRẦN ĐỨC**

**Sinh viên thực hiện:**

**Chang Chia Sheng - 2001150360 - 06DHTH1**

**Nguyễn Anh Vũ - 2001150110 - 06DHTH1**

**Nguyễn Ngọc Thuỳ Trang - 2001150075 - 06DHTH5**

*TP. HỒ CHÍ MINH, tháng 06 năm 2019*

BỘ CÔNG THƯƠNG

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------------------------

**ĐỒ ÁN TỐT NGHIỆP**

**“DỰ ĐOÁN XẾP HẠNG TỐT NGHIỆP DỰA TRÊN MÔ HÌNH HỌC CÓ GIÁM SÁT KẾT HỢP ASSOCIATION RULE”**

**Giáo viên hướng dẫn: TRẦN ĐỨC**

**Sinh viên thực hiện:**

**Chang Chia Sheng - 2001150360 - 06DHTH1**

**Nguyễn Anh Vũ – 2001150110 - 06DHTH1**

**Nguyễn Ngọc Thuỳ Trang - 2001150075 - 06DHTH5**

*TP. HỒ CHÍ MINH, tháng 06 năm 2018*

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Đồ án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Đồ án này   
đã được cảm ơn và các thông tin trích dẫn trong Đồ án đã được chỉ rõ nguồn gốc.

**Sinh viên thực hiện Đồ án**

*(Ký và ghi rõ họ tên)*

Sinh viên thực hiện 01 Sinh viên thực hiện 02 Sinh viên thực hiện 03

Chang Chia Sheng Nguyễn Anh Vũ Nguyễn Ngọc Thuỳ Trang

LỜI CẢM ƠN

Chúng em xin gửi đến quý thầy cô trường Đại học Công nghiệp Thực phẩm thành phố Hồ Chí Minh lòng biết ơn sâu sắc, đặc biệt là thầy cô trong khoa Công nghệ Thông tin với tri thức và tâm huyết của mình đã truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường.

Để có những kiến thức và kết quả học tập như ngày hôm nay, cũng như việc áp dụng những kiến thức đã học vào triển khai thực tế .

Các thầy cô trong khoa Công Nghệ Thông Tin đã dạy dỗ chúng em trong suốt 4 năm học vừa qua. Cám ơn các thầy cô trong khoa đã gợi ý và tạo điều kiện cho nhóm em thực hiện đồ án này.

Chúng em chân thành gửi đến thầy Trần Đức là giáo viên hướng dẫn đồ án tốt nghiệp của chúng em, người đã tận tình hướng dẫn, chỉ bảo nhóm chúng em trong suốt thời gian thực hiện đồ án với tình cảm trân trọng nhất và lòng biết ơn sâu sắc nhất.

Trong quá trình thực hiện đồ án không thể tránh khỏi những thiếu sót trong quá trình thực hiện. Kính mong quý thầy cô tận tình chỉ bảo và góp ý để nhóm có thể hoàn thiện một cách tốt hơn.

Một lần nữa, nhóm xin chân thành cảm ơn và luôn mong nhận sự đóng góp quý báu của tất cả mọi người.

Trân trọng!

Sinh viên thực hiện 01 Sinh viên thực hiện 02 Sinh viên thực hiện 03

Chang Chia Sheng Nguyễn Anh Vũ Nguyễn Ngọc Thuỳ Trang

TÓM TẮT

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 xếp loại tốt nghiệp dưới đây, để có thể hiểu được những môn có khả 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 đồ án tốt nghiệp.

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.

Nội dung đồ án được chia thành các chương sau:

[CHƯƠNG 1: TỔNG QUAN](#_Toc516484691)

CHƯƠNG 2: KHẢO SÁT

CHƯƠNG 3: PHÂN TÍCH

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

MỤC LỤC

[LỜI CAM ĐOAN 3](#_Toc10469756)

[LỜI CẢM ƠN 4](#_Toc10469757)

[TÓM TẮT 5](#_Toc10469758)

[MỤC LỤC 7](#_Toc10469759)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 13](#_Toc10469760)

[DANH MỤC CÁC BẢNG BIỂU 14](#_Toc10469761)

[DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ, HÌNH ẢNH 14](#_Toc10469762)

[CHƯƠNG 1: TỔNG QUÁT 16](#_Toc10469763)

[1.1. Thông tin đề tài 16](#_Toc10469764)

[1.1.1. Tên đề tài 16](#_Toc10469765)

[1.1.2. Giới thiệu chung về đề tài 16](#_Toc10469766)

[1.2. Mục đích và phạm vi đề tài 16](#_Toc10469768)

[1.2.1. Mục đích của đề tài 16](#_Toc10469769)

[1.2.2. Yêu cầu của đề tài 16](#_Toc10469771)

[1.2.3. Phạm vi của đề tài 16](#_Toc10469773)

[1.2.4. Môi trường triển khai của đề tài 16](#_Toc10469774)

[CHƯƠNG 2: KHẢO SÁT 17](#_Toc10469775)

[2.1. Dữ liệu thực hiện 17](#_Toc10469776)

[2.2. Quy tắc xét tốt nghiệp bậc đại học 18](#_Toc10469777)

[CHƯƠNG 3: PHÂN TÍCH 19](#_Toc10469778)

[3.1. Máy học và Dataminding 19](#_Toc10469779)

[3.1.1. Máy học 19](#_Toc10469780)

[3.1.2. Dataminding 19](#_Toc10469781)

[*3.2.* Thuật toán phân lớp 21](#_Toc10469782)

[3.2.1. Khái quát về học có giám sát và không giám sát 21](#_Toc10469783)

[3.2.1.1. Học có giám sát (Supervised Learning) 21](#_Toc10469784)

[3.2.1.2. Học không giám sát (Unsupervised Learning) 21](#_Toc10469785)

[3.2.2. KNN (K-nearest neighbor) 21](#_Toc10469787)

[3.2.2.1. Khái niệm 21](#_Toc10469788)

[3.2.2.2. Khái quát 22](#_Toc10469789)

[3.2.2.3. KNN cho Regression 23](#_Toc10469790)

[3.2.2.4. Chuẩn hoá dữ liệu 23](#_Toc10469791)

[3.2.2.5. Sử dụng các phép đo khoảng cách khác nhau 24](#_Toc10469792)

[3.2.2.6. Phân loại 24](#_Toc10469793)

[3.2.3. SVM (Suport Vector Machine) 24](#_Toc10469794)

[3.2.3.1. Khái niệm 25](#_Toc10469795)

[3.2.3.2. Ưu nhược điểm 27](#_Toc10469796)

[3.2.3.3. Xây dựng bài toán tối ưu cho SVM 28](#_Toc10469797)

[3.2.3.4. Soft Margin SVM 31](#_Toc10469799)

[3.2.3.5. Phân tích toán học 32](#_Toc10469801)

[3.2.4. C 4.5 (Information Gain Ratio) 33](#_Toc10469803)

[3.2.4.1. Khái niệm 33](#_Toc10469804)

[3.2.4.3. Ưu điểm 34](#_Toc10469806)

[3.2.4.4. Nhược điểm 35](#_Toc10469807)

[3.3. Luật kết hợp 35](#_Toc10469808)

[CHƯƠNG 4: KẾT LUẬN 36](#_Toc10469809)

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Thuật ngữ** | **Diễn giải tiếng Anh** | **Diễn giải 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 CÁC BẢNG BIỂU

[Bảng 2. 1: Danh sách các môn để dự đoán tốt nghiệp 14](#_Toc10469581)

DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ, HÌNH ẢNH

[Hình 3.1 Các lĩnh vực trong DataMining 19](#_Toc10469335)

[Hình 3.2 KNN cho bài toán Regression 23](#_Toc10469337)

[Hình 3.3 Phân lớp tuyến tính 25](#_Toc10469338)

[Hình 3.4 Các mặt phân cách hai classes linearly separable 26](#_Toc10469340)

[Hình 3.5 Phân lớp phi tuyến 26](#_Toc10469342)

[Hình 3.6 Margin của hai classes là bằng nhau và lớn nhất có thể 27](#_Toc10469343)

[Hình 3.7 Phân tích bài toán SVM 29](#_Toc10469344)

[Hình 3.8 Các điểm gần mặt phân cách nhất của hai classes được khoanh tròn. 31](#_Toc10469358)

[Hình 3.9 Soft margin SVM. Khi a) có nhiễu hoặc b) dữ liệu gần linearly separable, SVM thuần sẽ không hoạt động hiệu quả. 31](#_Toc10469359)

[Hình 3.10 Giới thiệu các biến slack ξn 32](#_Toc10469362)

[Hình 3.11 Ví dụ mô hình cây quyết định. 33](#_Toc10469364)

[3.2.4.2. Diễn giải 33](#_Toc10469367)

1. TỔNG QUÁT
   1. Thông tin đề tài
      1. Tên đề tài

DỰ ĐOÁN XẾP HẠNG TỐT NGHIỆP DỰA TRÊN MÔ HÌNH HỌC CÓ GIÁM SÁT KẾT HỢP ASSOCIATION RULE.

* + 1. 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, xếp loại gì và những môn nào có thể ảnh hưởng đến xếp loại.

* 1. Mục đích và phạm vi đề tài
     1. Mục đích của đề tài

Dự đoán kết quả xếp loại tốt nghiệp cho sinh viên khóa 06DHTH dựa trên điểm trung bình từng môn học của 3 năm học và tìm ra những môn có thể ảnh hưởng đến kết quả này.

Chạy thuật toán phân lớp trên MATLAB, ghi nhận kết quả.

Luật kết hợp hỗ trợ dự đoán hạn tốt nghiệp dựa vào điểm các môn quan trọng của ngành

* + 1. 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. Phạm vi của đề tài

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

* + 1. Môi trường triển khai của đề tài

MATLAB và WEKA.

1. KHẢO SÁT
   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

* 1. Quy 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 2.1 thì thể hiện ở cột 1 đến 10),
* Các môn có tích lũy tín chỉ thì phải >= 4 điểm trờ lên (trong bảng 2.1 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”}.

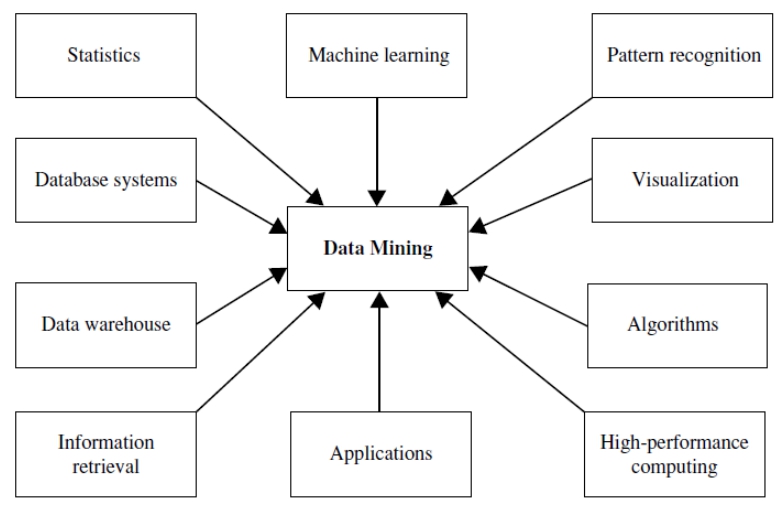
1. PHÂN TÍCH
   1. Máy học và Dataminding
      1. Máy học

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)".

* + 1. 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:

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 đoán dựa vào 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.
  1. Thuật toán phân lớp
     1. Khái quát về học có giám sát và không giám sát
        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,…

* + - 1. 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.

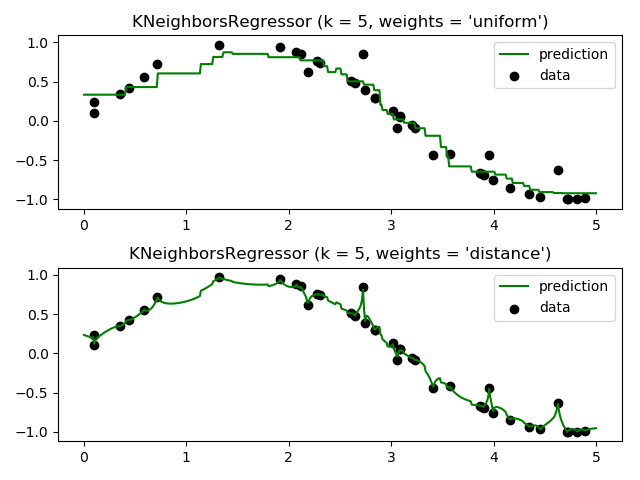
Nhãn lớp của tập huấn luyện không biết trước. Trong thuật toán này, chúng ta không biết trước được đầu ra hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán không có giám sát sẽ dựa vào cấu trúc của dữ liệu đề thực hiện một công việc nào đó để thuận tiện trong việc lưu trữ và tính toán.

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,…

* + 1. KNN (K-nearest neighbor)
       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).
  + - 1. 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.
* Nhân 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.
    - 1. 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.



1. KNN cho bài toán Regression
   * + 1. Chuẩn hoá dữ liệu

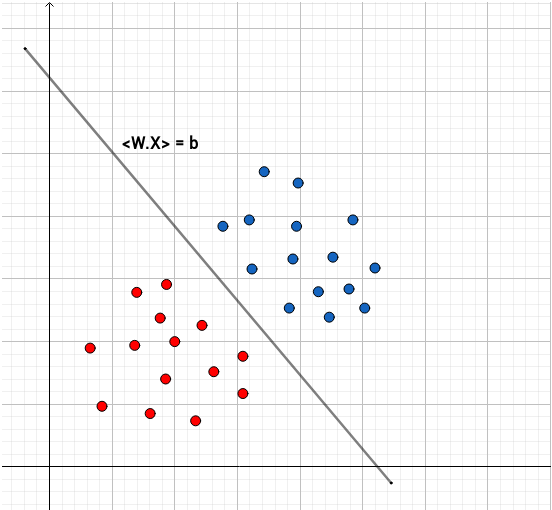
Khi có một thuộc tính trong dữ liệu (hay phần tử trong vector) lớn hơn các thuộc tính khác rất nhiều (ví dụ thay vì đo bằng cm thì một kết quả lại tính bằng mm), khoảng cách giữa các điểm sẽ phụ thuộc vào thuộc tính này rất nhiều. Để có được kết quả chính xác hơn, một kỹ thuật thường được dùng là Data 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 kỹ thuật chuẩn hóa được áp dụng với không chỉ KNN mà còn với hầu hết các thuật toán khác.

* + - 1. Sử dụng các phép đo khoảng cách khác nhau

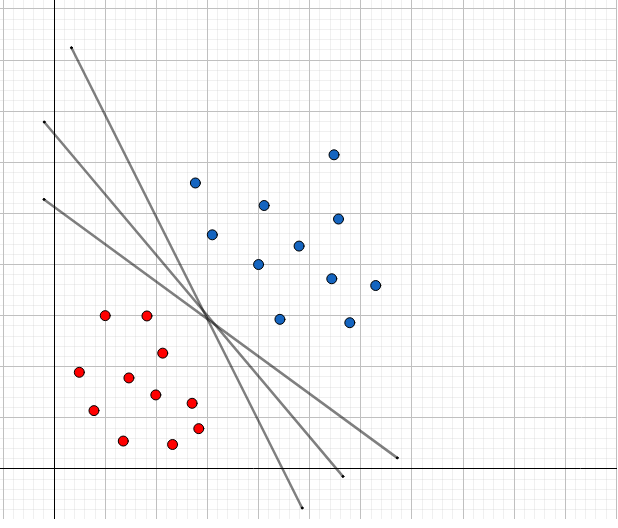
Ngoài norm 1 và norm 2 chúng tôi giới thiệu trong bài này, còn rất nhiều các khoảng cách khác nhau có thể được dùng. Một ví dụ đơn giản là đếm số lượng thuộc tính khác nhau giữa hai điểm dữ liệu. Số này càng nhỏ thì hai điểm càng gần nhau. Đây chính là giả chuẩn 0 mà chúng tôi đã giới thiệu trong Tab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

* + - 1. 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.
  + 1. SVM (Suport Vector Machine)
       1. Khái niệm

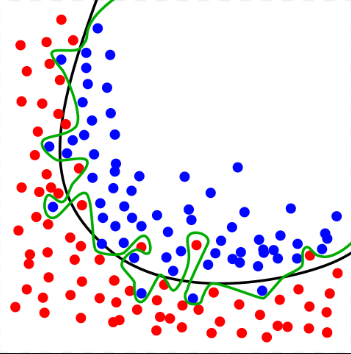
SVM là một thuật toán học có giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng em vẽ đồ thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng dữ liệu có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng em 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.



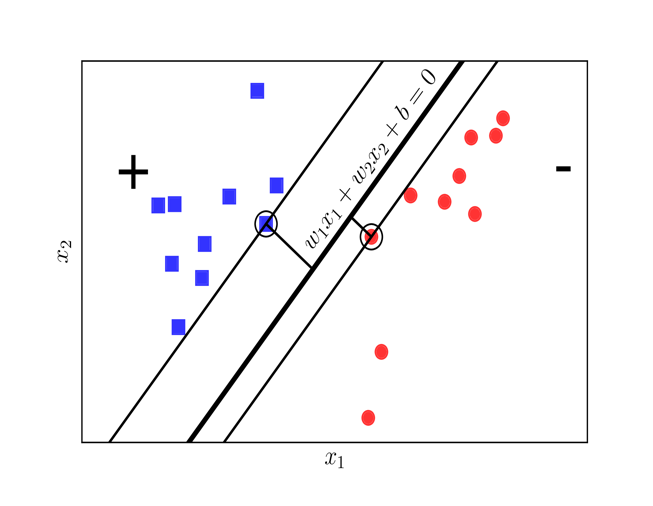
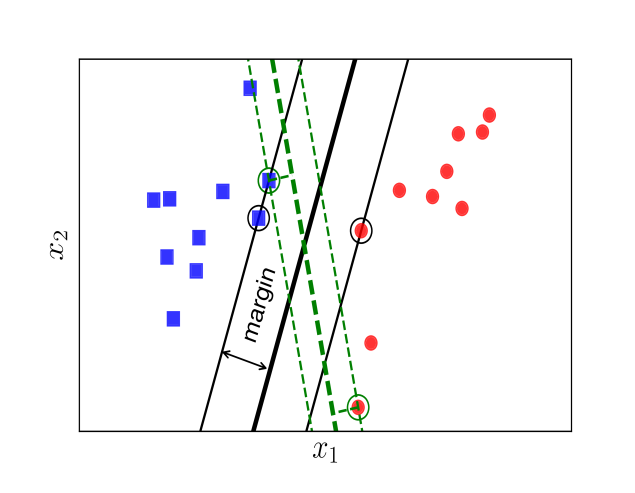
1. Phân lớp tuyến tính



1. Các mặt phân cách hai classes linearly separable



1. Phân lớp phi tuyến

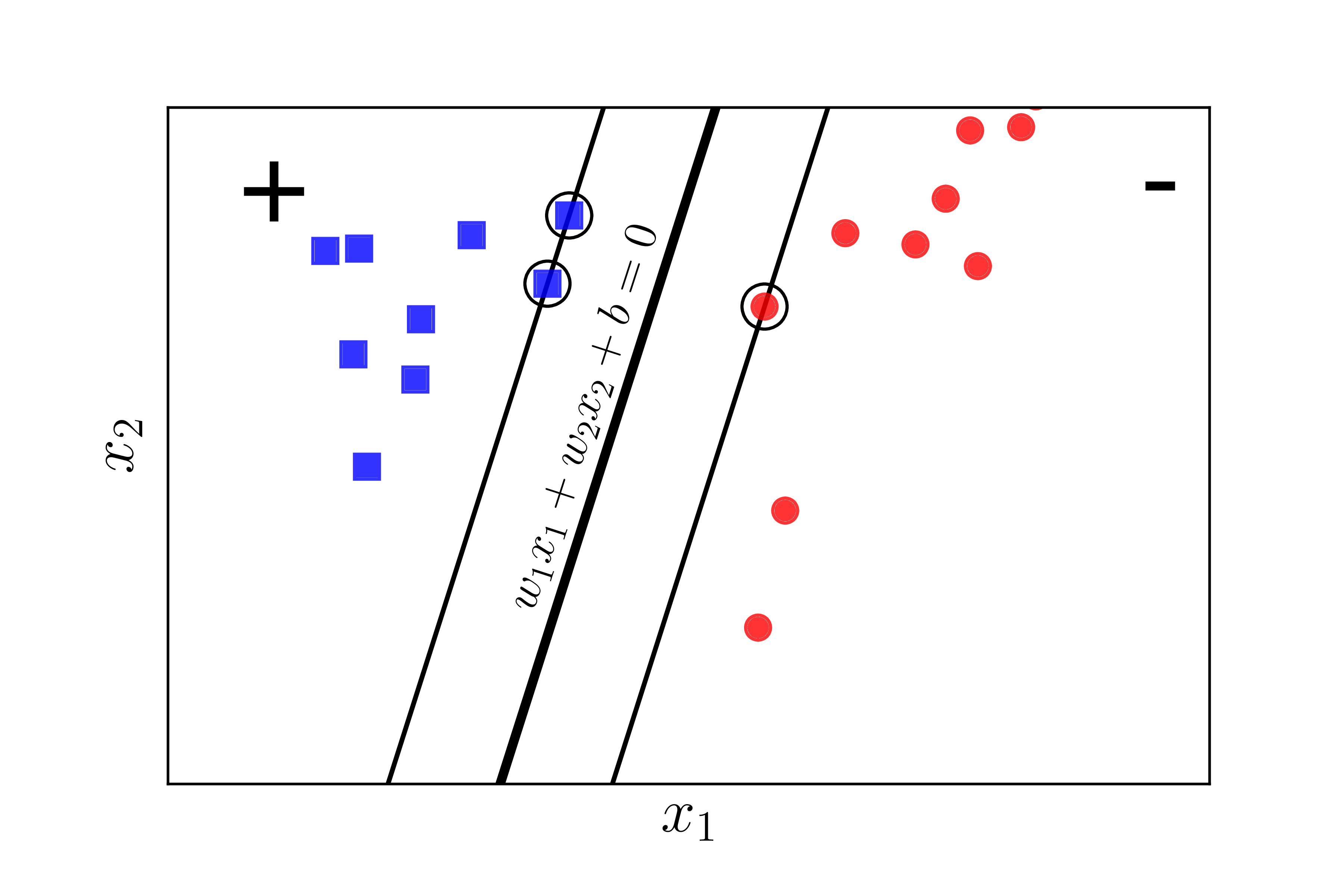


1. Margin của hai classes là bằng nhau và lớn nhất có thể
   * + 1. Ưu nhược điểm

* Ưu điểm: 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:
  + **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.
    - 1. 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.

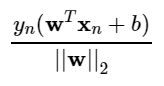
Để giúp các bạn dễ hình dung, chúng ta cùng xét trường hợp trong không gian hai chiều dưới đây. Không gian hai chiều để các bạn dễ hình dung, các phép toán hoàn toàn có thể được tổng quát lên không gian nhiều chiều.



1. Phân tích bài toán SVM

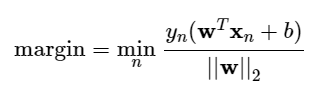
Giả sử rằng các điểm vuông xanh thuộc class 1, các điểm tròn đỏ thuộc class -1 và mặt wTx + b = w1x1 + w2x2 + b = 0 là mặt phân chia giữa hai classes (Hình 3.2.5). Hơn nữa, class 1 nằm về phía dương, class -1 nằm về phía âm của mặt phân chia. Nếu ngược lại, ta chỉ cần đổi dấu của w và b. Chú ý rằng chúng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

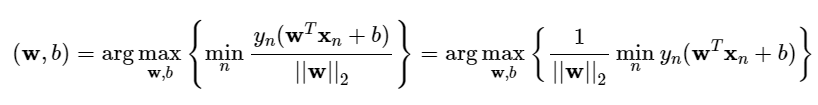


Điều này có thể dễ nhận thấy vì theo giả sử ở trên, yn luôn cùng dấu với phía của xn. Từ đó suy ra yn cùng dấu với wTx + b, và tử số luôn là 1 số không âm.

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):

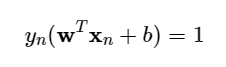


Bài toán tối ưu trong SVM chính là bài toán tìm w và b sao cho margin này đạt giá trị lớn nhất:

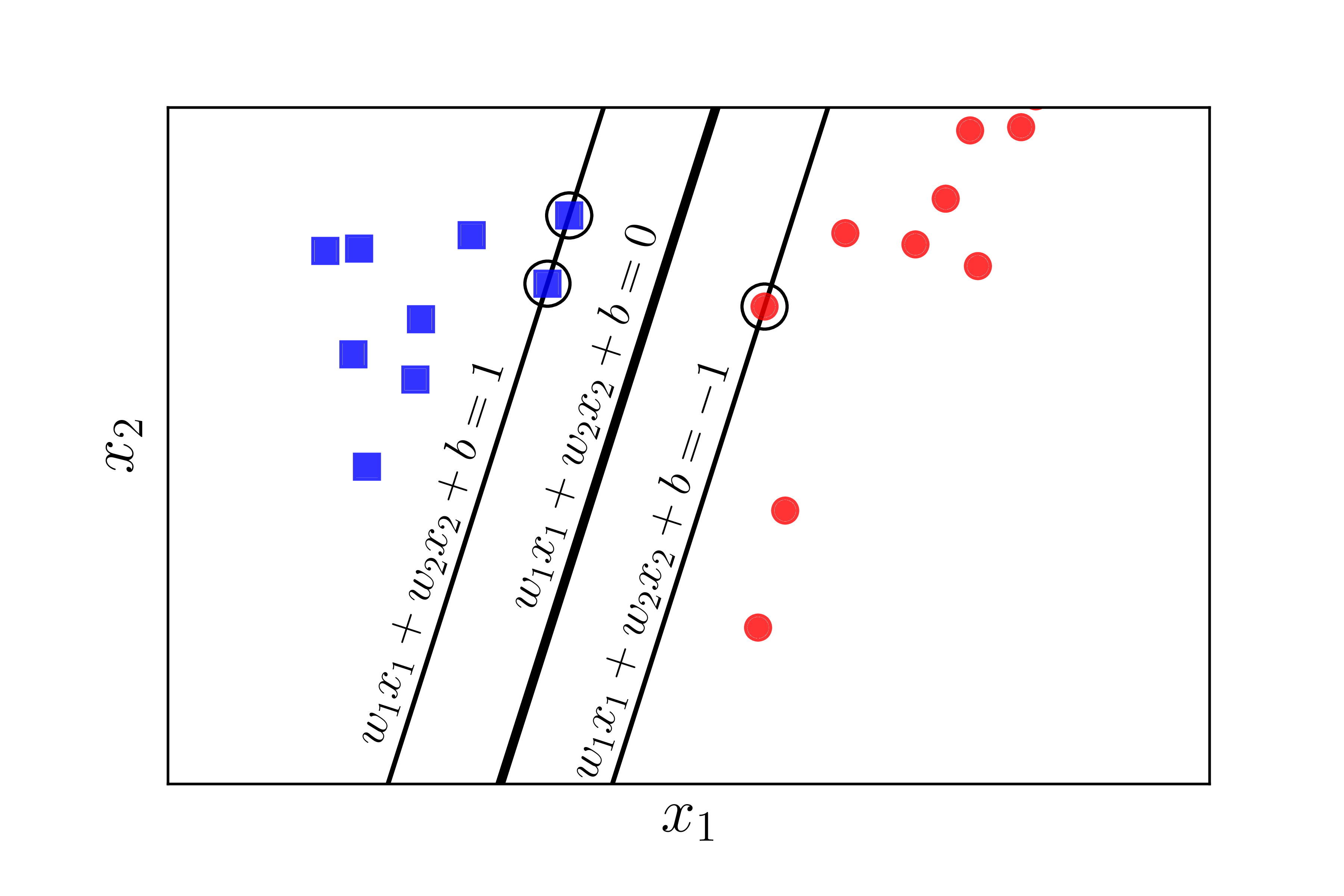


Việc giải trực tiếp bài toán này sẽ rất phức tạp, nhưng các bạn sẽ thấy có cách để đưa nó về bài toán đơn giản hơn.

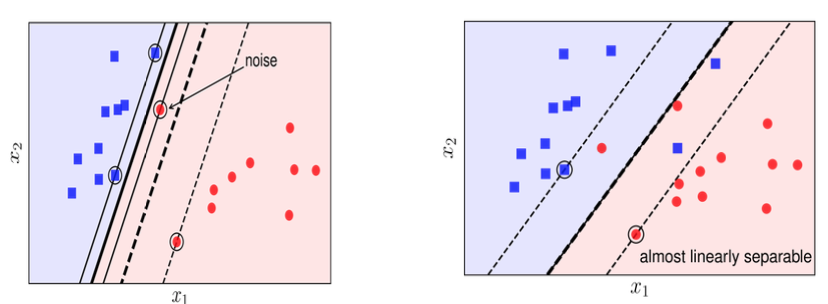
Nhận xét quan trọng nhất là nếu ta thay vector hệ số w bởi kw và b bởi kb trong đó k là một hằng số dương thì mặt phân chia không thay đổi, tức khoảng cách từ từng điểm đến mặt phân chia không đổi, tức margin không đổi. Dựa trên tính chất này, ta có thể giả sử:



**Với những điểm nằm gần mặt phân chia nhất**như Hình 3.8 dưới đây:



1. Các điểm gần mặt phân cách nhất của hai classes được khoanh tròn.
   * + 1. Soft Margin SVM



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

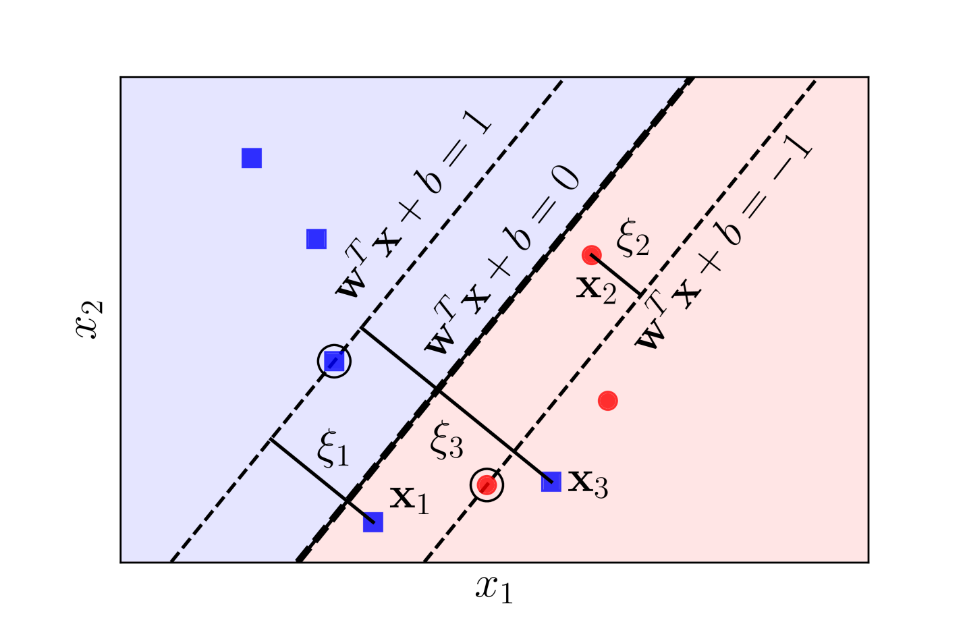
Trong cả hai trường hợp trên, margin tạo bởi đường phân chia và đường nét đứt mảnh còn được gọi là soft margin (biên mềm). Cũng theo cách gọi này, SVM thuần còn được gọi là Hard Margin SVM (SVM biên cứng).

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

* + - 1. 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.

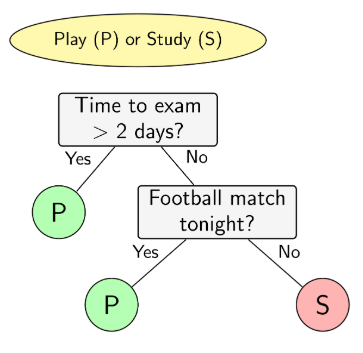
Giống như với Hard Margin SVM, việc tối đa margincó thể đưa về việc tối thiểu (||w||2)2. Để xác định sự hy sinh, chúng ta cùng theo dõi Hình 3.10 dưới đây:



1. 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.

* + 1. C 4.5 (Information Gain Ratio)
       1. Khái niệm



1. 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.

* + - 1. 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.

* + - 1. 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.
  + - 1. 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.
  1. Luật kết hợp
     1. Giới thiệu

Cho tập hạng mục (itemset) I = {i1, i2,…, in}; ii ∈ I được gọi là một hạng mục (item).

Cho CSDL giao dịch (transaction) T = {t1, t2,…, tm}; trong đó mỗi ti là một giao dịch và là một tập con của I.

Số lượng các giao dịch T ký hiệu là |T| hay card(T).

Tập k hạng mục (k - itemset) = {i1, i2,…, ik}; gồm k item trong I.

* Độ phổ biến:
* Supp(X) = |X|/|D|
* Supp(X⇒Y)=|{T⊆D:X∪Y⊆T}|/|D|
* Độ tin cậy
* 𝐶𝑜𝑛𝑓 𝑋⇒Y =
* Ví dụ: |D|=10;|X∪Y|=5;

|X|=7;|Y|=6 Supp(X⇒Y) = 5/10 = 0,5

Conf(X⇒Y) = 5/7 ≈ 0,71

Luật kết hợp X→Y được coi là một mẫu có giá trị nếu xảy ra đồng thời supp(X→Y) ≥ minsup và conf(X→Y) ≥ minconf( minsup và minconf là hai giá trị ngưỡng cho trước).

* Tập phổ biến (frequent itemsets - FIs): một tập hạng mục X có độ hỗ trợ vượt qua ngưỡng minsup được gọi là tập phổ biến
* Tập phổ biến tối đại (Max Pattern/Max Frequent Itemsets) X nếu thỏa:

1. Supp(X) ≥ minsupp và

2. ∄ 𝑋′ ⊃ 𝑋 mà X’ cũng phổ biến.

* Tập phổ biến đóng (Closed Pattern/Frequent Closed Itemsets) X nếu thỏa:

1. Supp(X) ≥ minsupp và

2. ∄ 𝑋′ ⊃ 𝑋 mà supp(X’) = supp(X).

* Qui trình khai thác luật kết hợp:
* Bước 1: Tìm các tập phổ biến thỏa ngưỡng minsupp.
* Bước 2: Xây dựng luật từ các tập phổ biến:
* Đối với mỗi tập phổ biến S, tạo mọi tập con khác rỗng của S.
* Đối với mọi tập con khác rỗng A của S. Luật A  (S –A) là luật kết hợp cần tìm nếu conf(A  (S –A)) ≥ minconf.
  + 1. Bài toán khai thác luật kết hợp

1. Khai thác tập phổ biến (FIs – Frequent Itemsets):

2. Khai thác luật từ các tập phổ biến (ARs – Association Rules)

* Bài toán khai thác tập phổ biến (frequent itemset) là bài toán rất quan trọng trong lĩnh vực data mining.

Bài toán khai thác tập phổ biến là bài toán tìm tất cả tập các hạng mục (itemset) S có độ phổ biến (support) thỏa mãn độ phổ biến tối thiểu minsupp: supp(S) ≥ minsupp

* Một số thuật toán khai thác tập phổ biến:
* Thuật toánApriori (Agrawal et al).
* Thuật toán IT-tree (Zaki – et al).
* Thuật toán FP-Tree (Han et al).
* Các thuật toán khác như LCM, DCI, PrePost, …
  + 1. Thuật toán Apriori
* Thuật toán Apriori là một thuật toán điển hình áp dụng trong khai phá luật kết hợp.
* Thuật toán dựa trên nguyên lý Apriori “tập con bất kỳ của một tập phổ biến cũng là một tập phổ biến và mọi tập cha của tập không phổ biến là không phổ biến”.
* Mục đích của thuật toán Apriori là tìm ra được tất cả các tập phổ biến có thể có trong cơ sở dữ liệu giao dịch D.
* Thuật toán hoạt động theo nguyên tắc quy hoạch động, nghĩa là từ các tập Fi = { ci | ci là tập phổ biến, |ci| = i} gồm mọi tập phổ biến có độ dài i (1 ≤ i ≤ k), đi tìm tập Fk+1 gồm mọi tập phổ biến có độ dài k+1. Các hạng mục i1, i2,…, in trong thuật toán được sắp xếp theo một thứ tự cố định.
* Ý nghĩa nguyên lý Apriori:
* Mọi tập con của tập phổ biến đều phổ biến, nghĩa là ∀ 𝑋 ⊆ 𝑌, nếu supp(Y) ≥ minsupp thì supp(X) ≥ minsupp.
* Mọi tập cha của tập không phổ biến đều không phổ biến, nghĩa là ∀ 𝑋 ⊆ 𝑌, nếu supp(X) ≤ minsupp thì supp(Y) ≤ minsupp.
* Thuật toán Apriori Input: CSDL giao dịch T và ngưỡng phổ biến minsupp. Output: FIs chưa tất cả các tập phổ biến của T.
* Ý tưởng thuật toán Apriori

1. Tìm tất cả các tập phổ biến 1 hạng mục (L1)
2. Tạo các tập ứng viên kích thước k hạng mục Ck từ các tập phổ biến Lk-1 hạng mục (nguyên lý Apriori). Ví dụ, tạo ứng viên C2 từ tập phổ biến L1.
3. Kiểm tra độ phổ biến của các ứng viên trên CSDL và loại các ứng viên không phổ biến ta được Li(i=1,2,..., k).
4. Dừng khi không tạo được tập phổ biến hay tập ứng viên

* Thuật toán Apriori
* Mã giả: Gọi Ck là tập các ứng viên có kích thước k Lk các tập phổ biến có kích thước k L1= { i ∈ I: supp(i) ≥ minsupp} for (k = 1; Lk !=∅; k++) do Ck+1 =apriori\_gen(Lk)
* //tập ứng viên được tạo từ Lk-1 for each t ∈ T do for each c ∈ Ck do if c ⊆ t then c.count++ Lk = {c ∈ Ck| c.count ≥ minSup} FIs = ∪𝑘Lk;
* Tạo tập ứng viên Ck+1 hạng mục từ tập Lk hạng mục: Gồm 2 bước:
  + - Kết và loại bỏ
    - Giả sử rằng Lk được xếp theo thứ tự từ điển Hàm apriori\_gen:
* Input: Lk tập phổ biến kích thước k ◦ Output: Ck+1 tập ứng viên kích thước k+1
* ◦ Mã giả:
* Procedure apriori\_gen(Lk) { For mỗi l1 ∈ Lk For mỗi l2 ∈ Lk If (l1[1]=l2[1] ⋀ … ⋀ l1[k-1]=l2[k-1] ⋀ l1[k]≠l2[k] ) then C = l1 ⋈ l2 If (có tập con của C ∉ không thuộc Lk ) then xóa C else them C vào Ck+1 Return Ck+1 }
* Ví dụ: Cho CSDL sau và minsupp = 50%.
* Mã giao dịch (TID)
* Nội dung giao dịch 1 A, C, D 2 B, C, E 3 A, B, C, E 4 B, E
* Itemset {A} {B} {C} {D} {E}
* Itemset Supp {A} 2 {B} 2 {C} 3 {E} 3
* C1 L1
* Itemset Supp {A} 2 {B} 2 {C} 3 {E} 3
* C2 L2 Itemset {A, B} {A, C} {A, E} {B, C} {B, E} {C, E}
* Itemset Supp {A, C} 2 {B, C} 2 {B, E} 3 {C, E} 2
* Chú ý: {A, C, B} không có trong C3 vì {A, B} không có trong L2.
* C3 L3 Itemset Supp {B, C, E} 2
* Itemset Supp {A, C} 2 {B, C} 2 {B, E} 3 {C, E} 2
* Itemset {B, C, E}
* Các tập phổ biến tìm được là: FIs = {{A}, {B}, {C}, {E}, {A, C}, {B, C}, {B, E}, {C, E}, {B, C, E}}.

35

* Thuật toán Apriori là thuật toán dễ hiểu, dễ cài đặt. Tuy nhiên có các nhược điểm sau:
* Phải duyệt CSDL nhiều lần.
* Số lượng tập ứng viên rất lớn (=2|T|- 1).
* Thực hiện việc tính độ phổ biến nhiều, đơn điệu.
* Cải tiến Apriori: Ý tưởng chung.
* Giảm số lần duyệt CSDL.
* Giảm số lượng (không sinh) tập ứng viên.
* Tìm qui trình tính độ phổ biến hiệu quả hơn.

1. KẾT LUẬN