KHAI PHÁ DỮ LIỆU

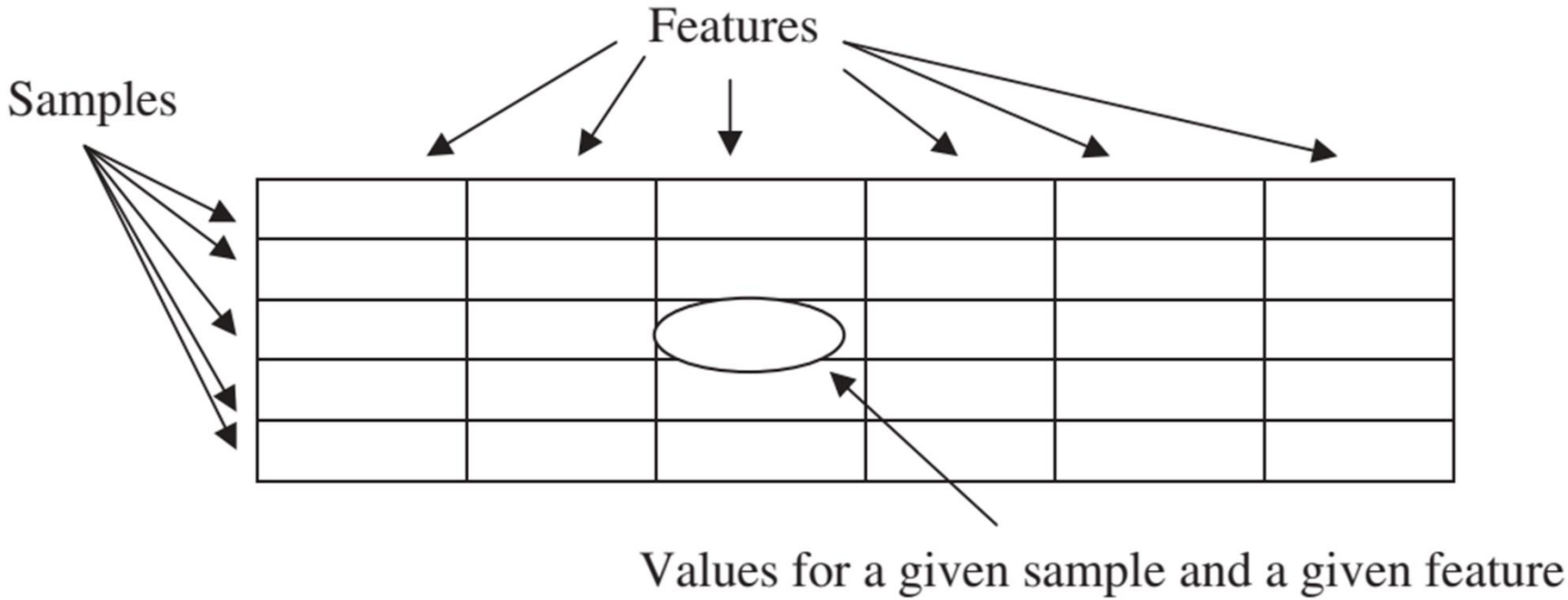
Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 2

CHUẨN BỊ DỮ LIỆU (Phần 1)

Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Dữ liệu thô:



Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Dữ liệu thô:
    - Features (đặc trưng):
      - Mỗi sample được mô tả bởi một số các features
      - Mỗi feature được biểu diễn ở nhiều dạng dữ liệu, trong đó có các dạng dữ liệu phổ biến:
        - Numeric:
          - số thực. VD: tuổi, tốc độ, chiều dài, . . .
        - Categorical:
          - Còn gọi là Symbolic
          - VD: màu mắt, giới tính, quốc tịch, …
          - Có thể chuyển đổi data kiểu categorical sang numeric

Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Dữ liệu thô:
    - Features (đặc trưng):
      - Mỗi sample được mô tả bởi một số các features
      - Mỗi feature được biểu diễn ở nhiều dạng dữ liệu, trong đó có các dạng dữ liệu phổ biến:
        - Categorical:
          - Có thể chuyển đổi data kiểu categorical sang numeric → gọi là các "dummy variable". VD:

Feature Value	Code
Black	1000
Blue	0100
Green	0010
Brown	0001

Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Dữ liệu thô:
    - Features (đặc trưng):
      - Mỗi feature được biểu diễn ở nhiều dạng dữ liệu, trong đó có các dạng dữ liệu phổ biến:
        - Continuous variables:
          - Hay Quantitative variable, hay metric variable
          - Là số nguyên hoặc số thực
        - Discrete variables:
          - Hay Qualitative variables
          - Gồm:
            - Nominal: không có thứ tự. VD: thông tin quảng cáo, nhân sự, …
            - Ordinal: có thứ tự. VD: thứ hạng của sinh viên, huy chương, …

Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Dữ liệu thô:
    - Features (đặc trưng):
      - Mỗi feature được biểu diễn ở nhiều dạng dữ liệu, trong đó có các dạng dữ liệu phổ biến:
        - Discrete variables:

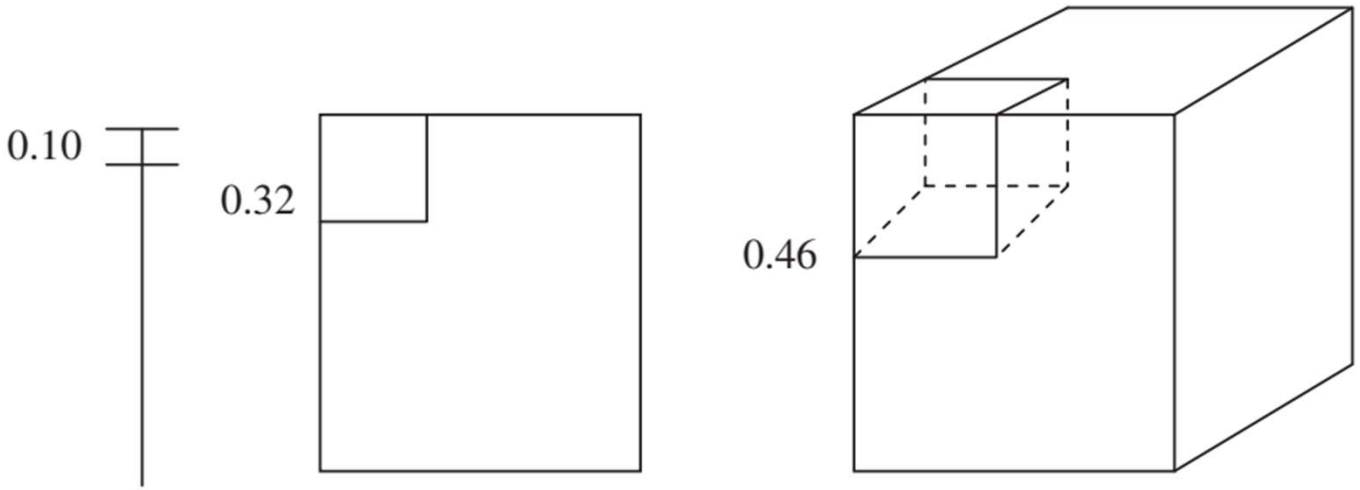
Type	Description	Examples	Operations
Nominal	Just label or different name to distinguish one object form another	Zip code, ID, gender	= or not =
Ordinal	The values provide the ordering of objects	Opinion, grades	< or >
Interval	Unit of measurement, but the origin is arbitrary	Celsius or Fahrenheit, calendar dates	+ or –
Ratio	Unit of measurement and the origin is not arbitrary	Temperature in Kelvin, length, counts, age, income	+, –, *, /

Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Dữ liệu thô:
    - Features (đặc trưng):
      - Mỗi feature được biểu diễn ở nhiều dạng dữ liệu, trong đó có các dạng dữ liệu phổ biến:
        - Static data
        - Dynamic data (Temporal data)
  - Tính chất phức tạp của dữ liệu thô:
    - Số lượng samples: rất lớn
    - Feature type: nhiều loại khác nhau

Chuẩn bị dữ liệu

- Chuẩn bị dữ liệu thô:
  - Tính chất phức tạp của dữ liệu thô:
    - High dimensional:
      - Kinh nghiêm của con người: low dimensional trong physical world. VD: 2-D, 3-D



Chuẩn bị dữ liệu

- Các đặc điểm của dữ liệu thô:
  - Số lượng dữ liệu thô dùng trong KPDL thường là lớn
  - Dữ liệu có thể có qui luật mà con người có thể nhận ra trực tiếp
  - Dữ liệu có thể có / có thể không có qui luật và rất rối rắm (messy), với dữ liệu này, cần phải tìm ra:
    - missing values
    - Distortions
    - Misrecording
    - inadequate sampling
    - . . .
  - Trong thực tế, các thông tin này rất khó tìm ra

Chuẩn bị dữ liệu

- Yêu cầu của công đoạn chuẩn bị dữ liệu:
  - Phải tổ chức dữ liệu sao cho:
    - Đạt được một dạng chuẩn nào đó
    - Sẵn sàng cho một quá trình KPDL
    - Tạo điều kiện cho các giải thuật KPDL đạt được hiệu quả và độ chính xác cao nhất
- Một số phép biến đổi dữ liệu thô:
  - Normalizations
    - Scale dữ liệu vào các khoảng giá trị hẹp nào đó: [0,1], [-1,1]
    - Nhằm giảm sự ảnh hưởng của các distance giữa các điểm dữ liệu
  - Data smoothing
    - Giảm sự khác biệt quá lớn giữa các điểm dữ liệu

Chuẩn bị dữ liệu

- Một số phép biến đổi dữ liệu thô:
  - Differences and Ratios
    - Differences:
      - Không lấy trực tiếp giá trị của data
      - Mà lấy giá trị sai khác giữa từng điểm dữ liệu với một giá trị mong muốn nào đó
    - Ratios:
      - Không lấy trực tiếp giá trị của data
      - Mà chia tất cả các điểm dữ liệu cho một giá trị đề xuất nào đó
  - Kết quả:
    - Các phép biến đổi trên có thể làm tăng hiệu quả và độ chính xác các của các thuật toán KPDL

Chuẩn bị dữ liệu

- Một số phép biến đổi dữ liệu thô:
  - Differences and Ratios
    - Differences:
      - Không lấy trực tiếp giá trị của data
      - Mà lấy giá trị sai khác giữa từng điểm dữ liệu với một giá trị mong muốn nào đó
    - Ratios:
      - Không lấy trực tiếp giá trị của data
      - Mà chia tất cả các điểm dữ liệu cho một giá trị đề xuất nào đó
  - Kết quả:
    - Các phép biến đổi trên có thể làm tăng hiệu quả và độ chính xác các của các thuật toán KPDL

Chuẩn bị dữ liệu

Q / A