Học máy
Giới thiệu

Lê Thành Sách

✉ ltsach@hcmut.edu.vn

Khoa Khoa học & Kỹ thuật Máy tính
Trường Đại học Bách Khoa - ĐHQG Tp.HCM

Tp.HCM. Ngày 1 tháng 9 năm 2019

Mục lục

ltsach@hcmut.edu.vn
Lê Thành Sách

Khái niệm

ltsach@hcmut.edu.vn
Lê Thành Sách

Khái niệm

Học máy

- **Học máy:**
  - là lĩnh vực nghiên cứu giúp cho máy tính học[1] để đưa ra cách thực hiện một công việc nào đó thay cho việc yêu cầu máy tính giải quyết công việc qua dãy lệnh cụ thể[2].
  - Ví dụ:
    - Dùng học máy để định danh người dùng
    - Dùng học máy để nhận dạng và tổng hợp tiếng nói
    - Dùng học máy để phân loại văn bản
    - Dùng học máy để dự báo nhu cầu về năng lượng, thực phẩm, v.v.

[1] **học từ kinh nghiệm trong quá khứ $\equiv$ học từ dữ liệu**

[2] nhiều trường hợp, nếu không dùng học máy cũng không có giải thuật nào khác để thay thế!

ltsach@hcmut.edu.vn
Lê Thành Sách

Khái niệm

Dữ liệu

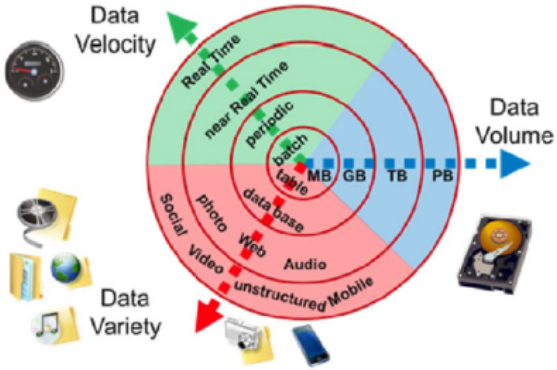
- **Dữ liệu:**
  - là tập các giá trị biểu diễn cho các đo đạc, quan sát, và mô tả về một tín hiệu, quá trình hay hoạt động nào đó
  - Ví dụ: video và hình ảnh trên Youtube, hồ sơ bệnh nhân trong bệnh viện, các tin nhắn người dùng, v.v.
- **Dữ liệu lớn:**
  - là dữ liệu nhưng có các tính chất sau:
    1. Độ lớn (**volume**): lượng dữ liệu là lớn tính theo dung lượng lưu trữ
    2. Tốc độ (**velocity**): dữ liệu được sản sinh liên tục và cần được xử lý liên tục; ví dụ, dữ liệu của các camera giám sát
    3. Tính đa dạng (**variety**): sự đa dạng về định dạng: văn bản, âm thanh, hình ảnh, chuỗi thời gian, v.v.

ltsach@hcmut.edu.vn

Lê Thành Sách

Khái niệm

Dữ liệu



Hình 1.1: Minh họa về dữ liệu lớn[1]

[1] nguồn: https://en.wikipedia.org/wiki/Big_data

ltsach@hcmut.edu.vn
Lê Thành Sách

Quan hệ hỗ tương

1. **Học máy cần dữ liệu để huấn luyện và chọn lựa mô hình, chọn lựa siêu tham số**. Chọn kỹ thuật học máy nào là tùy vào mục đích [1], sự sẵn sàng và tính chất của dữ liệu. Ví dụ:
   - Dữ liệu lớn: nên chọn kỹ thuật xử lý từng bó nhỏ dữ liệu thay cho phải phối hợp tất cả các điểm dữ liệu;
   - Dữ liệu ít có cấu trúc và quan hệ[2] đơn giản, không nên chọn mô hình quá phức tạp, như mạng nơron học sâu;
   - Dữ liệu có yếu tố thời gian → chọn mô hình có yếu tố thời gian.

[1] mục đích: để phân loại, hồi quy; để gom nhóm trên đặc tính dữ liệu; hay để chọn lựa hành động phù hợp với ngữ cảnh

[2] quan hệ giữa đầu vào và đầu ra của bài toán

Quan hệ hỗ tương

❷ **Dữ liệu cần học máy để phân tích**. Thu thập dữ liệu gì và như thế nào là tùy vào yêu cầu của bài toán và kỹ thuật học máy được chọn. Ví dụ:

- Đã chọn kỹ thuật học có giám sát thì cần làm nhãn [1] cho dữ liệu;
- Nếu kỹ thuật học giả thiết rằng các điểm dữ liệu được lấy mẫu độc lập thì việc chọn mẫu cũng tuân theo giả thiết đó;
- Yêu cầu của bài toán là định danh đến mức cá thể → phải gán nhãn là id của cá thể thay cho tên loại của cá thể.

[1] từ gốc là **label**, hay **target**: là đầu ra của mô hình dự báo

ltsach@hcmut.edu.vn
Lê Thành Sách

Các dạng học

ltsach@hcmut.edu.vn
Lê Thành Sách

Các dạng học

1. 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

ltsach@hcmut.edu.vn
Lê Thành Sách

Các dạng học

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.

Các dạng học

3. 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**)

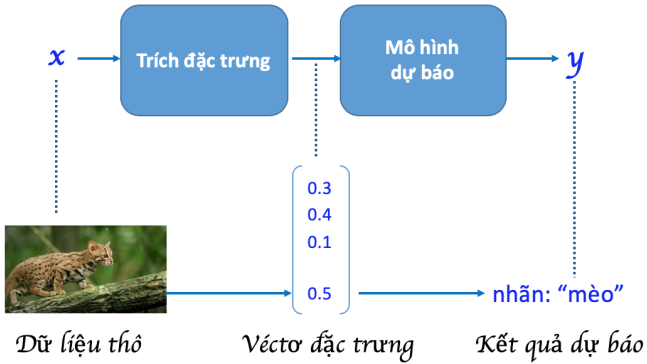
ltsach@hcmut.edu.vn
Lê Thành Sách

Các cách tiếp cận khi phân tích dữ liệu

ltsach@hcmut.edu.vn
Lê Thành Sách

Các cách tiếp cận khi phân tích dữ liệu

Hai giai đoạn

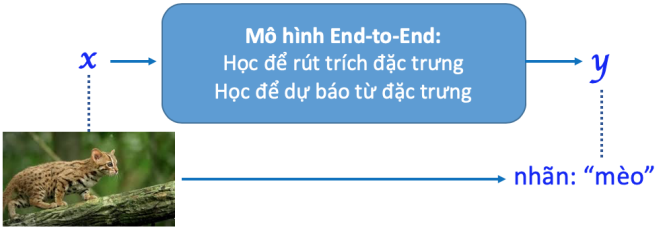


Hình 3.1: Cách tiếp cận hai giai đoạn[1]

[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

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

ltsach@hcmut.edu.vn
Lê Thành Sách

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

ltsach@hcmut.edu.vn
Lê Thành Sách

Lưu ý

Để phân tích dữ liệu hiệu quả cần hiểu rõ về tính chất dữ liệu và có cảm nhận về dữ liệu!

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

Chuẩn bị dữ liệu

Các đầu việc

- Thu thập, làm nhãn, làm sạch dữ liệu
- Thống kê và tóm tắt dữ liệu
- Rút trích đặc trưng
- Biến đổi và chuẩn hóa dữ liệu
- Hiển thị trực quan dữ liệu

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

Chuẩn bị dữ liệu

Các đầu việc

- Chia dữ liệu thành các tập con:
  - Nếu dùng phương pháp đánh giá **hold-out**, 3 tập: Tập huấn luyện (training set), tập kiểm thử (validation set), và tập kiểm tra (test set)
  - Nếu dùng phương pháp đánh giá **cross-validation** hay **leave-one-out**, 2 tập: Tập huấn luyện và kiểm thử, tập kiểm tra

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

Chuẩn bị dữ liệu

Lưu ý

- Không được đặt các điểm dữ liệu trong tập kiểm tra vào các tập kiểm thử và kiểm tra
- Cố tình thực hiện việc trên là một hình thức gian lận, nhằm có chỉ số đánh giá tốt cho mô hình

Các công việc khi phân tích dữ liệu

Xây dựng mô hình

Các đầu việc

- Lựa chọn kỹ thuật học máy
  - Mô hình tuyến tính
  - Máy véctơ hỗ trợ
  - Mạng nơron/học sâu, v.v.
- Huấn luyện và kiểm thử mô hình để thu được mô hình tốt nhất
- Kiểm tra mô hình thu được và ghi nhận/công bố/tuyên bố kết quả

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

Lựa chọn mô hình và siêu tham số

Khái niệm

- Mô hình được định nghĩa bởi:
  - Kỹ thuật học máy, ví dụ:
    - Mô hình tuyến tính
    - Máy véctơ hỗ trợ
    - Mạng nơron/học sâu, v.v.
  - Số lượng tham số, kiến trúc của các lớp tính toán (trong mạng nơron/học sâu)

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

Lựa chọn mô hình và siêu tham số

Khái niệm

- Siêu tham số (**hyper-parameter**): là tham số chọn trước khi huấn luyện mô hình, quá trình huấn luyện không cho ra giá trị của nó: ví dụ:
  - Số tham số của mô hình hồi quy tuyến tính;
  - Hệ số **weight decay**
  - Kiến trúc và số lượng tham số trong mỗi lớp tính toán của mạng nơron

ltsach@hcmut.edu.vn
Lê Thành Sách

Các công việc khi phân tích dữ liệu

Lựa chọn mô hình và siêu tham số

Các đầu việc

- Chọn ra danh sách các tổ hợp của các siêu tham số
- Huấn luyện mô hình vói từng bộ siêu tham số, trên tập huấn luyện
- Đánh giá mô hình thu được trên tập kiểm thử để cho ra các giá trị đánh giá
- Chọn lựa một tổ hợp làm cho giá trị đánh giá là tốt nhất => mô hình hay siêu tham số tối ưu

ltsach@hcmut.edu.vn
Lê Thành Sách