Nhập môn Học máy và Khai phá dữ liệu

(IT3190)

Nguyễn Nhật Quang

quang.nguyennhat@hust.edu.vn

Trường Đại học Bách Khoa Hà Nội

Viện Công nghệ thông tin và truyền thông

Năm học 2020-2021

Nội dung môn học:

- Giới thiệu về Học máy và Khai phá dữ liệu
- **Tiền xử lý dữ liệu**
- Đánh giá hiệu năng của hệ thống
- Hồi quy
- Phân cụm
- Phân lớp
- Phát hiện luật kết hợp

Tập dữ liệu

- Một tập dữ liệu (dataset) là một tập hợp các đối tượng (objects) và các thuộc tính của chúng
- Mỗi thuộc tính (attribute) mô tả một đặc điểm của một đối tượng
  - Vd: Các thuộc tính *Refund*, *Marital Status*, *Taxable Income*, *Cheat*
- Một tập các giá trị của các thuộc tính mô tả một đối tượng
  - Khái niệm "đối tượng" còn được tham chiếu đến với các tên gọi khác: bản ghi (record), điểm dữ liệu (data point), trường hợp (case), mẫu (sample), thực thể (entity), hoặc ví dụ (instance)

Các thuộc tính

Các đối tượng

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(Tan, Steinbach, Kumar - Introduction to Data Mining)

Các kiểu tập dữ liệu

- Bản ghi (Record)
  - Các bản ghi trong csdl quan hệ
  - Ma trận dữ liệu
  - Biểu diễn văn bản (document)
  - Dữ liệu giao dịch
- Đồ thị (Graph)
  - World Wide Web
  - Mạng thông tin, hoặc mạng xã hội
  - Các cấu trúc phân tử (Molecular structures)
- Có trật tự (Ordered)
  - Dữ liệu không gian (vd: bản đồ)
  - Dữ liệu thời gian (vd: time-series data)
  - Dữ liệu chuỗi (vd: chuỗi giao dịch)
  - Dữ liệu chuỗi di truyền (genetic sequence data)

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

(Han, Kamber - Data Mining: Concepts and Techniques)

Các kiểu giá trị thuộc tính

- Kiểu định danh/chuỗi (norminal): không có thứ tự
  - Lấy giá trị từ một tập không có thứ tự các giá trị (định danh)
  - Vd: Các thuộc tính như: Name, Profession, …
- Kiểu nhị phân (binary): là một trường hợp đặc biệt của kiểu định danh
  - Tập các giá trị chỉ gồm có 2 giá trị (Y/N, 0/1, T/F)
- Kiểu có thứ tự (ordinal):
  - Lấy giá trị từ một tập có thứ tự các giá trị
  - Vd1: Các thuộc tính lấy giá trị số như: Age, Height,…
  - Vd2: Thuộc tính Income lấy giá trị từ tập {low, medium, high}

Kiểu thuộc tính rời rạc vs. liên tục

- Kiểu thuộc tính rời rạc (Discrete-valued attributes)
  - Tập các giá trị là một tập hữu hạn
  - Bao gồm cả các thuộc tính có kiểu giá trị là các số nguyên
  - Bao gồm cả các thuộc tính nhị phân (binary attributes)
- Kiểu thuộc tính liên tục (Continuous-valued attributes)
  - Các giá trị là các số thực (real numbers)

Các đặc tính mô tả dữ liệu

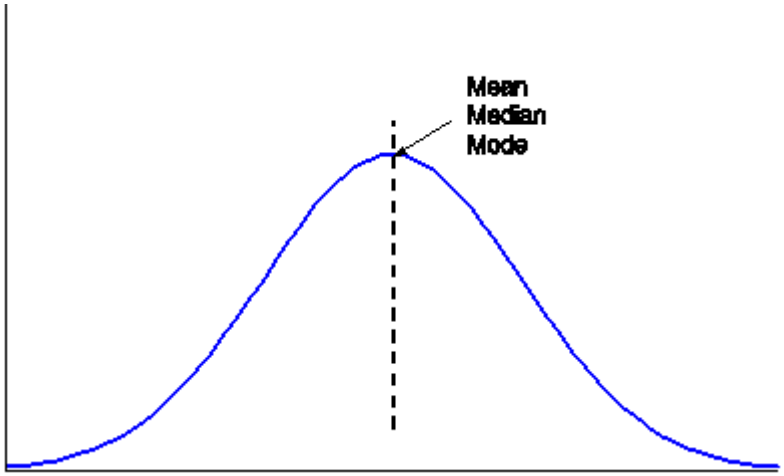
- Mục đích: Để hiểu rõ về dữ liệu có được (chiều hướng chính/trung tâm, sự biến thiên, sự phân bố)
- Sự phân bố của dữ liệu (Data dispersion)
  - Giá trị cực tiểu/cực đại (min/max)
  - Giá trị xuất hiện nhiều nhất (mode)
  - Giá trị trung bình (mean)
  - Giá trị trung vị (median)
  - Sự biến thiên (variance) và độ lệch chuẩn (standard deviation)
  - Các ngoại lai (outliers)

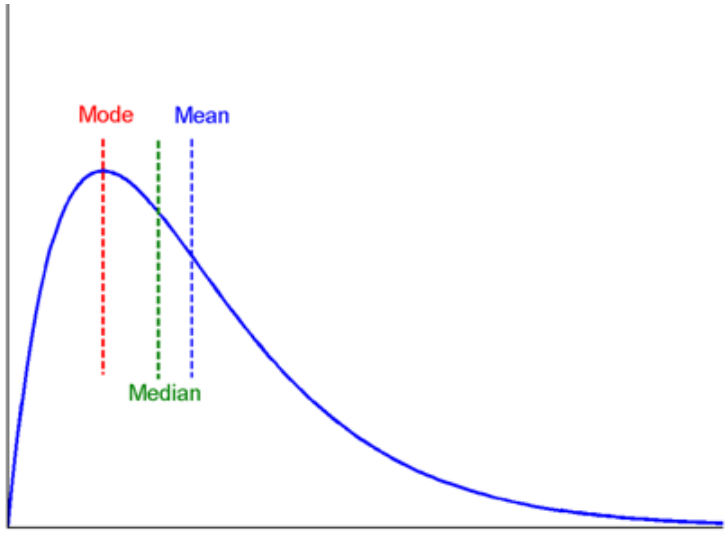
Hiển thị hóa dữ liệu (Data visualization)

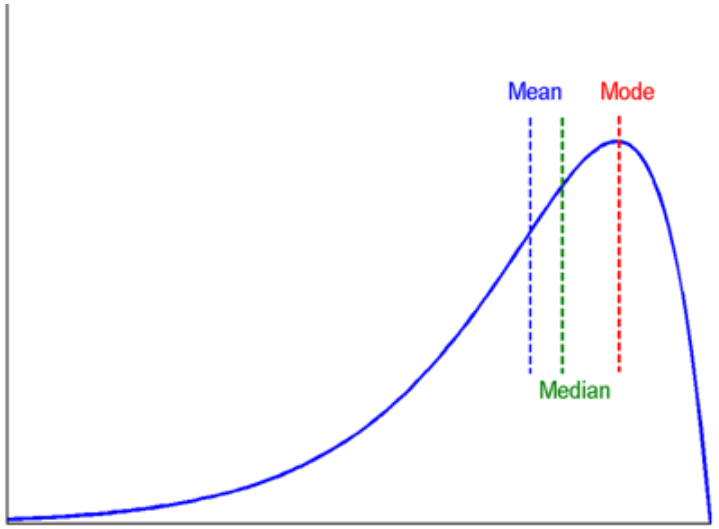
- Biểu diễn dữ liệu bằng các phương pháp hiển thị đồ họa, giúp hiểu rõ các đặc điểm của dữ liệu
- Cung cấp cái nhìn định tính đối với các tập dữ liệu lớn
- Có thể chỉ ra các mẫu, các xu hướng, các cấu trúc, các bất thường, và các quan hệ trong dữ liệu
- Hỗ trợ xác định các vùng dữ liệu quan trọng và các tham số phù hợp cho các phân tích định lượng tiếp theo
- Trong một số trường hợp, có thể cung cấp các chứng minh trực quan đối với các biểu diễn (tri thức) thu được

Dữ liệu cân đối vs. lệch

- Giá trị trung bình, giá trị trung vị, và giá trị xuất hiện nhiều nhất đối với
  - Dữ liệu cân đối
  - Dữ liệu lệch

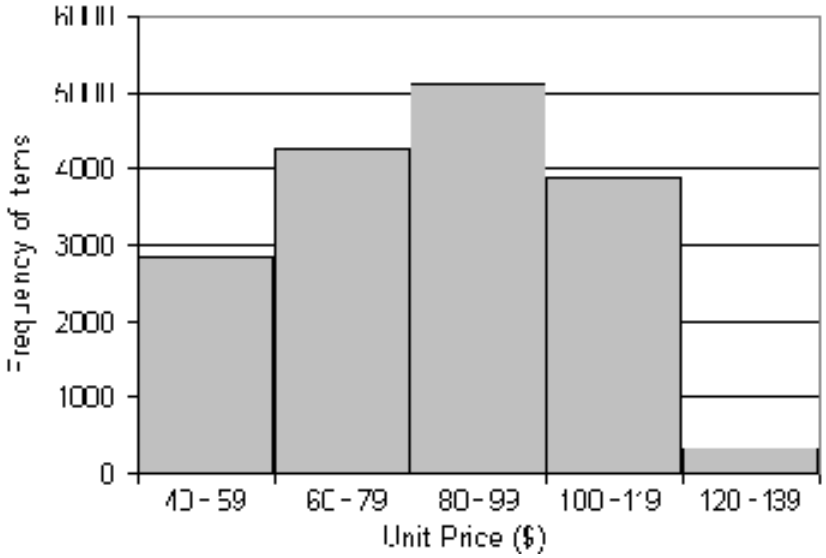






Biểu đồ histogram

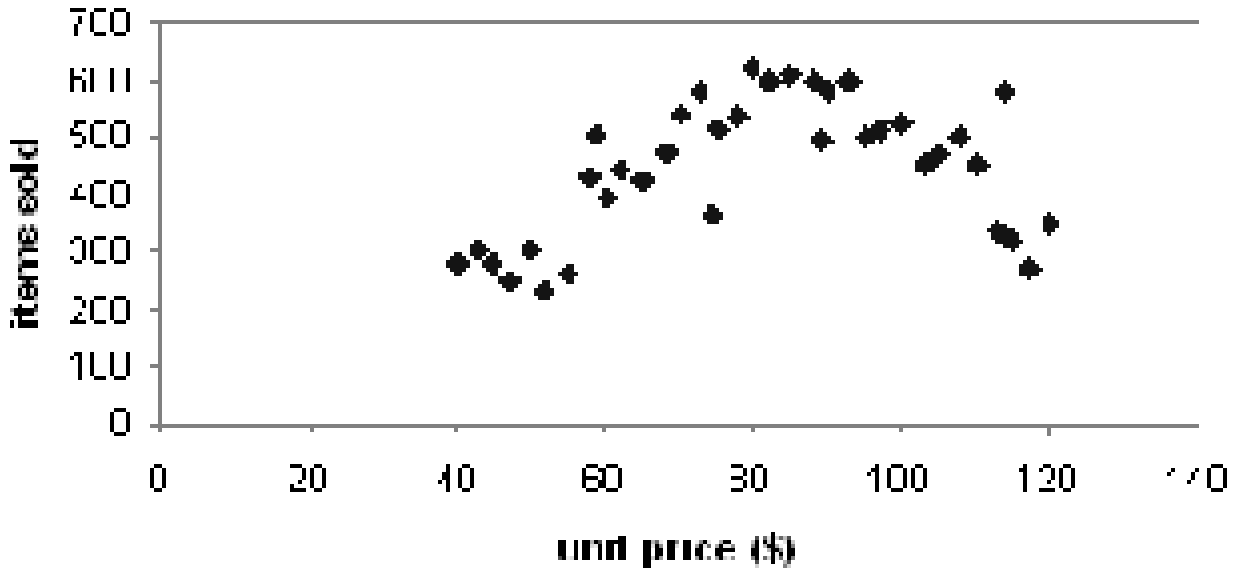
- Biểu đồ histogram là cách biểu diễn dựa trên đồ thị
- Được sử dụng rất phổ biến
- Hiển thị các mô tả thống kê xuất hiện (counts/frequencies) **theo một thuộc tính** nào đó



(Han, Kamber - Data Mining: Concepts and Techniques)

Đồ thị rải rác (Scatter plot)

- Cho phép hiển thị **quan hệ 2 chiều (giữa 2 thuộc tính)** của dữ liệu
- Cho phép quan sát (trực quan) các nhóm điểm, các ngoại lại,…
- Mỗi cặp giá trị của 2 thuộc tính được xét tương ứng với 2 tọa độ của điểm được hiển thị trên mặt phẳng



(Han, Kamber - Data Mining: Concepts and Techniques)

Tiền xử lý dữ liệu: Các nhiệm vụ chính

- **Làm sạch dữ liệu (Data cleaning)**
  - Gán các giá trị thuộc tính còn thiếu, Sửa chữa các dữ liệu nhiễu/lỗi, Xác định hoặc loại bỏ các ngoại lai (outliers), Giải quyết các mâu thuẫn dữ liệu
- **Tích hợp dữ liệu (Data integration)**
  - Tích hợp nhiều cơ sở dữ liệu, nhiều khối dữ liệu (data cubes), hoặc nhiều tập tin dữ liệu
- **Biến đổi dữ liệu (Data transformation)**
  - Chuẩn hóa (normalize) và kết hợp (aggregate) dữ liệu
- **Giảm bớt dữ liệu (Data reduction)**
  - Giảm bớt về biểu diễn (các thuộc tính) của dữ liệu, giảm bớt kích thước dữ liệu – nhưng vẫn đảm bảo thu được các kết quả khai phá dữ liệu tương đương (hoặc xấp xỉ)
  - **Rời rạc hóa dữ liệu (Data discretization)**
    - Là một thao tác trong giảm bớt dữ liệu
    - Được sử dụng đối với các dữ liệu có các thuộc tính kiểu số

Làm sạch dữ liệu (1)

- Các vấn đề của dữ liệu?
- Dữ liệu thu được từ thực tế có thể chứa nhiễu, lỗi, không hoàn chỉnh, có mâu thuẫn
  - **Không hoàn chỉnh (incomplete)**: Thiếu các giá trị thuộc tính, hoặc thiếu một số thuộc tính
    - Vd: salary = <undefined>
  - **Nhiễu/lỗi (noise/error)**: Chứa đựng những lỗi hoặc các ví dụ bất thường (abnormal instances)
    - Vd: salary = "-525" (giá trị của thuộc tính không thể là một số âm)
  - **Mâu thuẫn (inconsistent)**: Chứa đựng các mâu thuẫn (không thống nhất)
    - Vd: salary = "abc" (không phù hợp với kiểu dữ liệu số của thuộc tính salary)

Làm sạch dữ liệu (2)

- Nguồn gốc/lý do của dữ liệu không sạch?
- **Không hoàn chỉnh (incomplete)**
  - Giá trị của thuộc tính không có (not available) tại thời điểm được thu thập
  - Các vấn đề gây ra bởi phần cứng, phần mềm, hoặc người thu thập dữ liệu
- **Nhiễu/lỗi (noise/error)**
  - Do việc thu thập dữ liệu
  - Do việc nhập dữ liệu
  - Do việc truyền dữ liệu
- **Mâu thuẫn (inconsistent)**
  - Dữ liệu được thu thập từ nhiều nguồn khác nhau
  - Vi phạm các ràng buộc (điều kiện) đối với các thuộc tính

Làm sạch dữ liệu (3)

- Tại sao cần phải làm sạch dữ liệu?
- Nếu dữ liệu không sạch (có chứa lỗi, nhiễu, không đầy đủ, có mâu thuẫn), thì các kết quả khai phá dữ liệu sẽ bị ảnh hưởng và không đáng tin cậy
- Các kết quả khai phá dữ liệu (các tri thức khám phá được) không chính xác (không đáng tin cậy) sẽ dẫn đến các quyết định không chính xác, không tối ưu
  - Vd: Các dữ liệu chứa lỗi hoặc thiếu giá trị thuộc tính sẽ có thể dẫn đến các kết quả thống kê sai lầm

Thiếu giá trị thuộc tính

- Đối với một số thuộc tính, giá trị của chúng đối với một số bản ghi không có
  - Vd: Giá trị của thuộc tính Income không có (không được ghi lại) đối với một số bản ghi
- Thiếu giá trị thuộc tính có thể vì:
  - Lỗi của các thiết bị phần cứng
  - Không tương thích với các dữ liệu đã được ghi từ trước, do đó giá trị (mới) bị xóa đi
  - Dữ liệu không được nhập vào (lỗi của người nhập liệu)
- Các giá trị thuộc tính thiếu cần phải được gán (bằng một cơ chế suy diễn) – để đảm bảo tính chính xác của các kết quả khai phá dữ liệu

Thuộc tính thiếu giá trị: Các giải pháp

- Bỏ qua các bản ghi có các thuộc tính thiếu giá trị
  - Thường được áp dụng trong các bài toán phân lớp (classification)
  - Không hiệu quả, khi tỷ lệ % các giá trị thiếu đối với các thuộc tính (rất) khác nhau
- Một số người sẽ đảm nhiệm việc kiểm tra và gán các giá trị thuộc tính còn thiếu này (manually filling): công việc tẻ nhạt + chi phí cao
- Gán giá trị tự động bởi máy tính
  - Một giá trị (hằng) mặc định
  - Giá trị trung bình của thuộc tính đó
  - Giá trị trung bình của thuộc tính đó, xét đối với tất cả các ví dụ (các bản ghi) thuộc cùng lớp (class) với bản ghi đó
  - Giá trị có thể xảy ra nhất – dựa trên phương pháp xác suất (vd: công thức Bayes)

Dữ liệu chứa nhiễu

- Nhiễu: Lỗi ngẫu nhiên đối với giá trị của một thuộc tính
- Các giá trị thuộc tính bị lỗi (nhiễu) có thể vì:
  - Lỗi của các thiết bị thu thập dữ liệu
  - Các lỗi khi nhập dữ liệu
  - Lỗi trong quá trình truyền dữ liệu
  - Sự mâu thuẫn (không nhất quán) trong quy ước tên (thuộc tính/biến)

Dữ liệu chứa nhiễu: Các giải pháp

- Phân khoảng (Binning)
  - Sắp xếp dữ liệu, và phân chia thành các khoảng (bins) có tần số xuất hiện giá trị (frequency) như nhau
  - Sau đó, mỗi khoảng dữ liệu có thể được biểu diễn bằng trung bình(mean), trung vị (median), hoặc các giới hạn…của các giá trị trong khoảng đó
- Hồi quy (Regression)
  - Gắn dữ liệu với một hàm hồi quy (regression function)
- Phân cụm (Clustering)
  - Phát hiện và loại bỏ các ngoại lai (sau khi đã xác định các cụm)
- Kết hợp giữa máy tính và kiểm tra của con người
  - Máy tính tự động phát hiện các giá trị nghi ngờ (là nhiễu/lỗi)
  - Các giá trị nghi ngờ này sẽ được con người kiểm tra lại

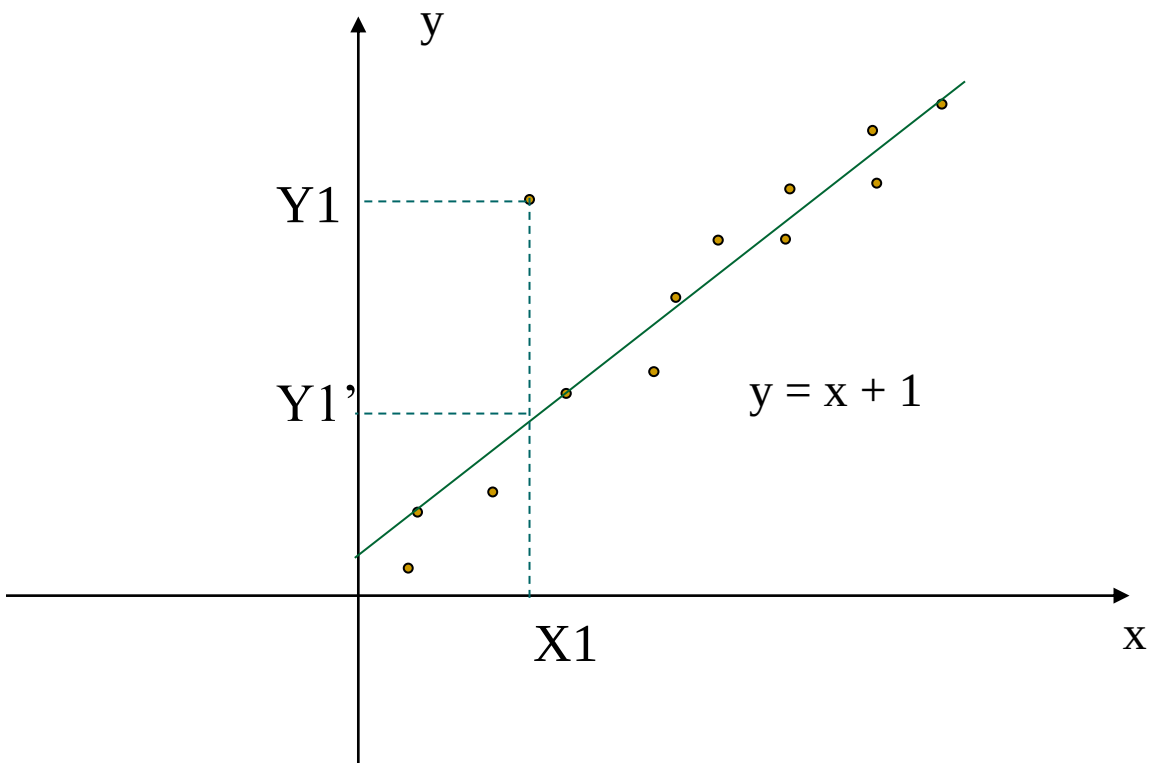
Phân khoảng (Binning)

- Phân chia với độ rộng (khoảng cách) bằng nhau
  - Chia khoảng giá trị thành N khoảng với kích thước (độ rộng) bằng nhau
  - Nếu min_i và max_i là giá trị lớn nhất và nhỏ nhất của thuộc tính, thì kích thước (độ rộng) của mỗi khoảng = $(max_i - min_i)/N$
  - Không phù hợp đối với các tập dữ liệu lệch (skewed data), hoặc có chứa các ngoại lai (outliers) – vì có thể một khoảng sẽ chỉ chứa một (hoặc một số) các ngoại lai
- Phân chia với độ sâu (tần suất xuất hiện) bằng nhau
  - Chia khoảng giá trị thành N khoảng (không nhất thiết bằng nhau), sao cho mỗi khoảng chứa xấp xỉ bằng nhau số lượng (tần xuất xuất hiện) của các ví dụ
  - Hiệu quả hơn cách phân chia với độ rộng (khoảng cách) bằng nhau

Phân khoảng (Binning) – Ví dụ

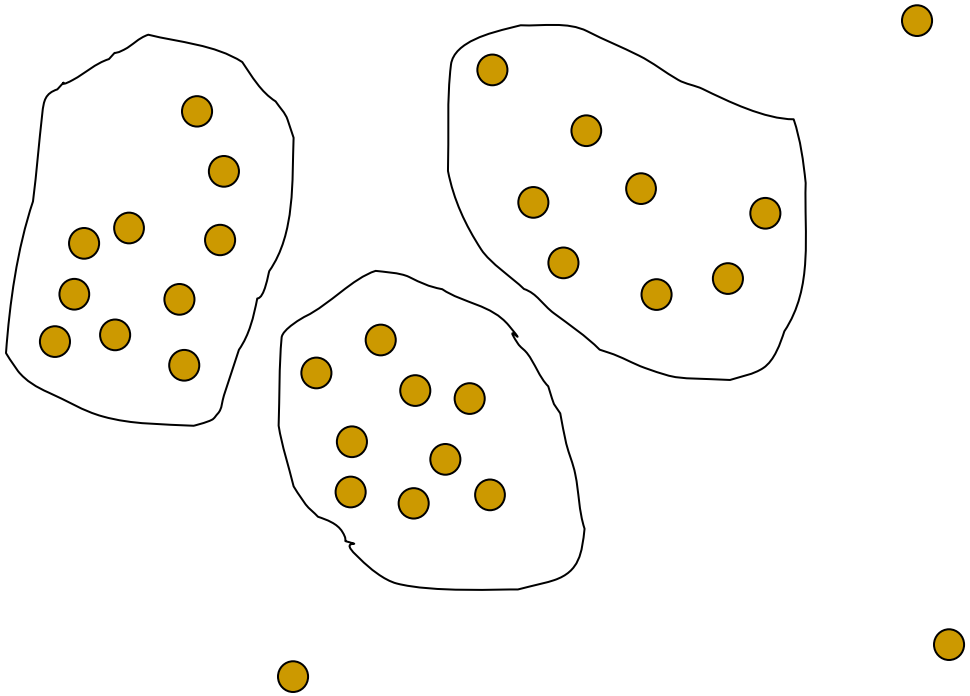
- Sắp xếp các giá trị của thuộc tính *Price*: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Phân chia thành các khoảng với độ sâu (tần xuất xuất hiện) bằng nhau
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Biểu diễn khoảng dữ liệu bởi giá trị trung bình
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

Hồi quy (Regression)



(Han, Kamber - Data Mining: Concepts and Techniques)

Phân tích các cụm (Cluster analysis)



(Han, Kamber - Data Mining: Concepts and Techniques)

Tích hợp dữ liệu

- Tích hợp dữ liệu (Data integration)
  - Kết hợp dữ liệu từ nhiều nguồn vào một kho dữ liệu thống nhất
- Tích hợp ở mức mô hình (Schema integration)
  - Tích hợp metadata từ các nguồn khác nhau
  - Vd: A.cust-id ≡ B.customID
- Vấn đề xác định thực thể (để tránh dư thừa dữ liệu)
  - Cần xác định các thực thể (identities) trên thực tế từ nhiều nguồn dữ liệu
  - Vd: Bill Clinton ≡ B. Clinton
- Phát hiện và xử lý các mâu thuẫn đối với giá trị dữ liệu
  - Đối với cùng một thực thể trên thực tế, nhưng các giá trị thuộc tính từ nhiều nguồn khác nhau lại khác nhau. Các lý do có thể:
    - Các cách biểu diễn khác nhau
    - Mức đánh giá, độ đo (scales) khác nhau – Vd: hệ đo lường mét vs. hệ đo lường của Anh

Tích hợp dữ liệu: Xử lý dư thừa dữ liệu

- Dư thừa dữ liệu (redundant data) thường xuyên xảy ra, khi tích hợp dữ liệu từ nhiều nguồn (vd: từ nhiều csdl)
  - Định danh đối tượng: Cùng một thuộc tính (hay cùng một đối tượng) có thể mang các tên (định danh) khác nhau trong các csdl khác nhau
  - Dữ liệu suy ra được: Một thuộc tính trong một bảng có thể là một thuộc tính được suy ra (derived attribute) trong một bảng khác – Vd: "Annual Revenue" và "Monthly Revenue"
- Các thuộc tính dư thừa có thể được phát hiện bằng phân tích tương quan (Correlation analysis): Pearson, Cosine, chi-square
- Yêu cầu chung đối với quá trình tích hợp dữ liệu: Giảm thiểu (tránh được là tốt nhất) các dư thừa và các mâu thuẫn
  - Giúp cải thiện tốc độ của quá trình khai phá dữ liệu, và nâng cao chất lượng của các kết quả (tri thức) thu được

Biến đổi dữ liệu (1)

- Biến đổi dữ liệu (Data transformation)
  - Việc chuyển (ánh xạ) toàn bộ tập giá trị của một thuộc tính sang một tập mới các giá trị thay thế, sao cho mỗi giá trị cũ tương ứng với một trong các giá trị mới
- Các phương pháp biến đổi dữ liệu
  - Làm trơn (Smoothing): Loại bỏ nhiễu/lỗi khỏi dữ liệu
  - Kết hợp (Aggregation): Sự tóm tắt dữ liệu, xây dựng các khối dữ liệu (data cubes)
  - Khái quát hóa (Generalization): Xây dựng các phân cấp khái niệm (concept hierarchies)
  - Chuẩn hóa (Normalization): Đưa các giá trị về một khoảng được chỉ định
    - Chuẩn hóa min-max
    - Chuẩn hóa z-score
    - Chuẩn hóa bởi thang chia 10
  - Xây dựng (tạo nên) các thuộc tính mới dựa trên các thuộc tính ban đầu

Biến đổi dữ liệu (2)

- Chuẩn hóa min-max: thành khoảng [new_min_i, new_max_i]

$$v^{new} = \frac{v^{old} - min_i}{max_i - min_i}(new\_max_i - new\_min_i) + new\_min_i$$

- Chuẩn hóa z-score
  - μ_i, σ_i: giá trị trung bình và độ lệch chuẩn đối với thuộc tính *i*

$$v^{new} = \frac{v^{old} - \mu_i}{\sigma_i}$$

- Chuẩn hóa bởi thang chia 10

$$v^{new} = \frac{v^{old}}{10^j}$$

  - *j* là giá trị số nguyên nhỏ nhất sao cho: $\max(\{v^{new}\}) < 1$

Giảm bớt dữ liệu

- Tại sao cần phải giảm bớt dữ liệu?
  - Một kho (tập) dữ liệu lớn có thể chứa lượng dữ liệu lên đến terabytes
  - Do đó, quá trình khai phá dữ liệu có thể sẽ chạy rất lâu (rất mất thời gian) đối với toàn bộ tập dữ liệu
- Giảm bớt dữ liệu (Data reduction)
  - Để thu được một biểu diễn thu gọn (giảm bớt); nhưng vẫn sinh ra cùng (hoặc xấp xỉ) các kết quả phân tích (khai phá) như với tập dữ liệu ban đầu
- Các chiến lược giảm bớt dữ liệu
  - **Giảm số chiều (Dimensionality reduction)**: Loại bỏ bớt các thuộc tính không (ít) quan trọng
  - **Giảm lượng dữ liệu (Data/Numerosity reduction)**
    - Kết hợp khối dữ liệu (Data cube aggregation)
    - Nén dữ liệu (Data compression)
    - Hồi quy (Regression)
    - Rời rạc hóa (Discretization)

Giảm số chiều

- Ảnh hưởng tiêu cực của số chiều (số thuộc tính) lớn
  - Khi số chiều tăng, dữ liệu trở nên thưa thớt hơn (more sparse)
  - Mật độ và khoảng cách giữa các điểm (quan trọng đối với việc phân cụm, phát hiện ngoại lai) trở nên ít có ý nghĩa
- Giảm số chiều (Dimensionality reduction) giúp:
  - Tránh (giảm bớt) ảnh hưởng tiêu cực của số chiều lớn
  - Loại bỏ các thuộc tính không liên quan, và giảm nhiễu/lỗi
  - Giảm chi phí về thời gian và bộ nhớ cần cho quá trình khai phá dữ liệu
  - Cho phép hiển thị hóa (visualize) dữ liệu một cách dễ dàng và hiệu quả hơn
- Một số ví dụ điển hình về kỹ thuật giảm số chiều:
  - Phân tích thành phần chính (Principal component analysis)
  - Lựa chọn tập con các thuộc tính (Feature subset selection)

Lấy mẫu dữ liệu

- Lấy mẫu dữ liệu (Data sampling) là phương pháp quan trọng đối với việc lựa chọn dữ liệu
- Việc lấy mẫu dữ liệu là cần thiết vì yêu cầu *thu thập* và *xử lý* toàn bộ một tập dữ liệu lớn sẽ đòi hỏi chi phí cao và tốn thời gian
- Các nguyên tắc quan trọng của việc lấy mẫu dữ liệu
  - Sử dụng một mẫu (sample) sẽ có tác dụng gần như sử dụng toàn bộ tập dữ liệu, nếu như mẫu đó đại diện cho tập dữ liệu
  - Một mẫu được gọi là đại diện cho một tập dữ liệu, nếu mẫu đó có (xấp xỉ) đặc tính của tập dữ liệu

Các phương pháp lấy mẫu dữ liệu

- Lấy mẫu ngẫu nhiên (Simple random sampling)
  - Mỗi ví dụ (bản ghi) được lựa chọn với một giá trị xác suất như nhau
- Lấy mẫu không thay thế (Sampling without replacement)
  - Khi một ví dụ (bản ghi) được lấy mẫu, nó sẽ được loại khỏi tập dữ liệu ban đầu (sẽ không thể được chọn thêm một lần nào nữa)
- Lấy mẫu có thay thế (Samping with replacement)
  - Khi một ví dụ (bản ghi) được lấy mẫu, nó không bị loại khỏi tập dữ liệu ban đầu (có thể được chọn nhiều hơn một lần)
- Lấy mẫu phân tầng (Stratified sampling)
  - Phân chia tập dữ liệu thành các phần (partitions)
  - Lấy ngẫu nhiên các ví dụ từ mỗi phần