**Câu trắc nghiệm: nội dung toàn bộ bài giảng (tới học mạng nơ ron, không kế học suốt đời và sau đó), đánh dấu sai bị trừ điểm**

**Câu viết.** (***Viết các nội dung cơ bản nhất càng đầy đủ càng tốt về hai trong các câu sau đây)***

1. Trình bày quy trình mô hình chuẩn công nghiệp khai phá dữ liệu CRISP-DM. Tại sao bước hiểu kinh doanh và hiểu dữ liệu là quan trọng. Nêu điểm giống và khác nhau (càng nhiều càng tốt) giữa bài toán khai phá dữ liệu và bài toán CSDL.

Quy trình crisp-dm gồm 6 bước chính:

* Hiểu kinh doanh
* Hiểu dữ liệu
* Chuẩn bị dữ liệu
* Khai phá dữ liệu
* Đánh giá mô hình
* Triển khai

Bước hiểu kinh doanh và hiểu dữ liệu là quan trọng trong quá trình khai phá dữ liệu vì nó giúp định hướng cho quá trình khai phá dữ liệu và đảm bảo rằng kết quả được sẽ hỗ trợ trong việc ra quyết định trong việc kinh doanh. Nếu không hiểu rõ bước này, có thể dẫn đến việc tạo ra các mô hình khai phá dữ liệu không hữu ích hoặc không thể áp dụng trong thực tế

Điểm giống giữa 2 bài toán

* liên quan đến dữ liệu
* mục tiêu tìm kiếm thông tin từ dữ liệu

Điểm khác

* KPDL tập trung vào việc tìm kiếm các mẫu hoặc quy luật trong dữ liệu, trong đó bài toán cơ sở dữ liệu tập trung vào việc lưu trữ hoặc truy xuất các dữ liệu
* KPDL sử dụng các phương pháp KPDL tìm kiếm thông tin, trong khi đó bài toán CSDL sử dụng các phương pháp quản lý dữ liêu để lưu trữ và truy xuất thông tin.
* KPDL thường được sử dụng để tìm ra các mẫu hoặc quy luật ẩn trong dữ liệu
* CSDL thường được sử dụng để lưu trữ và truy xuất thông tin có sẵn

1. Khái niệm luật kết hợp và các độ đo liên quan. Khái niệm tập phố biến và tiếp cận chung hai bước tìm luật kết hợp trong một CSDL giao dịch.

* Luật kết hợp
  + gọi a->b là một luật kết hợp nếu a là tập con của i, b là tập con của i và a và b không giao nhau
* Các độ đo của luật kết hợp
  + Độ hỗ trợ (support): là tần suất xuất hiện của một giao dịch trong một tập hàng. P(giao của a và b) >= 0 <= 1 (s(A->B))
  + Độ đáng tin cậy(confidence): là xác suất để xuất hiện của một mục trong cặp kết hợp xuất hiện khi cặp kia xuất hiện c(A->B)
* Luật mạnh: Luật A->B được gọi là đảm bảo hỗ trợ s nếu s(A->B)>=s, đảm bảo độ tin cậy c nếu c(A->B)>=c
* Khái niệm tập phổ biến: tập phổ biến thường xuất hiện thường xuyên trong giao dịch.
* Tiếp cận chung hai bước tìm tập kết hợp: tách tập dữ liệu lớn thành các tập dữ liệu nhỏ hơn. Nó gọi là tiếp cận hai bước vì nó sử dụng 2 giai đoạn để tìm các tập kết hợp: trích xuất tập phổ biến và kết hợp các tập phổ biến để tìm luật kết hợp

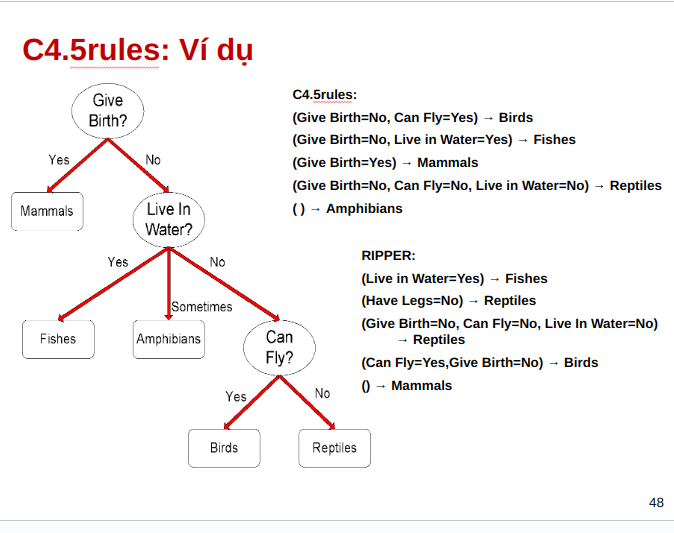
1. Phát biểu bàitoán tìm tập phổ biến và luật mạnh. Trình bày các nội dung cơ bản của thuật toán Apriori tìm tập mục phổ biến (input-output-nội dung với các bước chính). Giải thích sơ bộ về tính đúng đắn của thuật toán.

* Bài toán tìm tập phổ biến và luật mạnh là tìm các tập hợp trong tập xuất hiện thường xuyên và xác định các quy tắc kết hợp giữa các mục đó.
* Thuật toán apriori:
  + input: tập F(k) gồm mọi tập phổ biến có độ dài k
  + output: tập C(k+1) gồm mọi tập c(k+1) độ dài k+1 mọi tập ứng viên có độ dài k+1 từ tập F(k). Nguyên lý tỉa Apriori: Mọi c(k+1) phải được tạo ra từ F(k)
  + các bước:
    - bước ghép: Sinh tập c(k+1) bằng cách ghép hai mục P(k) và Q(k) thuộc F(k)
      * P(k)={i1, i2, …, ik-1, ik}
      * Q(k)={i1, i2, …, ik-1, ik’}
      * c(k+1) = {i1, i2, .., ik-1, ik, ik’}; i1 < i2 < … < ik-1 < ik < ik'
    - bước tỉa: nếu với X là tập con của c(k+1), số lượng của X là k, X thuộc F(k) thì bổ sung c(k+1) vào Ck+1, ngược lại bỏ qua
  + tính đúng đắn
    - Thuật toán Apriori đúng vì nó đảm bảo tìm ra tất cả các tập mục phổ biến đạt được ngưỡng hỗ trợ. Tuy nhiên, thuật toán có thể trả về một số tập mục không phổ biến, vì chúng có thể đạt được ngưỡng hỗ trợ vì sự xuất hiện ngẫu nhiên trong các giao dịch.

1. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp Bayes (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm).

* Bài toán phân lớp là bài toán xác định lớp dựa trên đặc trưng của nó. Xác định lớp có xác suất cao nhất, xấp xỉ tốt nhaát cho mục đích đó
* Phân lớp bayes:
  + Mô hình phân lớp: tập công thức tính xác suất
  + cơ sở khoa học: xác suất có điều kiện, định lý bayes. P(x) bằng nhau cho tất cả các lớp. Tìm c sao cho P(c|x) lớn nhất ⇔ tìm c sao cho P(x|c).P(c) lớn nhất. P(c) tần suất xuất hiện các dữ liệu thuộc lớp c. Vấn đề: cách thức tính P(x|c).
  + giải pháp chính: tính xác suất của mỗi mục dựa trên đặc trưng của mục đó. Xác định xác suất cao nhất của lớp cho mục đó.
* Giải pháp chính
  + n + 1 thuộc tính là các biến ngẫu nhiên
  + Cho một bản ghi với các giá trị thuộc tính (A\_1, A\_2, …, A\_n) là các giá trị quan sát được các biến ngẫu nhiên. Cần dự báo nhãn c. Và tìm lớp c để cực đại xác suất P(C|A1, A2, …, An)
  + Phân lớp naive bayess
    - Cho tập ví dụ Dexample = Dlearn + Dtest
    - tập lớp C.
    - Tính xác suất tiên nghiệm
* ưu: đơn giản, dễ hiểu. Hoạt động tốt trong nhiều trường hợp nếu các giả định độc lập được thỏa mãn. Là một phương pháp để tham chiếu với các thuật toán khác
* nhược: điều kiện độc lập thường không thỏa mãn trong thực tế. Khó khăn khi đối phó các đặc trưng liên quan đến nhau.

1. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp cây quyết định (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm, ví dụ đơn giản).

* Bài toán phân lớp là bài toán xác định lớp dựa trên đặc trưng của nó. Xác định lớp có xác suất cao nhất, xấp xỉ tốt nhaát cho mục đích đó
* thuật toán phân lớp cây quyết định:
  + Mô hình phân lớp: cây quyết định
  + Cơ sở lý thuyết: phương châm chia để trị. Đệ quy: mỗi nút tương ứng với một tập các ví dụ học. Một số thuật toán phổ biến: Hunt, họ IID3+C4.5+C.5x
  + Giải pháp chính:
    - Sử dụng cây quyết định: kiểm tra từ gốc theo điều kiện. Lá: giá trị lớp nhãn; chính xác một cung vào + không có cung ra.
    - Nút trong: tên thuộc tính; có chính xác một cung vào và một số cung ra, gắn với điều kiện kiểm tra giá trị thuộc tính của nút
    - gốc: kông có cung vào + không một số cung ra. (i) suy biến thành lá, nhãn, ngược lại: tên thuộc tính ở gốc, điều kiện ở cung ra.
  + Ưu: Dễ hiểu và giải thích, có thể xử lý các thuộc tính rời rạc và liên tục. có thể xử lý các mẫu dữ liệu bị thiếu
  + Nhược: dễ bị quá khớp (overfitting). không thể xử lý mối quan hệ phức tạp giữa các thuộc tính
  + Ví dụ:
  + 

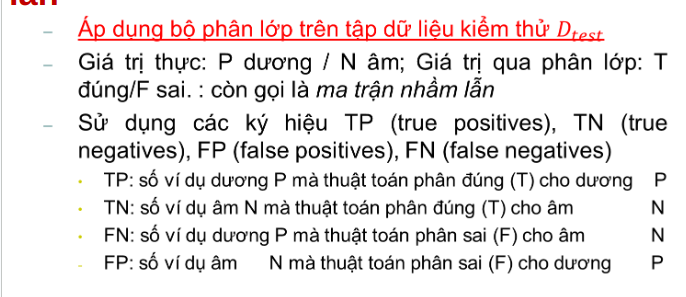
1. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp k-người láng giềng gần nhất (căn cứ, giải pháp chính, ưu điểm- nhược điểm, ví dụ đơn giản).

* Bài toán phân lớp là bài toán xác định lớp dựa trên đặc trưng của nó. Xác định lớp có xác suất cao nhất, xấp xỉ tốt nhất cho mục đích đó
* Căn cứ của bài toán KNN
  + phân lớp dựa trên khoảng cách hoặc dự đoán nhóm của một điểm cá nhân
* giải pháp chính
  + Cho một
    - tập D các đối tượng biểu diễn bản ghi đặc trưng
    - một độ đo khoảng cách (eculidean) hoặc tương tự
    - một sô k > 0 (láng giềng gần nhất)
  + Phân lớp đối tượng Xc mới được biểu diễn
    - tính khoảng cách (độ tương tự) từ X đến tất cả dữ liệu thuộc D
    - tìm k dữ liệu thuộc D gần X nhất
    - Dùng nhãn lớp của k láng giềng gần nhất để xác định nhãn lớp của X: nhãn nhiều nhất trong k láng giềng gần nhất
* Ưu điểm
  + đơn giản, dễ hiểu và độ chính xác cao
  + không cần yêu cầu phân phối đồng nhất của các lớp
  + có thể áp dụng cho tập dữ liệu lớn
* Nhược điểm
  + khi tập dữ liệu cao lên, nhiều chiều lên thì knn trở nên kém hiệu quả hơn ảnh hưởng đến hiệu suất tổng thể của mô hình
  + cần xác định trước giá trị k trước khi áp dụng thuật toán
  + yêu cầu tính toán khoảng cách giữa các mẫu dữ liệu, điều này có thể tốn nhiều thời gian đối với tập dữ liệu lớn
* Ví dụ
  + phân loại hoa iris. Mục tiêu là phân loại loài hoa iris dựa trên chiều dài và chiều rộng của cành hoa và lá hoa. tập dữ liệu huấn luyện có sẵn ở đây bao gồm các mẫu hoa và nhãn của chúng. Khi có một mẫu hoa mới, thuật toán KNN sẽ tìm k mẫu gần nhất trong tập dữ liệu huấn luyện và xác định lớp phổ biến nhất của chúng để gán cho mẫu hoa đó. nếu k = 3 và mẫu hoa mới có chiều dài cánh hoa là 5.8 và chiều rộng của cánh hoa là 2.7, thuật toán sẽ tìm 3 mẫu gần nhất trong tập dữ liệu huấn luyện và xác định lớp phổ biến nhất của chúng để dự đoán nhãn cho mẫu hoa mới đó.

1. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp SVM (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm, nhận xét).

* Bài toán phân lớp là bài toán xác định lớp dựa trên đặc trưng của nó. Xác định lớp có xác suất cao nhất, xấp xỉ tốt nhất cho mục đích đó
* Cơ sở lý thuyết: Dựa trên SRM (Strutural risk minimization) để giảm thiểu lỗi trên tập huấn luyện và tăng khả năng tổng quát hóa
* giải pháp chính
  + tập dữ liệu học D={(Xi, Ci), i=1…n}, Ci {-1, 1} xác định dữ liệu dương hay âm
  + tìm một siêu phẳng a\_svm.d + b phân chia dữ liệu thành hai miền
  + phân lớp một dữ liệu mới: xác định dấu của f(d) = a\_svm.d+b
  + thuộc lớp dương nếu f(d) > 0
  + thụôc lớp âm nếu f(d) < 0
* Ưu điểm: Chính xác, tổng quát tốt.
* - Nhược điểm: Tính toán nặng khi mẫu lớn, khó khi số lớp > 2
* - Ứng dụng: SVM được ứng dụng rộng rãi trong các bài toán như phân lớp ảnh, phát hiện xâm nhập, ...

1. Trình bày hai bộ độ đo phổ biến nhất (hồi tưởng-chính xác, chính xác-hệ số lỗi) được sử dụng trong đánh giá các bộ phân lớp dữ liệu. So sánh hai bộ độ đo này.



* Hai độ đo phổ biến nhất để đánh giá mô hình phân lớp là
  + độ hồi tưởng và độ chính xác
    - độ hồi tưởng R (Recall hay còn ký hiêuj p) hoặc TPR (true positive rate): R = TPR = TP/(TP+FN)
    - độ chính xác P (Precision) hoặc pi P = TP/(TP+FP)
  + độ chính xác và hệ số lỗi
    - A = Accuracy = (TP+TN)/(TP+TN+FP+FN)
    - E = Error = (FP + FN) / (TP + TN + FP + FN)
* So sánh 2 bộ đo
  + độ nhạy và độ đặc hiệu: cho biết khả năng phát hiện đúng mô hình ở từng lớp nhưng không cho biết tổng thể mô hìn
  + độ chính xác và hệ số lỗi: cho biết hiệu suát tổng thể của mô hình trên tập dữ liệu nhưng khong nói lên khả năng của mô hình ở từng lớp

1. Phát biểu bàitoán phân cụm. Trình bày thuật toán phân cụm K-mean gán cứng (nội dung cơ bản, ưu điểm- nhược điểm).

* Bài toán
  + cho tập dữ liệu D. Phân các dữ liệu thuộc D thành các cụm. Các dữ liệu trong một cụm tương tự nhau. Dữ liệu trong 2 cụm khác nhau không tương tự nhau
* thuật toán phân cụm k means
  + nội dung cơ bản
    - input và output
      * input: tập dữ liệu D, độ đo tương tự sum, k: số lượng cụm
      * Output, tập k cụm {ci, Ci} với ci là tâm cụm i, Ci là tập dữ liệu cụm i
    - thuật toán
      * + chọn k điểm làm tâm cụm
        + gắn mỗi điểm dữ liệu với trung tâm gần nhất
        + tính toán lại trung tâm cụm dựa vào các điểm đã phân vào phân cum
        + lặp lại bước 2 và 3 đến khi không còn thay đôi phân cụm
  + ưu điểm
    - đơn giản, dễ thực hiện, tính toán hiệu quả
  + nhược điểm
    - kết quả ban đầu phụ thuộc vào việc chọn tâm cụm, khả năng bị mắc cục bộ

1. Phát biểu bàitoán phân cụm. Trình bày thuật toán phân cụm phân cấp HAC (độ đo tương tự cụm, nội dung cơ bản, ưu điểm- nhược điểm).

* Bài toán
  + cho tập dữ liệu D. Phân các dữ liệu thuộc D thành các cụm. Các dữ liệu trong một cụm tương tự nhau. Dữ liệu trong 2 cụm khác nhau không tương tự nhau
* Thuật toán phân cụm phân cấp HAC
  + một số độ đo phân biệt cụm
    - độ tương tự hai dữ liệu
    - độ tương tự giữa 2 cụm
      * độ tương tự giữa 2 đại diện
      * độ tương tự cực đại giữa hai dữ liệu thuộc hai cụm
      * độ tương tự cực tiểu giữa hai dữ liệu thuộc hai cụm
      * độ tương tự trung bình giữa hai dữ liệu thuộc hai cụm
  + nội dung cơ bản
    - input: tập dữ liệu D
    - output: tập các tập cụm phân cấp của D
    - thuật toán:
      * gán mỗi điểm dữ liệu là một cụm riêng
      * tính độ tương tự giữa các cụm thông qua khoảng cách giữa các trung tâm cụm hoặc liên kết giữa các điểm dữ liệu
      * nhập hai cụm gần nhau thành một cụm
      * lặp lại bước 2 và 3 đến khi đạt được số cụm mong muốn
  + ưu ddiểm
    - đơn giản, dễ hiểu, không yêu cầu trước số cụm
  + nhược điểm
    - tốn thời gian tính toán, không phù hợp dữ liệu lớn, kết quả phụ thuộc vào thứ tự nhập cụm

1. Phát biểu bàitoán phân cụm. Trình bày các phương pháp đánh giá thuật toán phân cụm.

* Bài toán phân cụm
  + Cho tập dữ liệu D. Phân các dữ liệu thuộc D thành các cụm. Các dữ liệu trong một cụm tương tự nhau. Dữ liệu trong 2 cụm khác nhau không tương tự nhau
* Các phương pháp đánh giá thuật toán phân cụm
  + Người dùng kiểm tra
    - nghiên cứu trọng tâm và miền phủ
    - luật từ cây quyết định
    - đọc các dữ liệu trong cụm
  + Đánh giá theo các độ đo tương tự/khoảng cách
    - độ phân biệt giữa các cụm
    - phân ly theo trọng tâm
  + Dùng thuật toán phân lớp
    - coi mỗi cụm là một lớp
    - học bộ phân lớp đa lớp (cụm)
    - Xây dựng ma trận nhầm lẫn khi phân lớp
    - tính các độ đo: entropy, tinh khiết, chính xác, hồi tưởng, độ đo F và đánh giá theo các độ đo này

1. Định nghĩa hệ thống tư vấn, các tính chất và so sánh bài toán lọc cộng tác trong hệ tư vấn với bài toán phân lớp.

* định nghĩa: hệ thống tư vấn là các công cụ phần mềm và kỹ thuật cung cấp các tư vấn về các mục có khả năng cao là hữu ích nhất đối với một người dùng đích
* các tính chất của hệ thống tư vấn
  + tính có liên quan
  + tính mới lạ
  + tính may mắn bất ngờ
  + tính đa dạng gia tăng
  + tính giải trình
* So sánh bài toán lọc cộng tác với bài toán phân lớp
  + bài toán phân lớp
    - có các biến độc lập và các biến phụ thuộc
    - phân chia các dữ liệu thành 2 loại dữ liệu học và dữ liệu đánh giá
  + bài toán lọc
    - không chia ranh giới giữa biến phụ thuộc và biến độc lập
    - không chia ranh giới dòng dữ liệu học và dòng dữ liệu đánh giá

1. Giới thiệu chung về hệ thống các kỹ thuật lọc trong hệ tư vấn. Tư vấn xã hội (định nghĩa hẹp và rộng), tư vấn nhóm.

* các kỹ thuật lọc trong hệ tư vấn
  + kỹ thuật lọc cộng tác
  + kỹ thuật dựa trên cộng tác, trí thức, nhân khẩu học, kết hợp
  + một vài phân loại
* Tư vấn xã hội
  + định nghĩa hẹp: sử dụng mỗi quan hệ xã hội
  + định nghĩa rộng: sử dụng mọi dữ liệu từ phương tiện xã hội
* Tư vấn nhóm người dùng
  + kiểu nhóm: chính thức, không thường xuyên, ngẫu nhiên, tự động

1. Trình bày sơ đồ khối thủ tục lọc cộng tác hướng người dùng và giải thích sơ bộ.

Sơ đồ khối thủ tục lọc cộng tác (Collaborative Filtering) hướng người dùng:

Người dùng

|

|

|

Mô hình học

|

|

Dự đoán xếp hạng cho người dùng mới

Giải thích:

- Thu thập dữ liệu xếp hạng (rating) của người dùng cho các sản phẩm.

- Sử dụng mô hình học máy (thường là hồi quy, NNMF,...) để học mô hình từ dữ liệu xếp hạng của người dùng.

- Mô hình sẽ đưa ra dự đoán xếp hạng cho người dùng mới dựa trên mô hình đã học.

- Các xếp hạng dự đoán này được sử dụng để đưa ra khuyến nghị cho người dùng mới (nếu xếp hạng cao => khuyến nghị mạnh hơn).

Ưu điểm: Giúp tăng trải nghiệm người dùng bằng các khuyến nghị phù hợp với sở thích cá nhân của họ.

Nhược điểm: Cần lượng lớn dữ liệu để có hiệu quả và tránh overfitting. Thường khó mô hình hóa người dùng mới.

Ví dụ: Hệ thống khuyến nghị sản phẩm trên các trang thương mại điện tử như Amazon.

1. Trình bày sơ đồ khối thủ tục lọc cộng tác hướng mục và giải thích sơ bộ.

Sơ đồ khối thủ tục lọc cộng tác (Collaborative Filtering) hướng mục:

Mục (sản phẩm)

|

|

|

Mô hình học

|

|

Dự đoán xếp hạng cho người dùng mới

Giải thích:

- Thu thập dữ liệu xếp hạng (rating) của người dùng cho các sản phẩm.

- Sử dụng mô hình học máy (thường là hồi quy, NNMF,...) để học mô hình từ dữ liệu xếp hạng của sản phẩm.

- Mô hình sẽ dự đoán xếp hạng mà người dùng chưa đánh giá có thể cho sản phẩm đó.

- Các xếp hạng dự đoán này được sử dụng để đưa ra khuyến nghị cho người dùng (nếu xếp hạng cao => khuyến nghị mạnh hơn).

Ưu điểm: Tương tự hướng người dùng.

Nhược điểm: Khó học được đặc trưng của người dùng, chủ yếu tập trung vào sản phẩm.

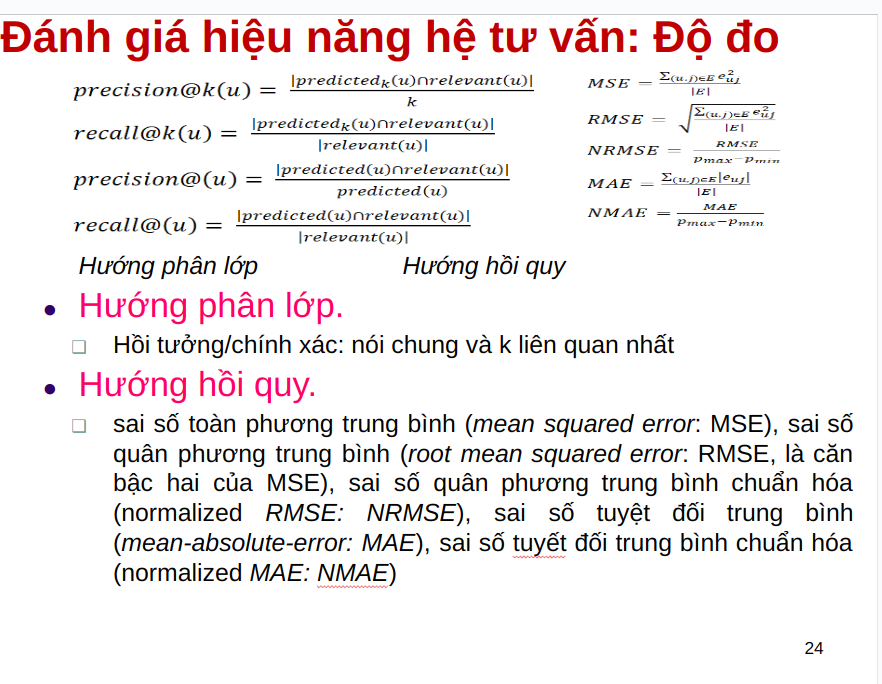
Ví dụ: Hệ thống khuyến nghị sản phẩm trên các trang thương mại điện tử.

So sánh 2 hướng trên: Hướng người dùng linh hoạt và hiệu quả hơn nhưng khó triển khai hơn. Hướng mục đơn giản hơn nhưng kém linh hoạt. Thường kết hợp cả 2 hướng cho kết quả tốt nhất

1. Các thành phần của bộ lọc dựa trên nội dung trong hệ tư vấn.

* Các thành phần chính của bộ lọc dựa trên nội dung (Content-based Filtering) trong hệ thống tư vấn:
* 1. Bộ thu thập nội dung: Thu thập dữ liệu mô tả các đối tượng (sản phẩm, tin tức, video,...) như tiêu đề, tóm tắt, thẻ tag, ảnh đại diện,...
* 2. Bộ biểu diễn nội dung: Biểu diễn các nội dung thu thập được ở dạng vector đặc trưng cho phép so sánh giữa các nội dung. Các kỹ thuật thường dùng là Bag-of-words, TF-IDF, Word2Vec, Doc2Vec,..
* 3. Bộ đo độ tương đồng: Đo độ giống nhau giữa các nội dung dựa trên vector đặc trưng. Các thước đo phổ biến là cosine similarity, Euclidean distance,...
* 4. Bộ lọc và xếp hạng: Lọc các nội dung giống với nội dung mà người dùng quan tâm dựa trên độ tương đồng. Xếp hạng các kết quả này dựa trên độ tương đồng để khuyến nghị cho người dùng.
* 5. Bộ đánh giá và phản hồi: Đánh giá độ chính xác của các khuyến nghị từ phản hồi của người dùng để cải tiến hệ thống.
* Ưu điểm: Không cần lượng lớn dữ liệu người dùng.
* Nhược điểm: Khó bắt được thông tin ngữ cảnh và sở thích tinh tế của người dùng.

1. Các phương thức đánh giá hiệu năng hệ thống tư vấn và các độ đo liên quan.



1. Cho tập ví dụ có nhãn nhị phân trên mặt phẳng trong đó là tọa độ của điểm dữ liệu, hoặc là nhãn của điểm dữ liệu tương ứng. Hãy xác định nhãn của điểm dữ liệu mới theo thuật toán 3-NN.
2. Cho một cơ sở dữ liệu giao dịch gồm 9 giao dịch (giao dịch ::=”(”<mục>[,<mục]>]”)”

(a,b,e), (b,d), (b,c), (a,b,d), (a,c), (b,c), (a,c), (a,b,c,e), (a,b,c). Cho độ hỗ trợ tối thiểu so=0.5.

Hãy mô tả hoạt động thuật toán Apriori theo F1, C2, F2, C3, F3, … với Fi là tập các tập mục phổ biến có độ dài i, Ci là tập ứng viên độ dài i.

1. Giả sử cần phân lớp nhị phân có 2 lớp với nhãn lớp tương ứng là 1, 0. Bảng sau đây trình bày một tập dữ liệu test gồm 42 ví dụ:

| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

Sau khi xây dựng xong một bộ phân lớp nhị phân, người ta cho kiểm tra với dữ liệu test trên đây và nhận được kết quả tổng hợp tương ứng như sau (thứ tự vị trí dữ liệu test giữ nguyên):

| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

Hãy đánh giá bộ phân lớp nhị phân nói trên qua tính toán (cần viết công thức tính)

* Cặp độ đo (độ chính xác, độ hồi tưởng) và độ đo F1.
* Cặp độ đo (độ chính xác, hệ số lỗi).

*Ghi chú: Số liệu trong trong bài thi đối với ba câu hỏi 18, 19, 20 sẽ khác đi.*MỘT BÀI THI MẪU NĂM HỌC HỌC KỲ 2 NĂM HỌC 2022-2023 (Thi 13h00 ngày 05/06/2023)

**Đề 1.** *Làm bài trong 25 phút lên tờ giấy thi gốm hai trang này*

**Câu 1. Mệnh đề đúng: điền ✔vào ◻ để nhận được 🗹** (**đánh dấu vào mệnh đề sai sẽ bị trừ điểm**)

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

◻

**Câu 2.** Giả sử cần phân lớp dữ liệu với miền dữ liệu có hai lớp với nhãn lớp tương ứng là 1, 0. Bảng sau đây trình bày một tập dữ liệu test gồm 10 ví dụ

| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Một bộ phân lớp nhị phân cho kết quả phân lớp đối với 10 dữ liệu test trên đây như sau

| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Tính giá trị các độ đo: độ chính xác, độ hồi tưởng và độ đo F1.

**Câu 3.** Trình bày cô đọng (ngắn gọn - đủ ý) về thuật toán cây quyết định: mô hình phân lớp, cây quyết định, tư tưởng chung của thuật toán cây quyết định, tối ưu hóa trong thuật toán cây quyét định.