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 4

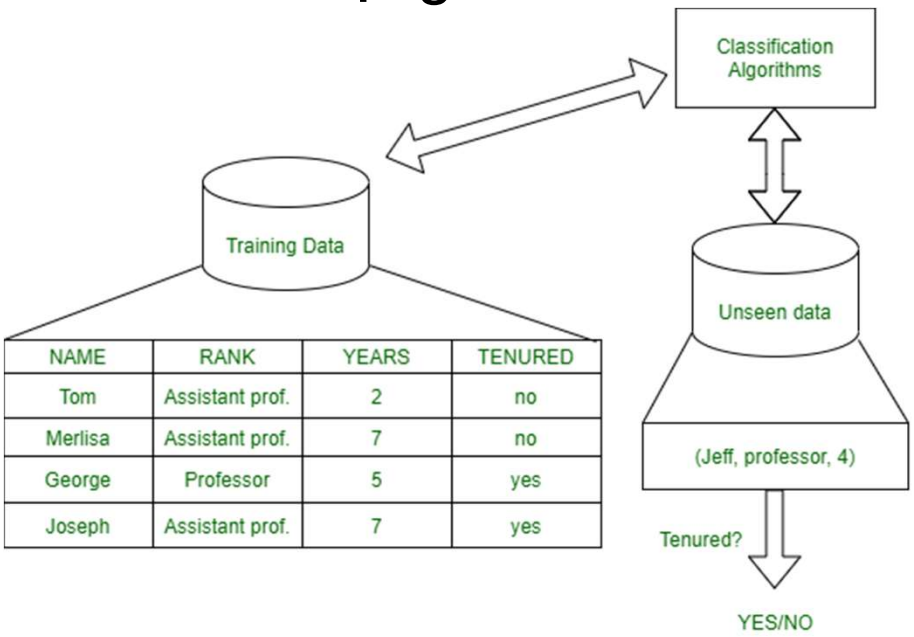
PHÂN LỚP

Phân lớp

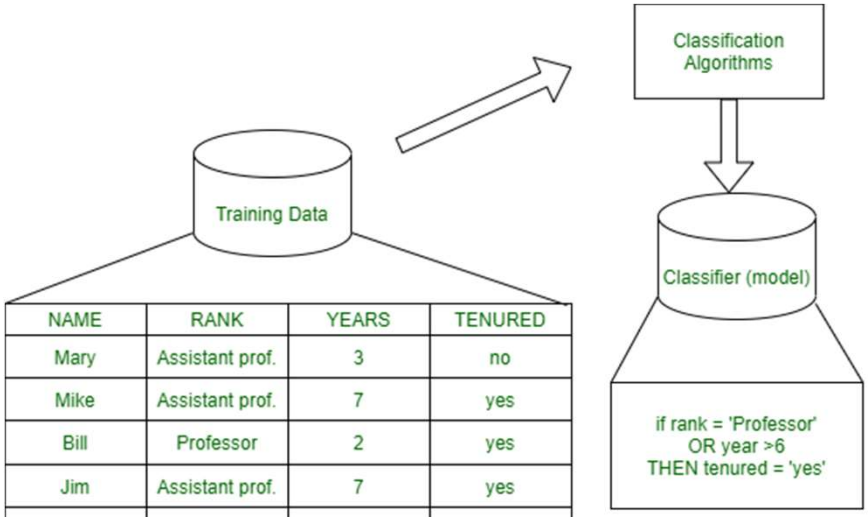
- Giới thiệu bài toán phân lớp
  - Phân lớp = classification
  - Là một trong các bài toán học theo dữ liệu (data-driven)
  - Training dataset (tập huấn luyện)
    - Tập dữ liệu được xây dựng sẵn
    - Từng dữ liệu được gán vào một lớp cụ thể cho trước → gán nhãn
  - Model (mô hình phân lớp)
    - Model = Classifier = là kết quả của quá trình huấn luyện
    - Dựa vào thông tin của tập huấn luyện → Xây dựng mô hình phân lớp sử dụng một giải thuật nào đó
  - Tác dụng
    - Từ một dữ liệu mới → dùng model (đã huấn luyện) để xác định dữ liệu đó có khả năng cao là
      - thuộc vào lớp nào trong số các lớp mà model đó đã biết
      - hoặc không thuộc lớp nào

Phân lớp

- Giới thiệu bài toán phân lớp
  - Tác dụng



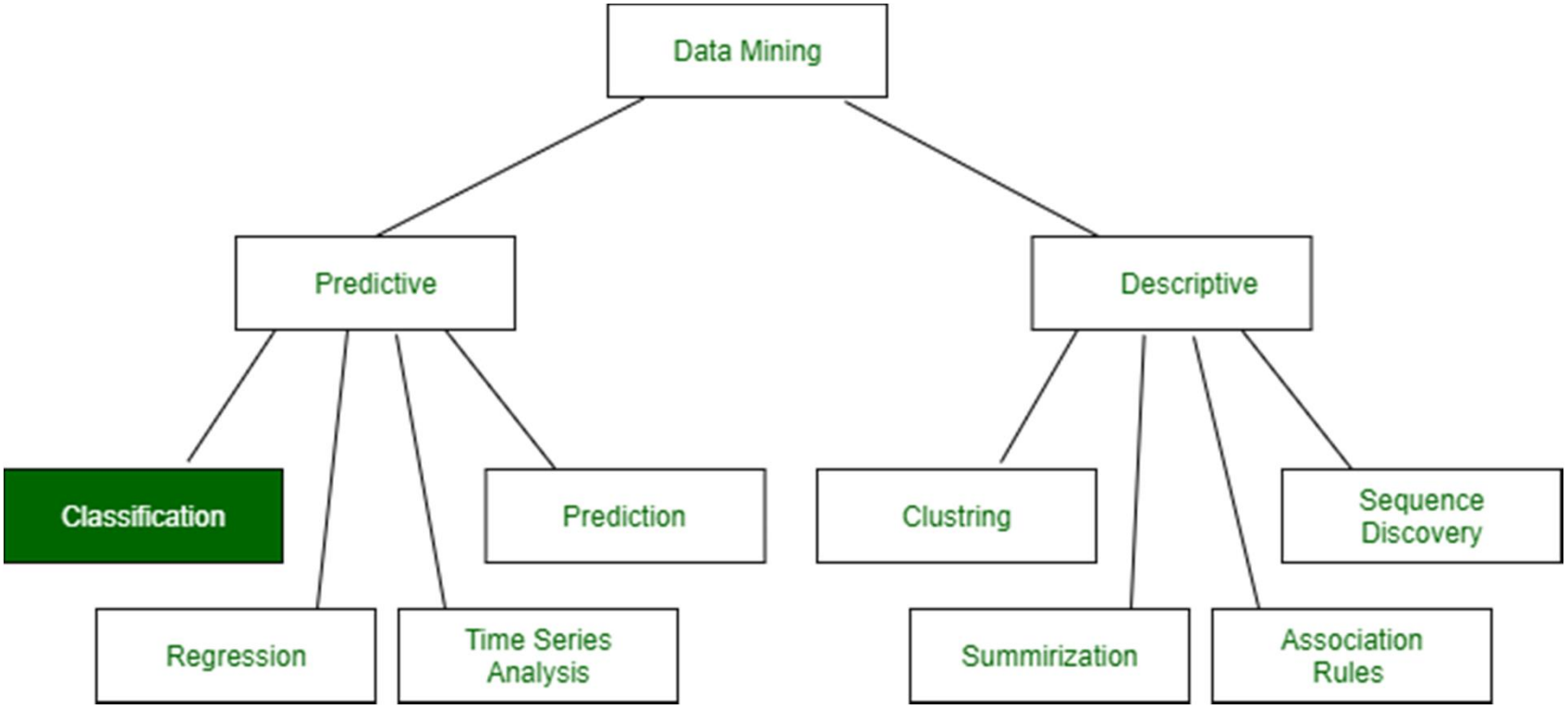
NAME	RANK	YEARS	TENURED
Tom	Assistant prof.	2	no
Merlisa	Assistant prof.	7	no
George	Professor	5	yes
Joseph	Assistant prof.	7	yes



NAME	RANK	YEARS	TENURED
Mary	Assistant prof.	3	no
Mike	Assistant prof.	7	yes
Bill	Professor	2	yes
Jim	Assistant prof.	7	yes
Dave	Assistant prof.	6	no
Anne	Assistant prof.	3	no

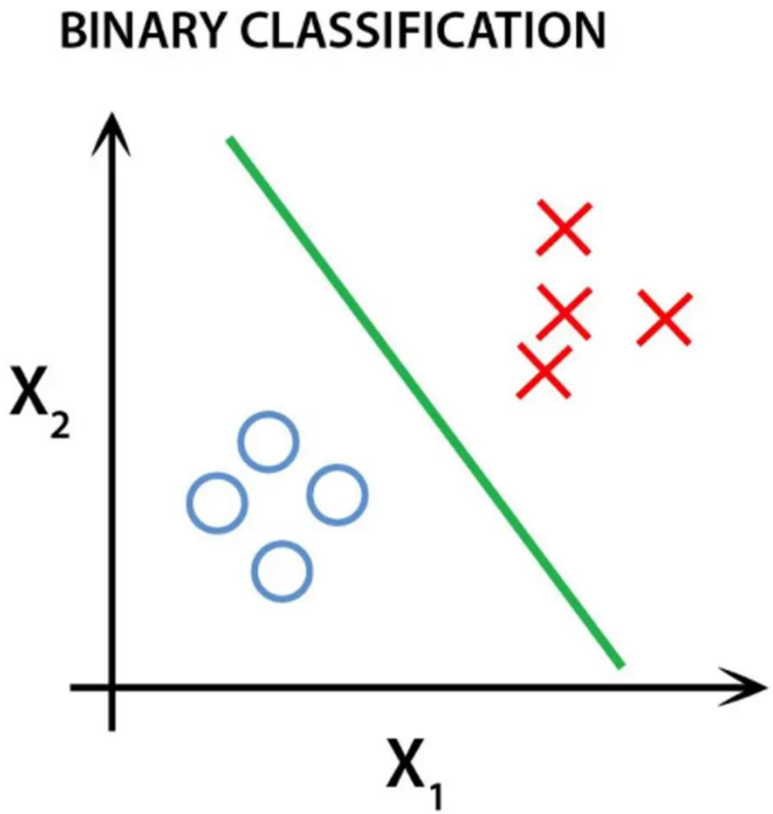
Phân lớp

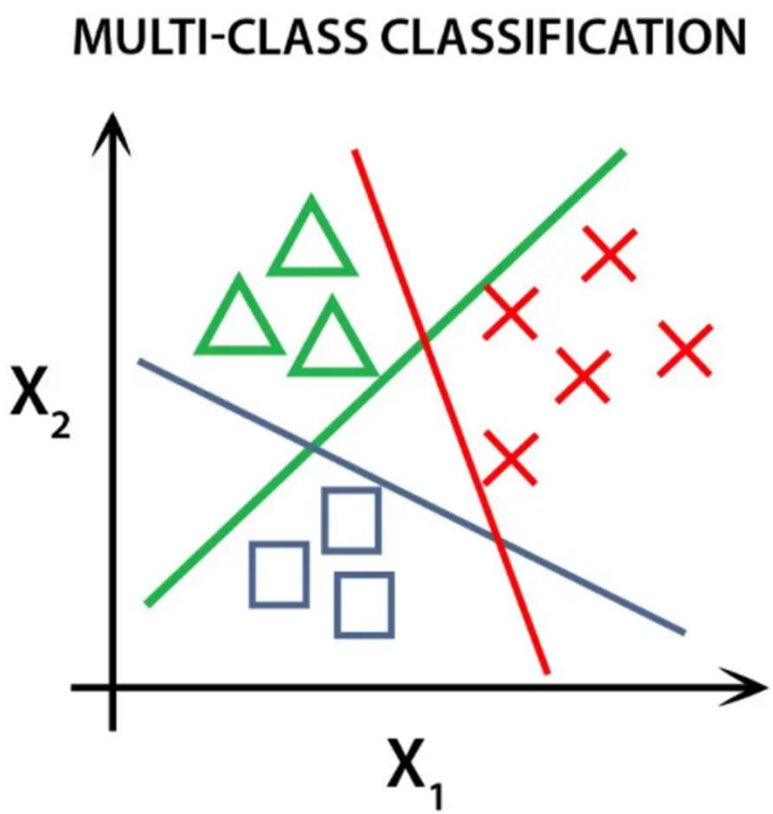
- Giới thiệu bài toán phân lớp
  - Vị trí của classification



Phân lớp

- Giới thiệu bài toán phân lớp
  - Các dạng phân lớp cơ bản



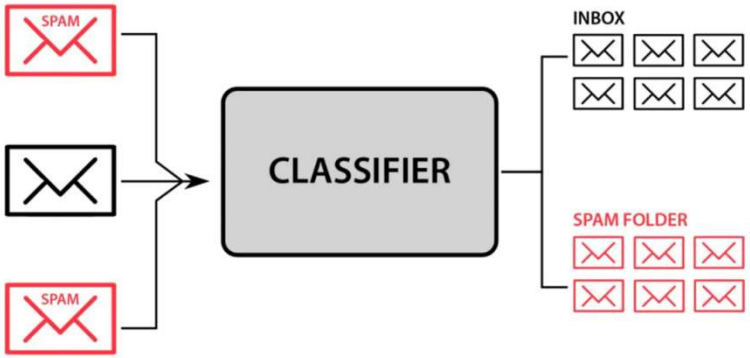


Phân lớp

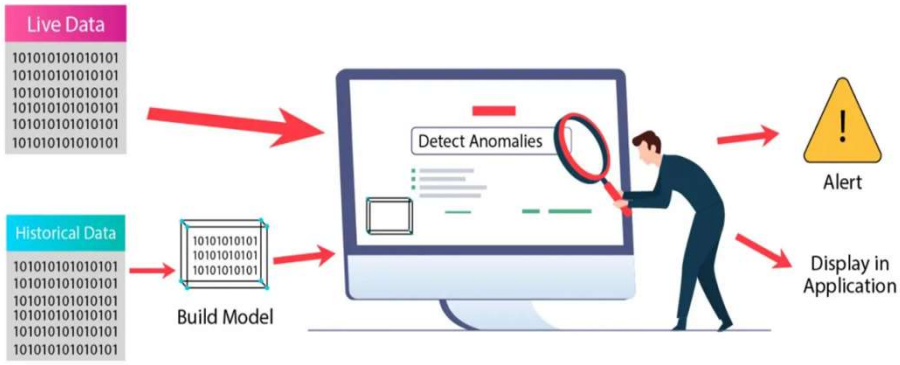
- Giới thiệu bài toán phân lớp
  - Một số giải thuật phân lớp phổ biến
    - Binary classification
      - k-Nearest Neighbors
      - Support Vector Machine
      - Decision Trees
      - Logistic Regression
      - Naive Bayes
      - . . .
    - Multi-class classification
      - k-Nearest Neighbors
      - Support Vector Machine
      - Decision Trees
      - Naive Bayes
      - Random Forest
      - Gradient Boosting
      - . . .

Phân lớp

- Giới thiệu bài toán phân lớp
  - Một số ứng dụng của bài toán phân lớp
    - Email classification



    - Anomaly / Fraud Detection



Phân lớp

- Giới thiệu bài toán phân lớp
  - Một số ứng dụng của bài toán phân lớp
    - Business data mining:
      - transaction data
    - Web mining:
      - web page classification
      - information extraction
    - Biological mining:
      - protein family classification
      - structure prediction
    - Autonomous driving
    - Speech recognition
    - Medical:
      - Based on patient records → who should be highly emergency

Phân lớp

- Giới thiệu bài toán phân lớp
  - Một số ứng dụng của bài toán phân lớp
    - Human face detection
    - Text categorization
    - Automatically collect documents of specific topics
    - Scientific Paper Header and Citation Extraction
      - Citation Index
      - Citation Database
    - DNA
      - DNA Sequence Modeling
      - DNA Database Search

Phân lớp

Q / A