HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 9(c): Khai phá dữ liệu

Lê Hải Hà

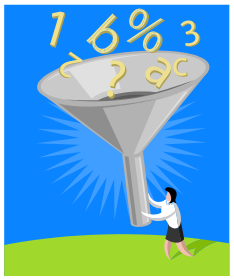
Nội dung

❶ Khái niệm khai phá dữ liệu
❷ Các ứng dụng của khai phá dữ liệu
❸ Các quá trình khai phá dữ liệu
❹ Các thuật toán khai phá dữ liệu
- Phân lớp
- Phân cụm
- Luật liên kết

Tại sao phải khai phá dữ liệu?

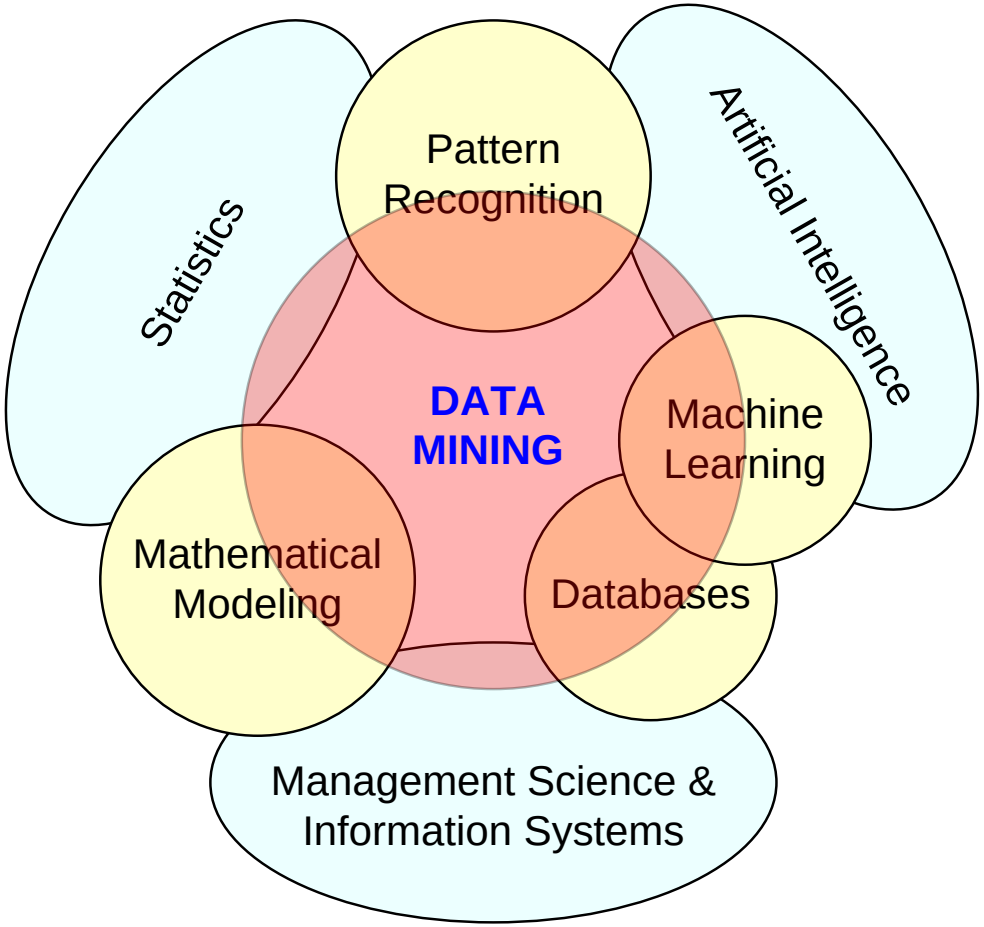
- Cạnh tranh khốc liệt hơn ở quy mô toàn cầu
- Ghi nhận giá trị trong các nguồn dữ liệu
- Tính sẵn dùng của các dữ liệu chất lượng về khách hàng, nhà cung cấp, giao dịch, Web, …
- Hợp nhất và tích hợp các kho dữ liệu đơn lẻ và các kho dữ liệu chung
- Khả năng xử lý dữ liệu và khả năng lưu trữ tăng theo cấp số nhân; và giảm giá thành
- Sự dịch chuyển từ các nguồn thông tin sang dạng phi vật lý

Định nghĩa khai phá dữ liệu



- Tiến trình không tầm thường để xác định các mẫu hợp lệ, mới, hữu dụng và có thể hiểu được trong dữ liệu được lưu trữ ở các dạng CSDL có cấu trúc.
  - Fayyad et al., (1996)
- Các từ khóa chính trong định nghĩa này: *Tiến trình, không tầm thường, hợp lệ, mới, hữu dụng, có thể hiểu.*
- Khai phá dữ liệu: liệu có nhầm lẫn gì không?
- Các tên khác: trích chọn tri thức, phân tích mẫu, khám phá tri thức, gặt hái thông tin, tìm kiếm mẫu, nạo vét dữ liệu,…

DM ở vùng giao của nhiều lĩnh vực

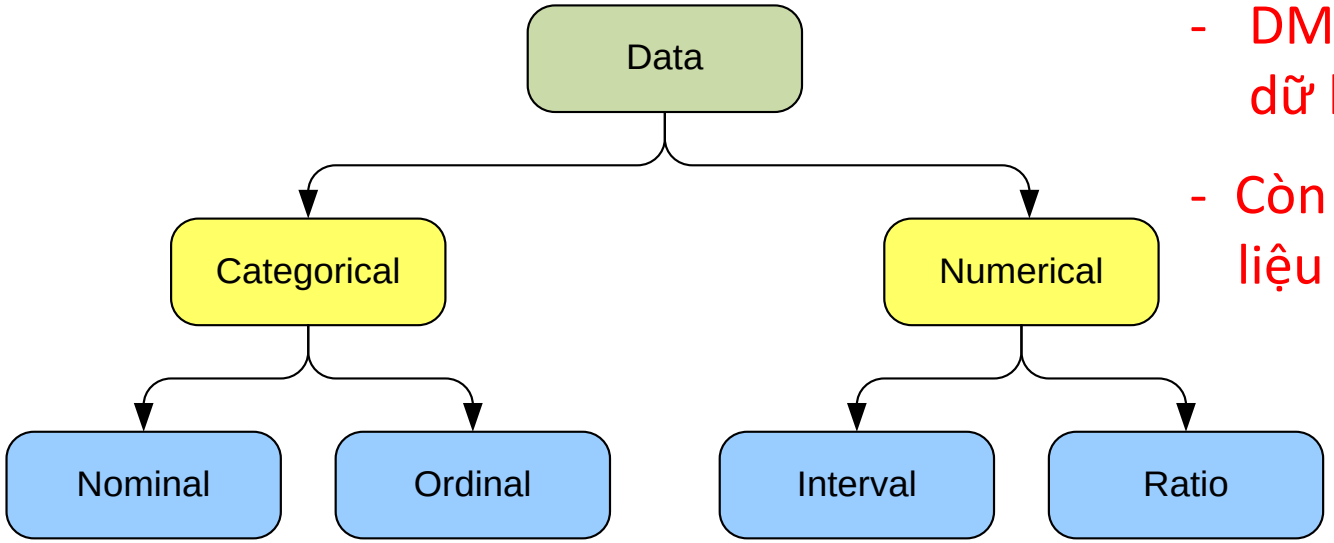


Các tính chất/mục tiêu của DM

- Nguồn dữ liệu cho DM thường là các Data Warehouse (DW) hợp nhất (không phải lúc nào cũng vậy!)
- Môi trường DM thường là kiến trúc client-server hay kiến trúc hệ thống thông tin dựa trên Web
- Dữ liệu là thành phần quan trọng nhất của DM, có thể bao gồm dữ liệu mềm/không cấu trúc
- Người khai phá thường là người dùng cuối
- Đòi hỏi sáng tạo thông qua các quy trình và diễn dịch kết quả tìm thấy
- Khả năng và tính dễ sử dụng của các công cụ DM là điều cần thiết (Web, xử lý song song, …)

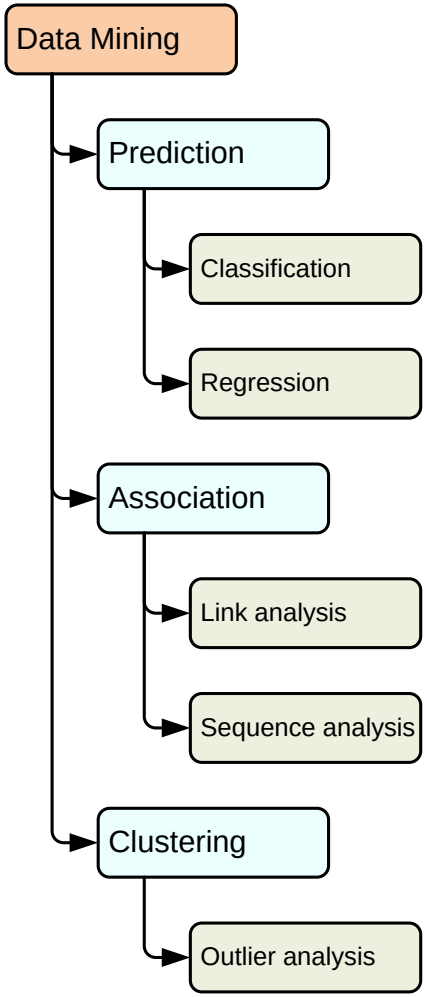
Dữ liệu trong DM

- Dữ liệu: thu thập của các sự kiện, thường có được như là kết quả của các trải nghiệm, quan sát, hay thực nghiệm
- Dữ liệu có thể chứa các con số, các từ, các hình ảnh, …
- Dữ liệu: mức trừu tượng thấp nhất (từ đó trích rút ra các thông tin và tri thức)



- DM với các dạng dữ liệu khác nhau?
- Còn các dạng dữ liệu khác không?

Phân loại các nhiệm vụ của DM



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

DM làm gì?

- DM trích các mẫu từ dữ liệu
  - Mẫu? Quan hệ (số hay biểu tượng) toán học giữa các mục dữ liệu

- Các dạng mẫu
  - Liên kết/Association
  - Dự báo/Prediction
  - Phân cụm/Cluster (segmentation)
  - Các quan hệ tuần tự (hay chuỗi thời gian)

Các ứng dụng của DM

- Quản lý quan hệ khách hàng
  - Tối đa hóa lợi nhuận trên các chiến dịch tiếp thị
  - Cải thiện khả năng giữ chân khách hàng (churn analysis)
  - Tối đa hóa giá trị khách hàng (bán kèm, bán thêm)
  - Nhận diện và ứng xử với các khách hàng có giá trị nhất
- Ngân hàng và tổ chức tài chính khác
  - Tự động hóa quy trình duyệt đơn vay
  - Phát hiện giao dịch gian lận
  - Tối đa hóa giá trị khách hàng (bán kèm, bán thêm)
  - Tối ưu hóa dự trữ tiền mặt với các dự báo

Các nhiệm vụ của DM (tiếp)

- Dự báo chuỗi thời gian
  - Một phần của phân tích tuần tự hay liên kết?
- Trực quan hóa
  - Còn các nhiệm vụ khác của DM?

- Các dạng DM
  - Khai phá dữ liệu theo hướng giả thuyết
  - Khai phá dữ liệu theo hướng khám phá

Các ứng dụng của DM (tiếp)

- Bán lẻ và Logistic
  - Tối ưu hóa mức tồn kho ở các vị trí khác nhau
  - Cải thiện việc bố trí (layout) trong các cửa hàng và khuyến mãi
  - Tối ưu hóa hậu cần dựa trên các dự báo về ảnh hưởng mùa
  - Giảm thiểu tổn thất do hạn sử dụng
- Sản xuất và bảo trì
  - Dự đoán/ngăn ngừa hỏng máy móc
  - Xác định các điểm bất thường trong hệ thống sản xuất để tối ưu khả năng sản xuất
  - Khám phá các mẫu mới để cải tiến chất lượng sản phẩm

Các ứng dụng của DM (tiếp)

- Môi giới và giao dịch chứng khoán
  - Dự đoán các thay đổi của giá một trái phiếu nào đó
  - Dự đoán hướng biến động của chứng khoán
  - Đánh giá ảnh hưởng của các sự kiện đến diễn biến thị trường
  - Xác định và ngăn chặn các hoạt động gian lận trong giao dịch
- Bảo hiểm
  - Dự báo chi phí bồi thường để có kế hoạch tốt hơn
  - Xác định các kế hoạch tỉ lệ tối ưu
  - Tối ưu hóa việc tiếp thị tới các khách hàng cụ thể
  - Xác định và ngăn chặn các hoạt động yêu cầu bồi thường gian lận

Các ứng dụng của DM (tiếp)

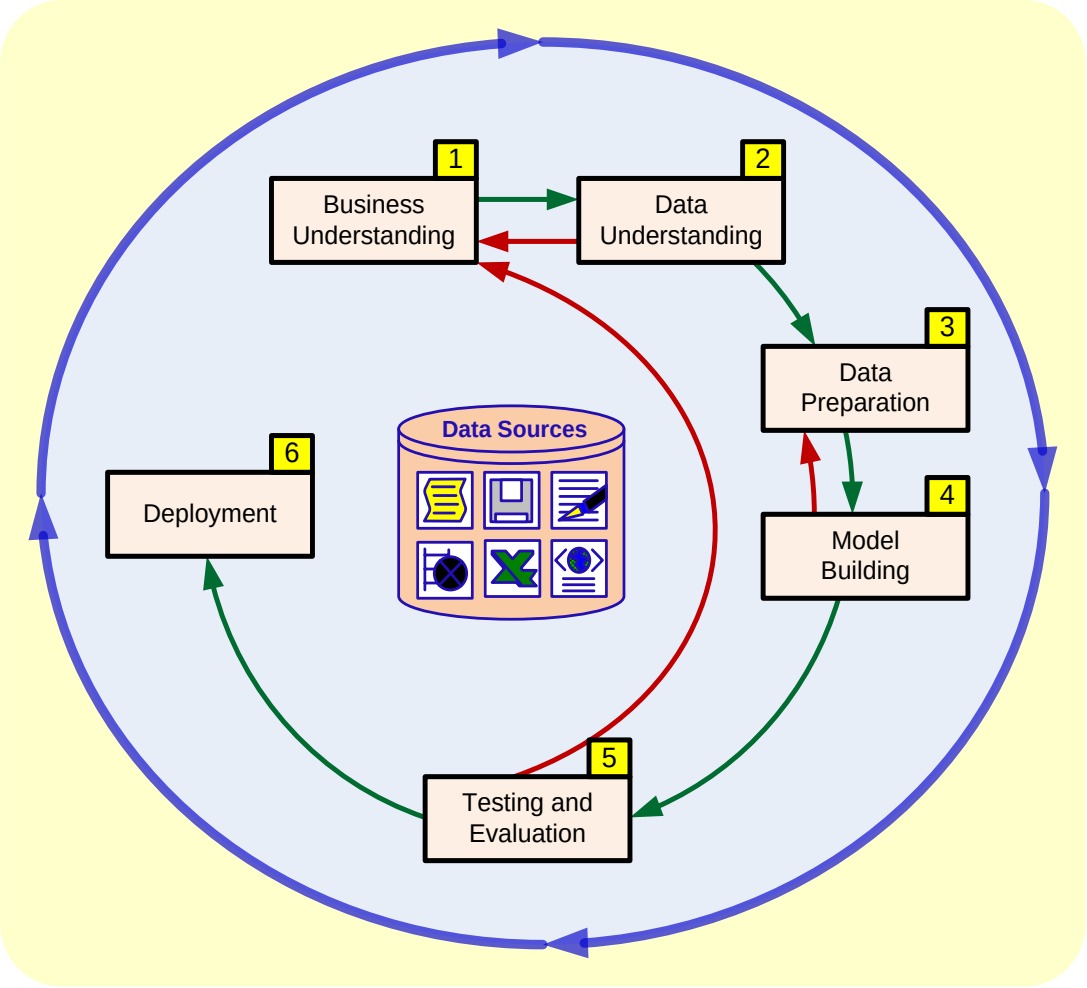
- Phần cứng và phần mềm máy tính
- Khoa học và công nghệ
- Chính phủ và quốc phòng
- An ninh nội địa và thực thi pháp luật
- Du lịch
- Chăm sóc sức khỏe
- Y tế
- Công nghiệp giải trí
- Thể thao
- V.v…

Các lĩnh vực ứng dụng phổ biến cho DM

Quy trình khai phá dữ lệu

- Cách có hệ thống để tiến hành các dự án DM
- Dựa trên các thực hành tốt nhất
- Các nhóm khác nhau có các phiên bản khác nhau
- Các quy trình chuẩn chung nhất:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Quy trình DM: CRISP-DM



Quy trình DM: CRISP-DM

Bước 1: Hiểu nghiệp vụ

Bước 2: Hiểu dữ liệu

Bước 3: Chuẩn bị dữ liệu (!)

Bước 4: Xây dựng mô hình

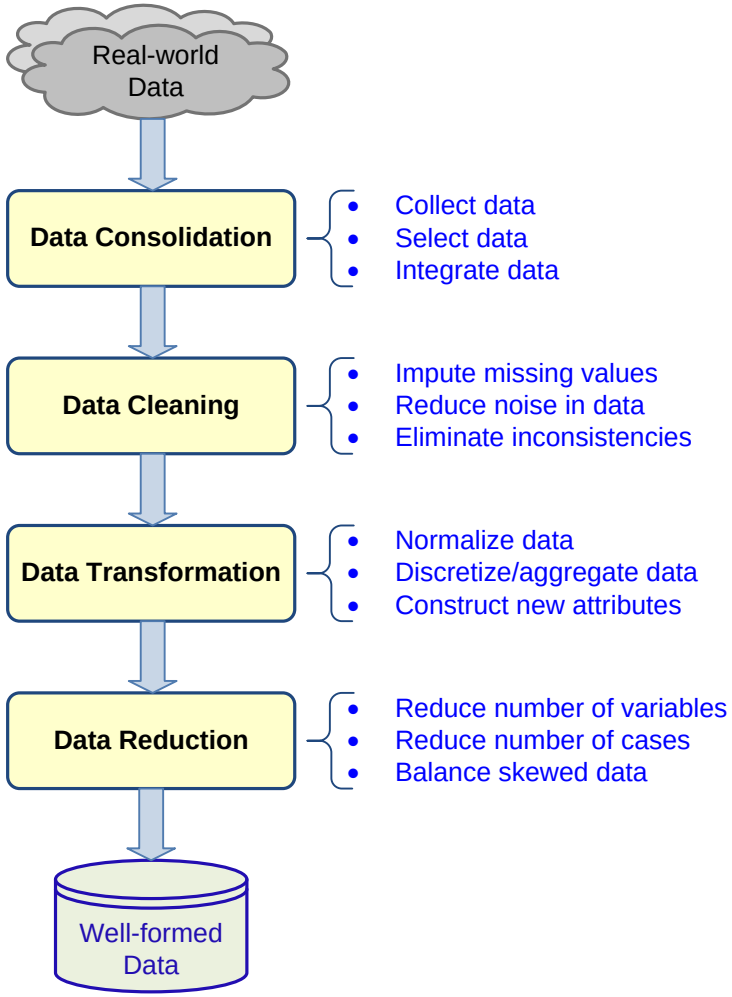
Bước 5: Kiểm định và đánh giá

Bước 6: Triển khai

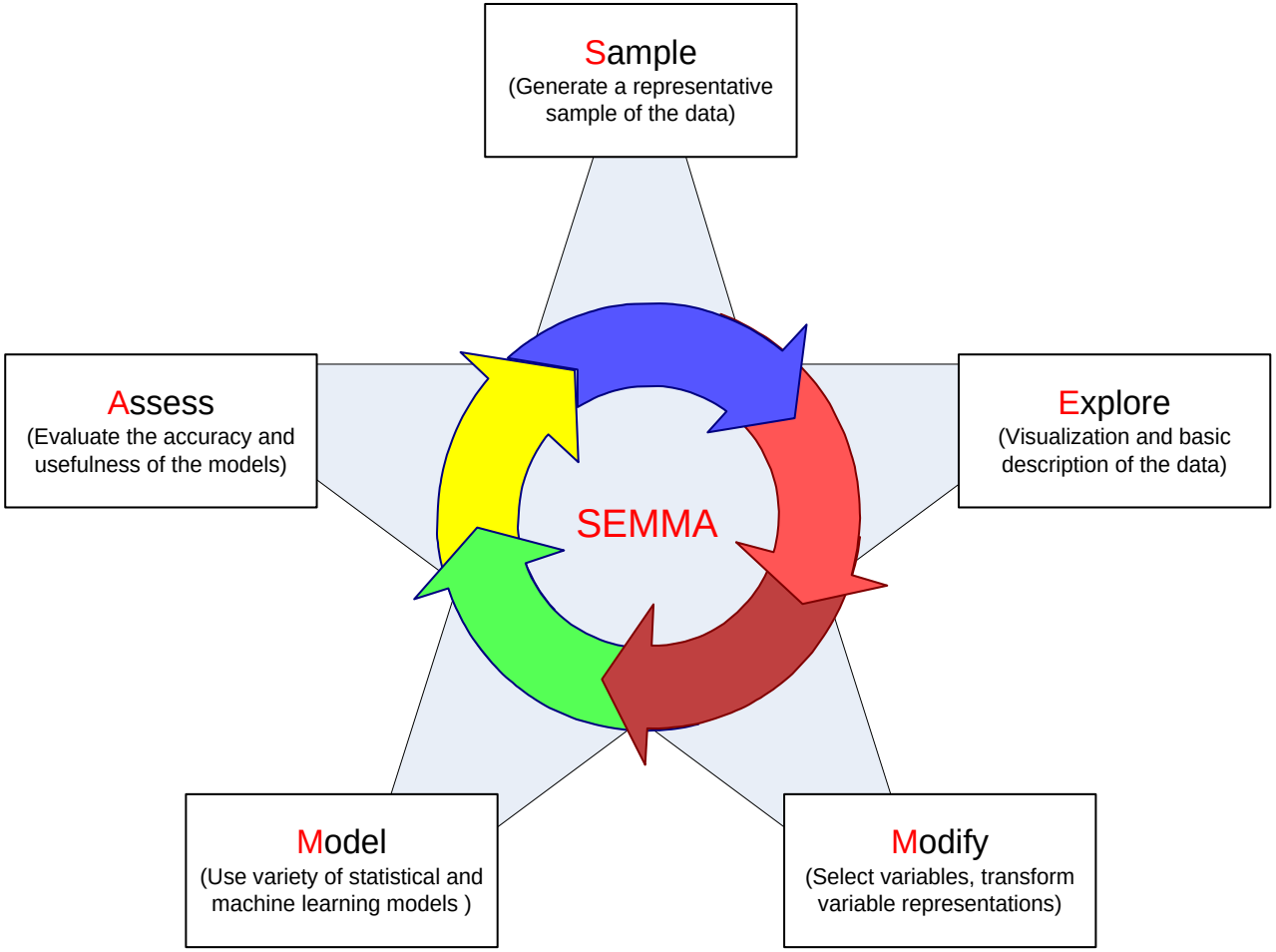
Chiếm xấp xỉ 85% thời gian dự án

- Quy trình có tính thử nghiệm và lặp cao (DM: nghệ thuật và khoa học?)

Chuẩn bị dữ liệu – Công việc DM quan trọng



Quy trình DM: SEMMA



Các phương pháp DM: Phân lớp/Classification

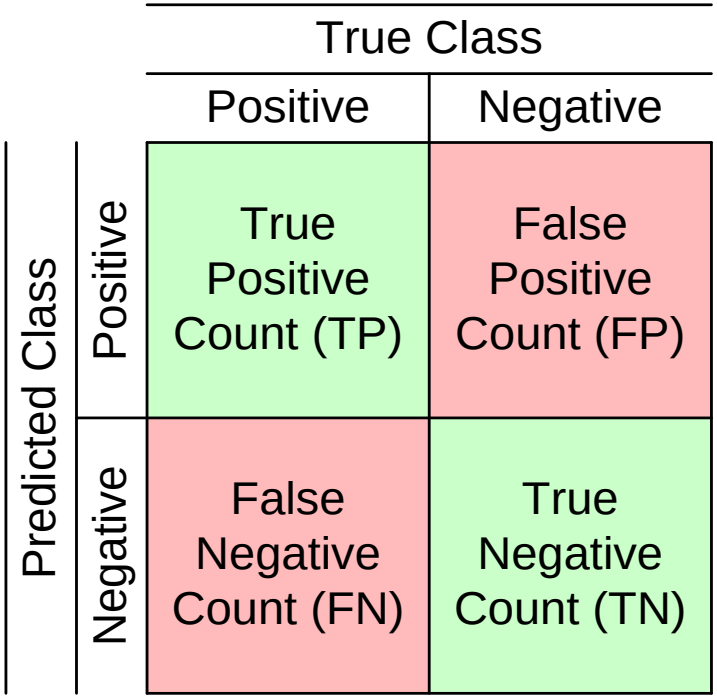
- Là phương pháp DM được sử dụng phổ biến nhất
- Là một phần trong họ máy học
- Sử dụng học có giám sát
- Học từ dữ liệu quá khứ, phân lớp dữ liệu mới
- Biến đầu ra có bản chất phân loại (danh nghĩa hoặc thứ tự - nominal or ordinal)
- Phân lớp và hồi quy?
- Phân lớp và phân cụm?

Các phương pháp đánh giá phân lớp

- Độ chính xác dự báo
  - Tỷ lệ đúng
- Tốc độ
  - Xây dựng mô hình; dự báo
- Độ mạnh
- Khả năng mở rộng
- Khả năng diễn giải
  - Tính minh bạch, khả năng giải thích

Độ chính xác của các mô hình phân lớp

- Trong bài toán phân lớp, nguồn chính để ước lượng độ chính xác là confusion matrix



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

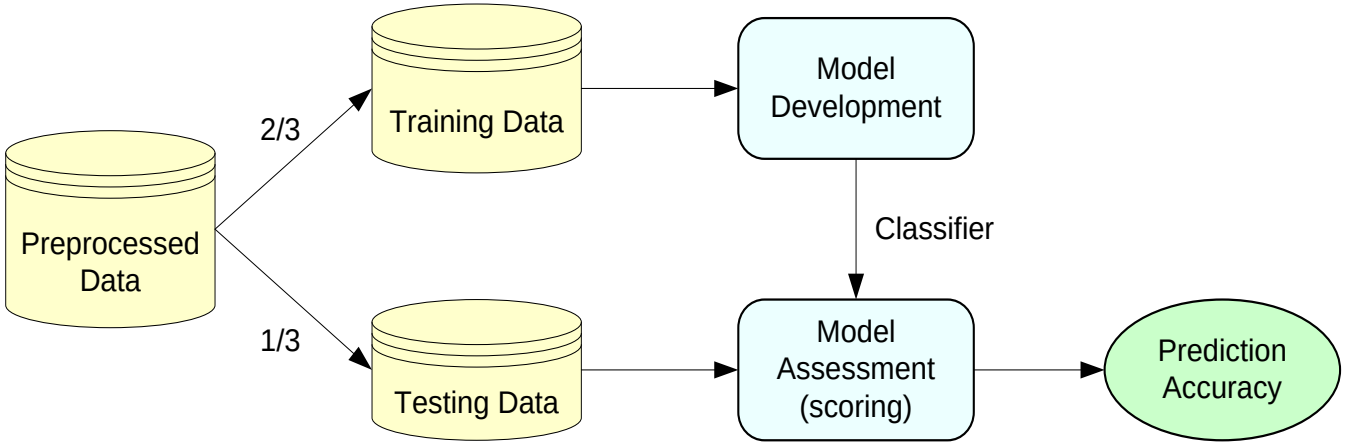
$$True\ PositiveRate = \frac{TP}{TP + FN}$$

$$True\ NegativeRate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Các phương pháp ước lượng đối với việc phân lớp

- Phân tách đơn giản (ước lượng mẫu kiểm định hay giữ lại)
  - Chia dữ liệu vào 2 phần không giao nhau: dữ liệu luyện (~70%) và dữ liệu kiểm định (30%)

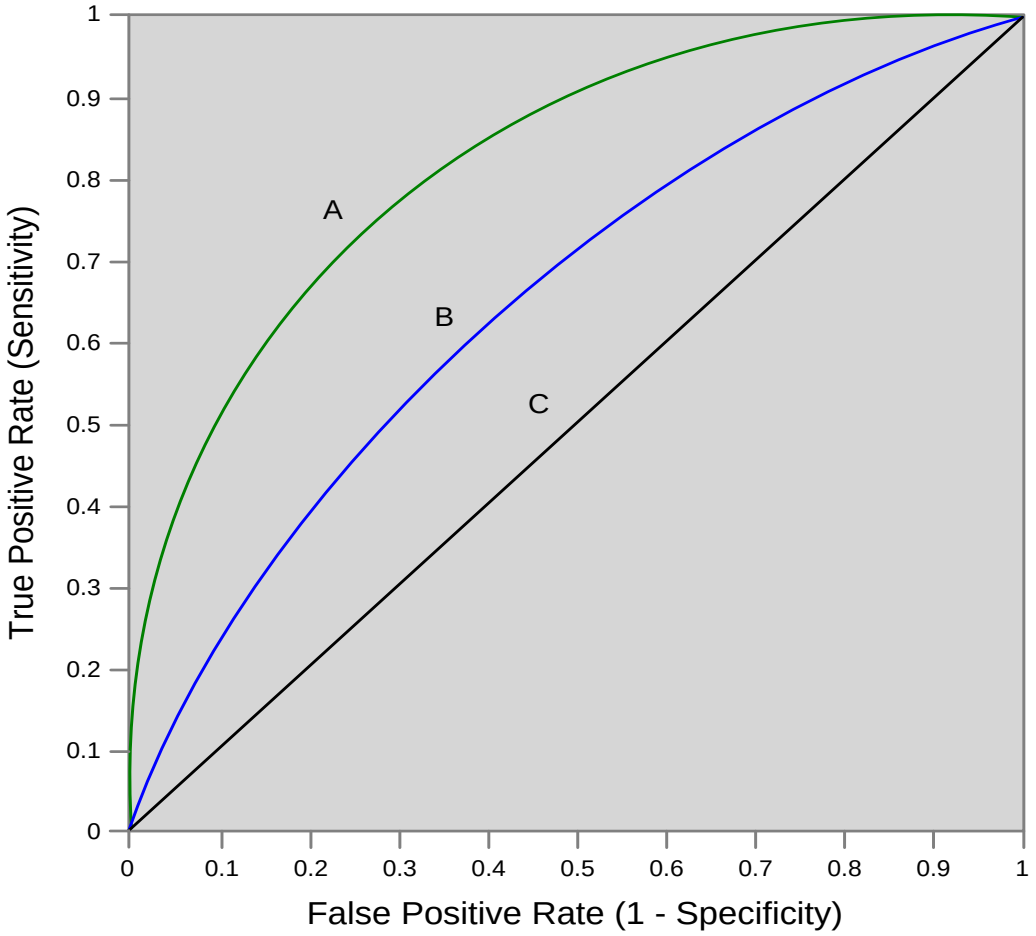


  - Đối với ANN (mạng neural nhân tạo), dữ liệu được chia vào 3 tập con (training [~60%], validation [~20%], testing [~20%])

Các phương pháp ước lượng đối với việc phân lớp

- *k*-Fold Cross Validation (ước lượng quay vòng)
  - Chia dữ liệu thành *k* tập con không giao nhau
  - Sử dụng mỗi tập con làm dữ liệu kiểm định, các tập con còn lại làm dữ liệu luyện
  - Lặp lại thực nghiệm *k* lần
  - Tổ hợp các kết quả kiểm định để có ước lượng đúng về độ chính xác quá trình luyện
- Một số phương pháp luận ước lượng khác
  - Leave-one-out, bootstrapping, jackknifing
  - Area under the ROC curve

Các phương pháp ước lượng đối với việc phân lớp – ROC Curve



Các kỹ thuật phân lớp

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

Cây quyết định

- Sử dụng phương pháp chia để trị
- Chia tập luyện tới khi mỗi phần chia chỉ chứa toàn bộ (hay phần chính) các mẫu của một lớp

Giải thuật chung để xây dựng cây quyết định

1. Tạo nút gốc và gán tất cả dữ liệu luyện vào nó
2. Chọn thuộc tính phân chia tốt nhất
3. Thêm 1 nhánh cho mỗi giá trị của thuộc tính phân chia. Phân dữ liệu vào các tập con không giao nhau theo các nhánh chia xác định
4. Lặp lại bước 2 và 3 đối với mỗi nút lá tới khi đạt được tiêu chuẩn dừng

Cây quyết định

- Các giải thuật cây quyết định khác nhau ở các điểm chính sau:
  - Tiêu chuẩn chia
    - Biến/thuộc tính nào được dùng để phân chia?
    - Chia theo những giá trị nào?
    - Tách mỗi nút thành bao nhiêu phần?
  - Tiêu chuẩn dừng
    - Khi nào dừng việc xây dựng cây
  - Cắt tỉa (phương pháp tổng quát hóa)
    - Pre-pruning và post-pruning
- Các giải thuật DT phổ biến nhất gồm
  - ID3, C4.5, C5; CART; CHAID; M5

Cây quyết định

- Một số tiêu chuẩn chia
  - Gini index xác định sự cắt tỉa một lớp cụ thể như kết quả của quyết định phân nhánh theo thuộc tính/giá trị cụ thể
    - Được sử dụng trong CART
  - Information gain sử dụng entropy để đo lường mức độ không chắc chắn hoặc ngẫu nhiên của một thuộc tính/giá trị chia cụ thể
    - Được sử dụng trong ID3, C4.5, C5
  - Chi-square statistics (được sử dụng trong CHAID)

Information gain

- Sử dụng Entropy để đo mức độ không chắc chắn trong 1 tập hợp
- Thí dụ tập *S* với *p* mẫu lớp *P* và n mẫu lớp *N*
- Số thông tin cần để quyết định một mẫu tùy ý trong *S* thuộc lớp *P* hay *N* được xác định là

$$I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

- Nếu sử dụng thuộc tính *A* chia *S* thành các tập con $S_1, S_2, \ldots, S_v$. Kỳ vọng thông tin

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)$$

- Thông tin đạt được

$$Gain(A) = I(p, n) - E(A)$$

- *A* được chọn sao cho Gain đạt giá trị lớn nhất

Chỉ số Gini

- Đo mức độ đa dạng của quần thể
- Tập S với n lớp có

$$gini(S) = 1 - \sum_{j=1}^{n} p_j^2$$

- Với p_j là tần suất tương đối của lớp j trong S
- Nếu S được chia thành 2 tập con S_1, S_2 với kích thước N_1, N_2 tương ứng

$$gini_{split}(S) = \frac{N_1}{N} gini(S_1) + \frac{N_2}{N} gini(S_2)$$

- Thuộc tính và giá trị tách được chọn sao cho chỉ số Gini phân tách nhỏ nhất

Phân tích phân cụm trong DM

- Được sử dụng để xác định sự phân nhóm tự nhiên của các đối tượng
- Là phần của họ học máy
- Sử dụng học không giám sát
- Học các cụm đối tượng từ dữ liệu quá khứ, sau đó gán các thực thể mới vào cụm tương ứng
- Không có biến đầu ra
- Cũng được biết là dạng phân mảnh (segmentation)

Phân tích phân cụm trong DM

- Các kết quả phân cụm có thể được sử dụng để
  - Xác định các nhóm khách hàng
  - Xác định các luật phân nhóm các tình huống mới đối với các mục đích chuẩn đoán
  - Cung cấp đặc điểm, định nghĩa, gán nhãn các quần thể
  - Giảm kích thước và độ phức tạp của vấn đề cho các phương thức khai phá dữ liệu khác
  - Xác định các ngoại lai (outliers) trong các lĩnh vực cụ thể (thí dụ: phát hiện sự kiện hiếm)

Phân tích phân cụm trong DM

- Các phương pháp phân tích
  - Các phương pháp thống kê (cả phân cấp và không phân cấp), như *k*-means, *k*-modes
  - Mạng neural (adaptive resonance theory [ART], self-organizing map [SOM])
  - Logic mờ (thí dụ: giải thuật fuzzy c-means)
  - Giải thuật di truyền

- Các phương pháp có thể nhóm vào nhóm phân chia hay nhóm tổng hợp

Phân tích phân cụm trong DM

- Có bao nhiêu cụm?
  - Không có cách nào "thực sự tối ưu" để tính nó
  - Thường sử dụng Heuristic
    - Xem xét mức thưa của các cụm
    - Số cụm = $(n/2)^{1/2}$ (n: số điểm dữ liệu)
    - Sử dụng tiêu chí thông tin Akaiken (Akaiken Information Criterion - AIC)
    - Sử dụng tiêu chí thông tin Bayes (Bayesian Information Criterion - BIC)
- Hầu hết các phương pháp phân tích sử dụng số đo khoảng cách để tính mức độ gần giữa các cặp điểm
  - Khoảng cách Euclidian và Manhattan (khoảng cách vuông góc)

Phân tích phân cụm trong DM

- *k*-Means Clustering Algorithm
  - *k* : số cụm xác định trước
  - Giải thuật (Bước 0: xác định giá trị *k*)

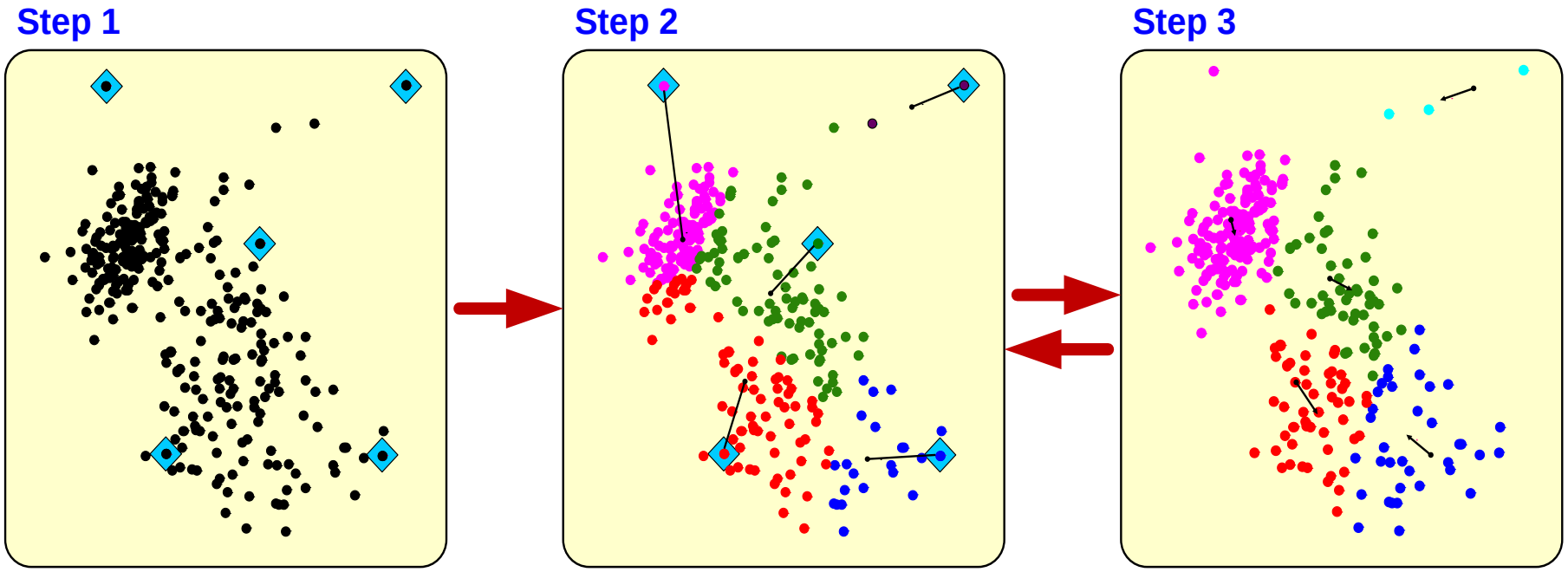
  Bước 1: Tạo ngẫu nhiên *k* điểm như là tâm cụm ban đầu

  Bước 2: Gán mỗi điểm tới tâm cụm gần nhất

  Bước 3: Tính lại các tâm cụm mới

  Lặp: Lặp lại bước 2 và 3 tới khi đạt được tiêu chuẩn hội tụ (thông thường là tới khi việc gắn các điểm tới tâm cụm ổn định)

Phân tích phân cụm trong DM - Giải thuật phân cụm *k*-Means



Khai phá luật liên kết

- Một phương pháp DM rất phổ biến trong kinh doanh
- Tìm kiếm các quan hệ thú vị giữa các biến (các mục hay sự kiện)
- Là phần của họ máy học
- Sử dụng học không giám sát
- Không có biến đầu ra
- Cũng được gọi là phân tích rổ thị trường (market basket analysis)
- Thường được sử dụng như thí dụ điển hình để mô tả DM với người bình thường. Thí dụ "quan hệ giữa tã trẻ em và bia!"

Khai phá các luật liên kết

- Đầu vào: dữ liệu giao dịch POS (point-of-sale)
- Đầu ra: Mối quan hệ thường xuyên nhất giữa các mục
- Thí dụ: theo dữ liệu giao dịch…

  "70% khách hàng mua máy tính và phần mềm chống virus cũng mua thêm gói dịch vụ."
- Sử dụng mẫu/tri thức như vậy thế nào?
  - Xắp đặt các mặt hàng gần nhau để dễ tìm
  - Bán các mặt hàng theo gói
  - Đặt các mặt hàng xa nhau để khách hàng phải đi bộ trên các lối đi để tìm kiếm nó và như vậy có khả năng nhìn thấy và mua các mặc hàng khác

Khai phá các luật liên kết

- Các ứng dụng đại diện của khai phá luật liên kết gồm
  - Trong kinh doanh: thị trường chéo/cross-marketing, bán kèm/cross-selling, thiết kế cửa hàng/store design, thiết kế danh mục/catalog design, thiết kế website thương mại điện tử, tối ưu hóa quản cáo trực tuyến, định giá sản phẩm, cấu hình bán hàng/khuyến mãi
  - Trong y học: quan hệ giữa triệu chứng và bệnh tật; đặc điểm của bệnh nhân và chuẩn đoán với phương pháp điều trị (được sử dụng trong DSS y học); và quan hệ giữa gen với chức năng của chúng (được sử dụng trong các dự án gen)…

Khai phá các luật liên kết

- *Có phải tất cả các luật liên kết đều thú vị và hữu dụng?*

  Quy tắc chung: **X ⇒ Y [S%, C%]**

  X, Y: sản phẩm hay dịch vụ

  X: bên trái (LHS)

  Y: bên phải (RHS)

  S: Support/hỗ trợ: tần suất **X** và **Y** xuất hiện cùng nhau

  C: Confidence/tin cậy: tần suất **Y** xuất hiện với **X**

  Thí dụ: {Laptop Computer, Antivirus Software} ⇒
  {Extended Service Plan} [30%, 70%]

Khai phá các luật liên kết

- Một số giải thuật phát thiện các luật liên kết
  - Apriori
  - Eclat
  - FP-Growth
  - + Các dẫn xuất và lai của 3 giải thuật trên
- Các giải thuật giúp xác định tần suất các tập phần tử, và sau đó cần được chuyển vào các luật liên kết

Khai phá các luật liên kết

- Giải thuật Apriori
  - Tìm tập con phổ biến cho ít nhất một số tối thiểu trong tập mục
  - Sử dụng tiếp cận từ dưới lên
    - Các tập con được mở rộng từng mục một (kích thước của tập con tăng từ tập con 1 mục tới tập con 2 mục, sau đó là tập con 3 mục, …), và
    - Các nhóm ứng cử ở mỗi mức được kiểm tra lại đã đạt mức support tối thiểu
    - Xem hình…

Khai phá các luật liên kết

- Giải thuật Apriori

Raw Transaction Data

Transaction No	SKUs (Item No)
1	1, 2, 3, 4
1	2, 3, 4
1	2, 3
1	1, 2, 4
1	1, 2, 3, 4
1	2, 4

One-item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-item Itemsets

Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Các lầm tưởng về DM

- Khai phá dữ liệu …
  - *Cung cấp các giải pháp/dự báo trực tuyến/tức thì*
  - *Chưa khả thi đối với các ứng dụng kinh doanh*
  - *Yêu cầu CSDL riêng, chuyên dụng*
  - *Chỉ có thể được làm bởi các chuyên gia*
  - *Chỉ dành cho các công ty lớn với nhiều dữ liệu khách hàng*
  - *Là tên khác của thống kê lạc hậu*

Các lỗi chung của DM

1. Chọn không đúng vấn đề cho DM
2. Bỏ qua những gì nhà tài trợ nghĩ là DM và cái gì DM có thể/hay không thể làm
3. Không dành đủ thời gian cho việc thu thập, lựa chọn và chuẩn bị dữ liệu
4. Chỉ xem xét các kết quả tổng hợp mà không phải các bản ghi/dự báo đơn lẻ
5. Không cẩn thận trong việc theo dõi quy trình và kết quả khai phá dữ liệu

Các lỗi chung của DM

6. Bỏ qua các phát hiện đáng ngờ (tốt hay xấu) và nhanh chóng chuyển tiếp
7. Thực hiện các giải thuật khai pháp liên tục và mù quáng, không nghĩ tới giai đoạn tiếp
8. Ngây thơ tin tưởng mọi thứ về dữ liệu
9. Ngây thơ tin tưởng mọi thứ về phân tích khai phá dữ liệu
10. Đo lường kết quả khác với cách nhà tài trợ đo lường

Tổng kết

- Xác định khai phá dữ liệu như một công nghệ hỗ trợ kinh doanh thông minh
- Hiểu mục tiêu và lợi ích của phân tích kinh doanh và khai phá dữ liệu
- Ghi nhận một dải rộng các ứng dụng của khai phá dữ liệu
- Học các quy trình khai phá dữ liệu chuẩn
  - CRISP-DM,
  - SEMMA,
  - KDD, …

Tổng kết (tiếp)

- Hiểu các bước liên quan tới tiền xử lý dữ liệu cho DM
- Học các phương pháp và giải thuật DM khác nhau
- Xây dựng nhận thức về các công cụ phần mềm DM hiện tại
  - Thương mại hay nguồn mở/miễn phí
- Hiểu các lầm tưởng và lỗi của DM