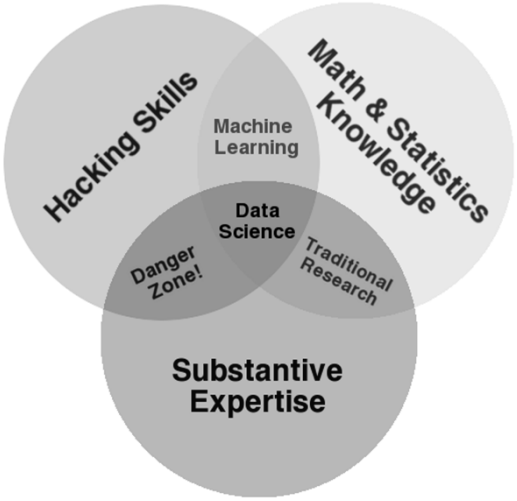
THƯ VIỆN TRONG PYTHON HỖ TRỢ PHÂN TÍCH DỮ LIỆU

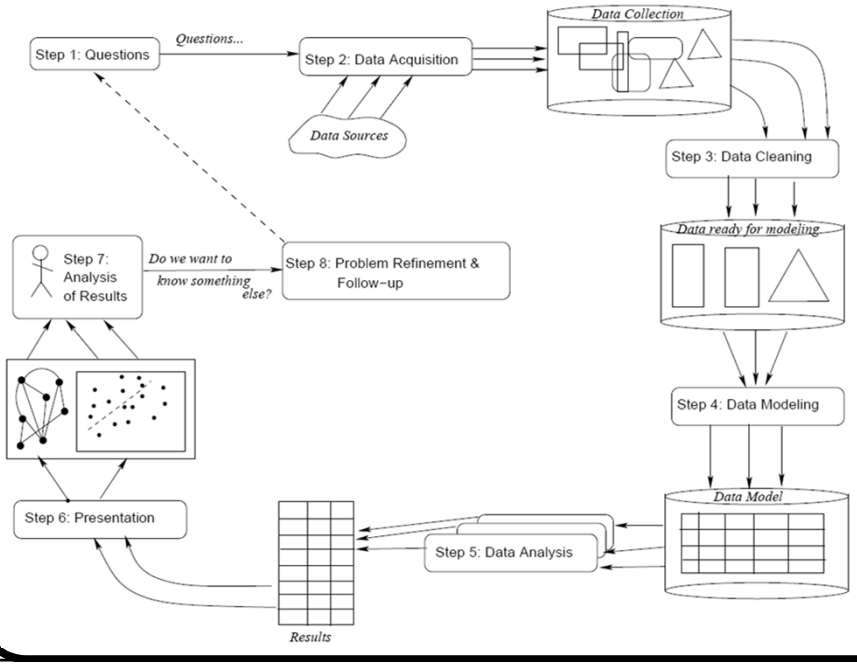
Nội dung

1. Thư viện scikit-learn

Khoa học Dữ liệu và máy Học



Quá trình xử lý của khoa học dữ liệu



Ví dụ: hệ thống phát hiện thư rác

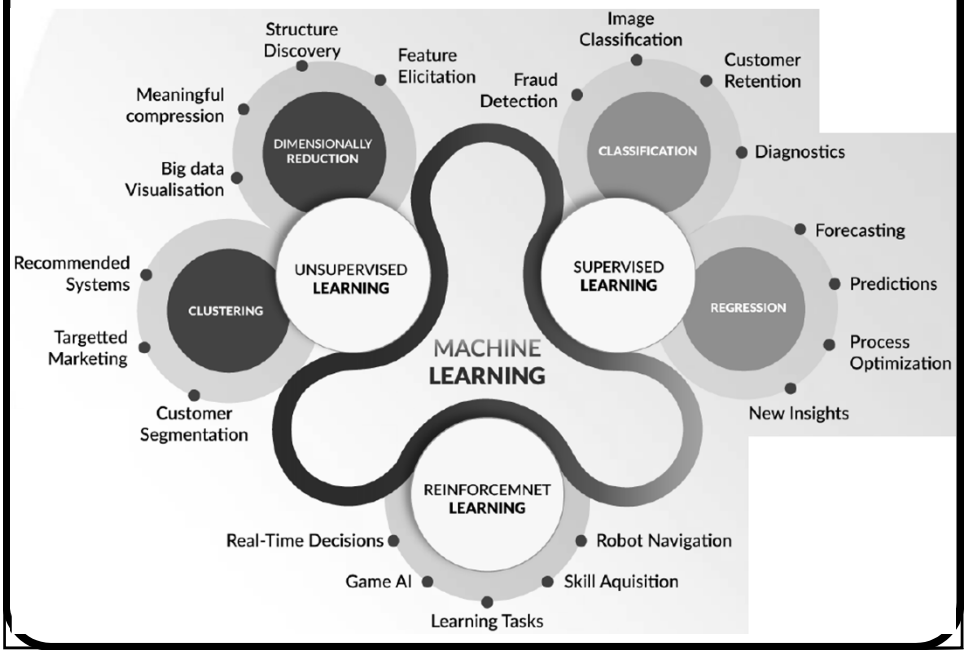
1. Thu thập mẫu thư (gồm cả thư rác và thư thường)
2. Xác định đề bài (phân lớp hay đánh giá)
3. Xử lý dữ liệu
4. Chọn mô hình học máy phù hợp với bài toán phân loại thư rác
5. Huấn luyện mô hình
6. Hiệu chỉnh, tinh chỉnh mô hình
7. Áp dụng thực tế (chạy trên email server thực)
8. Tiếp tục cập nhật theo phản hồi của người dùng

Học có giám sát (supervised learning)

Học không giám sát (unsupervised learning)

Học bán giám sát

Bài toán cơ bản



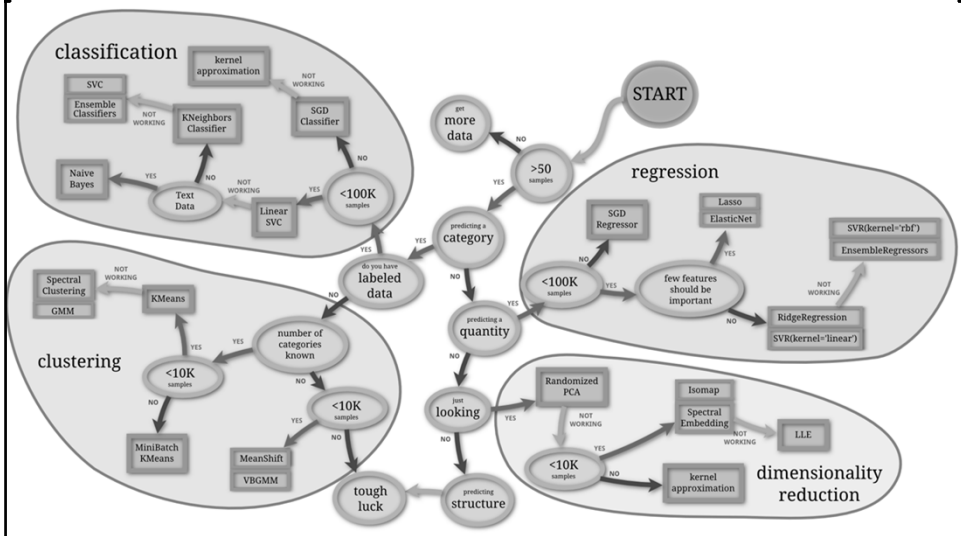
Thư viện học máy scikit-learn

Scikit-learn cung cấp gần như tất cả các loại thuật toán học máy cơ bản cùng với các kỹ thuật xử lý dữ liệu đã được chuẩn hóa

Cài đặt: **pip install scikit-learn scipy** hoặc

py- m pip install scikit-learn scipy

Chọn thuật toán học máy phù hợp



Các dạng học

❶ Học có giám sát (**Supervised learning**)

- Đầu vào là dữ liệu $\boldsymbol{X}$ và nhãn $\boldsymbol{t}$
  - $\boldsymbol{X}$: ma trận, kích thước: $N \times M$
  - $\boldsymbol{t}$: véctơ, kích thước N
  - mỗi hàng thứ n của $\boldsymbol{t}$ là nhãn của điểm dữ liệu ở hàng n trong $\boldsymbol{X}$
- Mục đích: tìm ra một mô hình $f_{\boldsymbol{w}}(\boldsymbol{x})$[1] từ tập học $<\boldsymbol{X}, \boldsymbol{t}>$ để $f_{\boldsymbol{w}}(\boldsymbol{x})$ dự báo được nhãn cho điểm dữ liệu mới $\boldsymbol{x}$ là tốt nhất[2].
- Một số dạng bài toán quan trọng:
  - Hồi quy (**Regression**)
  - Phân loại (**Classification**)
  - Định danh, nhận dạng (**Identification, Recognition**)
  - Phát hiện (**Detection**)
  - Phân đoạn (**Segmentation**)
  - Hỏi-đáp (**Question-Answering**)

[1] bài toán quy về làm tìm $\boldsymbol{w}$; $\boldsymbol{w}$: tham số của mô hình; nếu $\boldsymbol{w}$ không có phần tử nào, $f_{\boldsymbol{w}}(\boldsymbol{x})$ được gọi là mô hình phi tham số

[2] tính trên tập dữ liệu kiểm tra và sử dụng một độ đo đánh giá nào đó

❷ Học không giám sát (**Unsupervised learning**)

- Đầu vào là dữ liệu $\boldsymbol{X}$, không có nhãn đi kèm
  - $\boldsymbol{X}$: ma trận, kích thước: $N \times M$
- Mục đích: tìm ra những đặc tính quan trọng[1] trên dữ liệu đầu vào.
- Một số dạng bài toán quan trọng:
  - Gom nhóm (**Clustering**)
  - Thu giảm số chiều (**Dimensionality reduction**)
  - Trực quan hóa (**Visulization**)
  - Phát hiện bất thường (**Anormaly detection**)
  - Tìm kiếm luận kết hợp (**Association rule mining**)

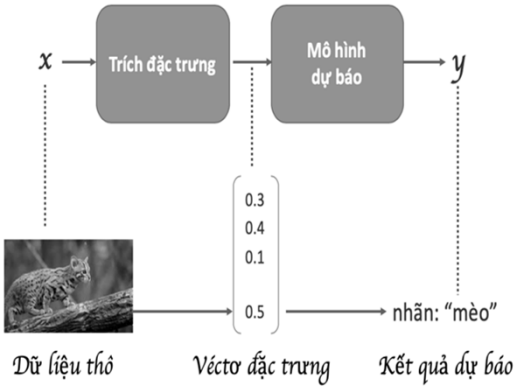
[1] tùy vào tiêu chí: sự tương quan giữa các điểm dữ liệu, tính khả tách của dữ liệu sau biến đổi, những điểm dữ liệu dị biệt so với nhóm, v.v.

❸ Học tăng cường (**Reinforcement learning**)

- Đầu vào:
  - Môi trường của bài toán: không gian trên đó có thể thực hiện các hành động;
  - Danh sách các trạng thái của bài toán: ngữ cảnh cụ thể để thực hiện hành động;
  - Danh sách các hành động: một dạng tương tác với môi trường, có thể thay đổi tình trạng của môi trường;
  - Tiêu chí để phản hồi (dạng điểm) khi đối tượng thực hiện một hành động cụ thể nào đó.
- Mục đích: Thực hiện "thử-sai" nhiều lần để khi học xong thì có thể chọn lựa hành động tối ưu trên ngữ cảnh (trạng thái) cụ thể.
  - Trò chơi (**Games**)
  - Robotics (**Người-máy, xe và thiết bị vận hành/bay tự động**)
  - Thương mại (**Trading**)

Các cách tiếp cận khi phân tích dữ liệu

Hai giai đoạn



Hình 3.1: Cách tiếp cận hai giai đoạn[1]

[1]Giai đoạn 1: rút trích đặc trưng từ dữ liệu thô; giai đoạn 2: sử dụng học máy để xây dựng mô hình dự báo

Các cách tiếp cận khi phân tích dữ liệu

Một giai đoạn

Hình 3.2: Cách tiếp cận một giai đoạn, **end-to-end** [1]

[1]Đây là cách tiếp hiện đang phổ biến, mô hình học máy là mạng nơron học sâu

Các công việc khi phân tích dữ liệu

1. Chuẩn bị dữ liệu
2. Xây dựng mô hình
3. Lựa chọn mô hình và siêu tham số
4. Triển khai ứng dụng