Machine Learning - Concepts

Lê Thành Sách

NỘI DUNG

- AI, ML, DL
- Các kiểu học
  - Giám sát
  - Không giám sát
  - Bán giám sát
  - Học tăng cường
- Bài toán
  - Phân loại
  - Hồi quy
- Các tập dữ liệu
  - Tập huấn luyện, tập kiểm thử, tập kiểm tra

Học có giám sát
supervised learning

- Mô hình toán:
  - y = f(x; W)
  - f(.) là hàm chưa biết, nhưng có thể quan sát (lấy mẫu) đầu vào (x) và đầu ra (y) của nó.
- Đầu vào của quá trình học
  - Tập các bộ {(x_i, y_i)}; i là chỉ số của cặp
    - x_i : dữ liệu (quan sát)
    - y_i: nhãn của x_i, giá trị kỳ vọng của hàm f(.) khi tính cho x_i
- Mục tiêu của học:
  - Ước lượng hàm f(.) từ dữ liệu {(x_i, y_i)}
    - Học W
- Ứng dụng:
  - Khi đã có hàm f(.) ➔ dùng f(.) để dự báo hay tính toán y cho những giá trị x chưa có trong quá trình lấy mẫu

Học có không giám sát unsupervised learning

- Đầu vào của quá trình học
  - Tập các bộ {(x_i)}; i là chỉ số của dữ liệu
    - Lưu ý: học không giám sát thì không cần dùng đến nhãn của dữ liệu
- Mục tiêu của học:
  - Gom các điểm dữ liệu x_i gần nhau (trên độ đo nào đó) về cùng một nhóm
- Ứng dụng:
  - Khi đã có các nhóm ➔ có thể suy ra tính chất của các điểm dữ liệu chưa biết từ tính chất của các điểm dữ liệu trong nhóm nó thuộc vào

Học có bán giám sát
semi-supervised learning

- Thách thức của học có giám sát
  - Dán nhãn cho dữ liệu
  - Ví dụ:
    - Vẽ bounding-box cho các đối tượng
    - Dán nhãn cho các vùng ảnh trong bài toán phân đoạn
    - …
- Nhu cầu:
  - Có kỹ thuật học tận dụng được cả dữ liệu có nhãn và không có nhãn (lĩnh vực đang còn nghiên cứu)

Học tăng cường (incremental learning)

- Mục tiêu:
  - Học để chọn hành động (actions)
  - => Phù hợp cho lĩnh vực: game, robot, xe tự hành, …
- Đầu vào:
  - Danh mục hành động + mô tả
  - Mô trường thực hiện hành động
  - Hàm đánh giá

Các bài toán

Bài toán phân loại

- Mục tiêu:
  - Học quan hệ y = f(x)
  - y: thuộc tập hợp nhóm (categorical set); không có nhu cầu định thực hiện phép toán trên các giá trị của y
    - Ví dụ: y thuộc {Chó, Mèo, Gà}
    - => các phép toán sau đây không có ý nghĩa: Chó + Mèo; Chó * Mèo; 2*Chó; …
- Ba cách mã hoá nhãn y thông dụng:
  - Nhãn văn bản: Classes = {Chó, Mèo, Gà}
  - Nhãn chỉ số: 0: Chó; 1: Mèo; 2: Gà
  - Nhãn one-hot:
    - {1, 0, 0} : Chó
    - {0, 1, 0} : Mèo
    - {0, 0, 1}: Gà

Các bài toán

Bài toán Hồi quy

- Mục tiêu:
  - Học quan hệ y = f(x)
  - y: thuộc tập số thực => các phép toán số học với các giá trị y là có ý nghĩa
- Ví dụ:
  - Dự báo lượng mưa
  - Dự báo giá chứng khoán
  - Dự báo mức tiêu hao nhiên liệu, …

Các bài toán

Bài toán Phát hiện (đối tượng)

- Phát hiện = Phân loại + Hồi quy

Các bài toán

Bài toán Phân đoạn

- Phân loại ảnh = phân loại cho từng điểm ảnh

Các bài toán

Bài toán nhận dạng

- Có thể thực hiện thông qua bài toán phân loại; ví dụ các mô hình nhận dạng khuôn mặt

Các tập dữ liệu

- Các tập:
  - Tập huấn luyện (training set): tập dữ liệu dùng để học ra các tham số của mô hình
  - Tập kiểm thử (validation set): tập dữ liệu dùng để chọn các siêu tham số của mô hình
    - Ví dụ:
      - Với mạng nơron: số lớp nơron, hàm truyền, …;
      - Với SVM: hàm kernel, hệ số C, …
  - Tập kiểm tra (test set): tập dữ liệu dùng để đánh giá và công bố (độ chính xác, độ triệu hồi, …)
- Lưu ý:
  - Cả ba tập dữ liệu trên phải được lấy mẫu theo nguyên tắc "iid" (independent and identical distribution)

Các tập dữ liệu

- Lưu ý:
  - identical: các điểm dữ liệu phải được lấy mẫu cùng một phân phối
    - Ví dụ: tập kiểm tra đánh giá trên ảnh chụp ban đêm => các tập khác cũng phải chứa dữ liệu lấy mẫu trên ngữ cảnh này
  - independent: các điểm dữ liệu phải được lấy mẫu độc lập; không được nhân bản dữ liệu bằng phụ thuộc toán học, ví dụ như x_2 = 5*x_1 + 3