# Chương 1: Tổng quan về KTDL

# Chương 2: Tiền xử lý dữ liệu

## Một số khái niệm

* Attribute:

Nominal: Danh mục, trạng thái, name of things, Binary: Attribute have 2 status,Ordinal: Thứ tự ó ý nghĩa,Numeric: Định lượng, có thể đo được

* Thuộc tính rời rạc, liên tục:

Chỉ có một bộ giá trị hữu hạn hoặc vô hạn đếm được,Đôi khi được biểu diễn dưới dạng số nguyên,Binary là trường hợp đặc biệt của rời rạc,EX: Zip Code, giới tính, nghề nghiệp, Giá trị là số thực,Trên thực tế, các giá trị thực chỉ có thể được đo lường và biểu diễn bằng số lượng chữ số hữu hạn,Thường được biểu diễn dưới dạng các biến dấu phẩy động, EX: Nhiệt độ, chiều cao, cân nặng

## Làm sạch dữ liệu

Dữ liệu nhiễu, ngoại lai: Loại bỏ

Dữ liệu mâu thuẫn: Chỉnh sửa

Dữ liệu thiếu, chưa đầy đủ:

* + Điền thiếu bằng tay
  + Điền thiếu tự động
    - Hằng số chung, 1 lớp mới EX: unknown
    - GTrị TB của thuộc tính
    - GTrị TB của các mẫu thuộc cùng lớp đó
    - GTrị có xác suất, khả năng lớn nhất: suy ra từ Hồi quy, Cây quyết định, Công thức suy diễn Bayesian, Giải thuật EM

# Chương 3: Tập phổ biến & Luật kết hợp

* Mẫu phổ biến: là mẫu (tập các item, chuỗi con, cấu trúc con,…) xuất hiện thường xuyên trong tập dữ liệu
* Mục đích: Tìm các hiện tượng thường xuyên xảy ra trong dữ liệu 🡺 Phân tích CSDL bán hàng, quảng cáo, chiến dịch bán hàng, DNA, … (Ứng dụng)
* Tính chất tập phổ biến: Tất cả tập con của tập phổ biến đều là tập phổ biến
* Tập phổ biến tối đại: Là tập phổ biến không có tập nào bao nó
* Tập phổ biến đóng: Là tập phổ biến không có tập nào bao nó cùng độ phổ biến với nó

## Apriori

* Thách thức:
  + Phải duyệt CSDL nhiều lần
  + Số lượng tập ứng viên rất lớn
  + Thực hiện việc tính độ phố biến nhiều
  + EX: 100item 🡺 Duyệt 100 lần 🡺 Tập ứng viên 2100-1
* Cải tiến: Ý tưởng chung
  + Giảm số lần duyệt CSDL
  + Giảm số lượng tập ứng viên
  + Tính độ phổ biến thuận tiện hơn

# Chương 4: Dãy phổ biến

# Chương 5: Tập thô

## Ứng dụng

* Khắc phục hiện tượng dữ liệu bị nhiễu
* Rút gọn dữ liệu (khử dữ liệu thừa)
* Tạo luật phân lớp
* Nhận diện phụ thuộc riêng phần và toàn phần của các thuộc tính

## Nhận xét quan hệ tương đương

* Mỗi đối tượng đều thuộc 1 lớp duy nhất
* Hai đối tượng trong 1 lớp có quan hệ với nhau
* Hai đối tượng thuộc 2 lớp khác nhau không quan hệ với nhau

## Mục đích xấp xỉ tập hợp

* Chỉ ra những đối tượng có thuộc tính quyết định dương/ không dương.
* Những đối tượng nào thuộc vùng biên giữa các trường hợp chắc chắn

# Chương 6: Phân lớp dữ liệu

## Phân lớp dữ liệu

* Dự đoán nhãn lớp (discrete hoặc nominal)
* Xây dụng mô hình phân lớp dựa trên tập huấn luyện và các nhãn lớp của thuộc tính phân lớp và sử dụng mô hình đó để phân lớp cho dữ liệu mới
* Ứng dụng: Chuẩn đoán y tế, phát hiện gian lận, phân loại trang web

## So sánh độ đo chọn thuộc tính

|  |  |  |
| --- | --- | --- |
| Information Gain (Entropy) | Gain Radio | Gini Index |
| Thiên về thuộc tính có nhiều giá trị | Thiên về việc phân chia không cân bằng giữa các vùng | Thiên về thuộc tính có nhiều giá trị  Gặp vấn dề khi số lớp lớn  Có xu hướng ưu tiên cho các thử nghiệm tạo ra kết quả phân chia có cùng kích thước và cùng sự đồng nhất |

## Ưu nhược của Navie Bayes

|  |  |
| --- | --- |
| Ưu điểm | Nhược điểm |
| Dễ dàng cài đặt  Thời gian thi hành nhanh  Đạt kết quả tốt trong phần lớn các trường hợp | Giả thiết về tính độc lập điều kiện của thuộc tính làm giảm độ chính xác |

## Mạng neural nhân tạo

Artificial neural network(ANN) (mạng neural nhân tạo):

* Mô phỏng các hệ thần kinh sinh học (não người)
* Một cấu trúc/mạng được tạo thành từ sự kết nối của các neural nhân tạo, mỗi kết nối có một trọng số được liên kết với nó

## Ưu nhược của ANN

|  |  |
| --- | --- |
| Ưu điểm | Nhược điểm |
| Hỗ trợ tính toán song song rất cao  Đạt độ chính xác cao trong nhiều bài toán  Kiến trúc mạng linh hoạt  Khả năng thích ứng tốt với data nhiễu | Thời gian huấn luyện dài  Không có quy tắc chung để xác định cấu trúc mạng và các tham số tối ưu cho một vấn đề nhất định. Thường dựa trên exp  Khả năng diễn giả kém |

# Chương 7: Gom cụm dữ liệu

## Supervised Learning

Supervision: Dữ liệu huấn luyện (quan sát, đo lường, ...) được kèm theo nhãn lớp

Dữ liệu mới được phân lớp dựa trên tập huấn luyện (classification)

## UnSupervised Learning

Nhãn lớp của dữ liệu huấn luyện không xác định

Đưa ra một tập hợp các phép đo, quan sát, ... với mục đích thiết lập sự tồn tại của các lớp hoặc cụm trong dữ liệu (clustering)

## So sánh gom cụm truyền thống và phân lớp

|  |  |
| --- | --- |
| Gom cụm truyền thống | Phân lớp |
| Mục tiêu: xác định các cụm đối tượng tương tự. Các cụm được phát hiện  Bộ dữ liệu gồm các thuộc tính  Không giám sát (nhãn lớp phải học)  Đánh giá tính tương đồng, “hàm khoảng cách” là rất quan trọng, bởi vì các cụm được phát hiện dựa trên khoảng cách / mật độ. | Mục tiêu là dự đoán các lớp từ các thuộc tính/giá trị thuộc tính của đối tượng. Các lớp được xác định trước  Bộ dữ liệu gồm các thuộc tính và một nhãn lớp  Được giám sát (nhãn lớp đã biết)  Bộ phân lớp được học từ các tập ví dụ đã phân lớp  Bộ phân lớp cần có độ chính xác cao |

## Tiêu chuẩn gom cụm

* Tạo ra cụm có chất lượng Giữa các đối tượng trong cùng 1 cụm sự giống nhau cao (intra-class)
* Giữa các cụm sự giống nhau thấp (inter-class)
* Chất lượng của kết quả gom cụm phụ thuộc vào:
  + Độ đo sự tương tự
  + Thuật toán gom nhóm
  + Khả năng phát hiện một vài hay tất cả các mẫu bị che (hidden patterns)

## Yêu cầu và thách thức

* Khả năng mở rộng (Scalability): Phân cụm tất cả dữ liệu thay vì chỉ trên một số mẫu
* Khả năng xử lý với các loại thuộc tính khác nhau: Numerical, binary, categorical, ordinal, linked hay hỗn hợp
* Phân cụm dựa trên ràng buộc: Người dùng có thể cung cấp đầu vào về các ràng buộc
* Khả năng diễn giải và khả năng sử dụng
* Khác:
  + Khám phá các cụm với hình dạng tùy ý
  + Khả năng xử lý dữ liệu nhiễu
  + Phân cụm tăng dần và không nhạy cảm với thứ tự đầu vào
  + Số chiều lớn

## Ưu nhược

|  |  |
| --- | --- |
| Ưu điểm | Nhược điểm |
| Đơn giản, dễ hiểu, tương đối hiệu quả  Độ phức tạp: O(tkn), với n: số đối tượng, k: số cụm, t: số lần lặp (k, t << n).  So sánh với PAM: O(k(n-k)2), CLARA: O(ks2 + k(n-k))  Các đối tượng được tự động gán vào các nhóm  Thường đạt được tối ưu cục bộ | Chỉ áp dụng cho các đối tượng trong không gian n chiều liên tục.  Thuộc tính không phải dạng số?  + Sử dụng phương pháp k-mode cho dữ liệu danh mục  + k-medoids có thể được áp dụng cho nhiều loại dữ liệu  Cần xác định trước số cụm k. Có nhiều cách để xác định k tốt nhất  (Hastie et al., 2009)  Phụ thuộc vào việc khởi tạo các cụm đầu tiên  Nhạy cảm với dữ liệu nhiễu, cá biệt  Không phù hợp để khám phá các cụm có kích thước, mật độ khác  nhau hoặc có hình dạng không phải là hình cầu |

## So sánh PAM – K-Means

PAM hiệu quả hơn so với K-means với dữ liệu nhiễu, cá biệt.

PAM hiệu quả với tập dữ liệu nhỏ nhưng không mở rộng tốt với tập dữ liệu lớn

Cải tiến:

* CLARA – Clustering LARge Applications (Kaufmann & Rousseeuw, 1990): dựa trên phương pháp lấy mẫu
* CLARANS – Clustering LARge Applications based upon RANdomized Search (Ng & Han, 1994) : lấy mẫu động