

Chủ đề: Học có giám sát

Vũ Lê Thế Anh,
Nguyễn Mạc Nam Trung,
Nguyễn Đình Hoàng Phúc

TRẠI HÈ PiMA ONLINE 2021



Trình bày: Vũ Lê Thế Anh, ĐH Khoa học Tự nhiên, ĐHQG-HCM

Ngày 28 tháng 7 năm 2021

1 Đặt vấn đề

2 Học có giám sát

Table of Contents

1 Đặt vấn đề

2 Học có giám sát



Nhắc lại kiến thức

Cho một tập các điểm dữ liệu.

- **Giảm chiều:** đưa mỗi điểm xuống một không gian ít chiều hơn
- **Phân cụm:** gom nhóm các điểm



Nhắc lại kiến thức

Cho một tập các điểm dữ liệu.

- **Giảm chiều:** đưa mỗi điểm xuống một không gian ít chiều hơn
- **Phân cụm:** gom nhóm các điểm

Điểm chung:

- đi tìm một cấu trúc tiềm ẩn
- không có “hướng dẫn”



Ví dụ

Điểm thi đại học

Cho điểm thi đại học của một số bạn học sinh.

Tên	Lý	Hóa	Sinh
Thư	9	8	5
Nguyễn	6	8	9
Linh	8	7	3
Trung	2	5	6
Nah	5	8	8



Ví dụ

Điểm thi đại học

Cho điểm thi đại học của một số bạn học sinh.

Tên	Lý	Hóa	Sinh
Thư	9	8	5
Nguyễn	6	8	9
Linh	8	7	3
Trung	2	5	6
Nah	5	8	8

Câu hỏi #1: Gom các bạn học sinh thành các nhóm.



Ví dụ - Gom cụm 3 nhóm

Tên	Lý	Hóa	Sinh
Thư	9	8	5
Nguyễn	6	8	9
Linh	8	7	3
Trung	2	5	6
Nah	5	8	8



Ví dụ - Gom cụm 2 nhóm

Tên	Lý	Hóa	Sinh
Thư	9	8	5
Nguyễn	6	8	9
Linh	8	7	3
Trung	2	5	6
Nah	5	8	8



Ví dụ

Điểm thi đại học

Cho điểm thi đại học của một số bạn học sinh.

Tên	Lý	Hóa	Sinh
Thư	9	8	5
Nguyễn	6	8	9
Linh	8	7	3
Trung	2	5	6
Nah	5	8	8

Câu hỏi #2: Xác định khối thi dự định của các bạn (biết có 2 khối là A hoặc B).



Ví dụ - Phân lớp

Tên	Lý	Hóa	Sinh	Dự đoán 1	Dự đoán 2
Thư	9	8	5	B	A
Nguyễn	6	8	9	A	B
Linh	8	7	3	B	A
Trung	2	5	6	A	B
Nah	5	8	8	A	B



Ví dụ

Điểm thi đại học

Cho điểm thi đại học của một số bạn học sinh.

Tên	Lý	Hóa	Sinh
Thư	9	8	5
Nguyễn	6	8	9
Linh	8	7	3
Trung	2	5	6
Nah	5	8	8

Câu hỏi #3: Xác định điểm thi môn Toán của các bạn.



Ví dụ - Hồi quy

Tên	Lý	Hóa	Sinh	Dự đoán (mò)
Thư	9	8	5	cao
Nguyễn	6	8	9	trung bình
Linh	8	7	3	hơi cao (?)
Trung	2	5	6	thấp
Nah	5	8	8	trung bình (?)



Đặt vấn đề

- Hai bài toán mới: **phân loại** và **hồi quy**



Đặt vấn đề

- Hai bài toán mới: **phân loại** và **hồi quy**
- **Giống**: tìm cấu trúc tiềm ẩn



Đặt vấn đề

- Hai bài toán mới: **phân loại** và **hồi quy**
- **Giống**: tìm cấu trúc tiềm ẩn
- **Khác**: cần hướng dẫn



Table of Contents

1 Đặt vấn đề

2 Học có giám sát



Khái niệm

- **Biết:** tập dữ liệu **có nhãn**

Khái niệm

- **Biết**: tập dữ liệu **có nhãn**
- **Muốn**: một **mô hình** đưa ra dự đoán cho các điểm dữ liệu **mới** một cách **tốt nhất**



Tập dữ liệu có gán nhãn

Mỗi điểm dữ liệu:

- Đặc trưng (feature)
- Nhãn (label)



Tập dữ liệu có gán nhãn

Mỗi điểm dữ liệu:

- Đặc trưng (feature)
- Nhãn (label)

Nhãn và đặc trưng

Nhãn là đặc trưng mà chúng ta quan tâm.



Tập dữ liệu có gán nhãn

Mỗi điểm dữ liệu:

- Đặc trưng (feature)
- Nhãn (label)

Nhãn và đặc trưng

Nhãn là đặc trưng mà chúng ta quan tâm.

Tập dữ liệu

- Điểm thi đại học và nguyện vọng khối thi của các bạn
- Bài báo có tiêu đề, nội dung, và các hình ảnh
- Mạng xã hội có người dùng và quan hệ bạn bè



Mô hình

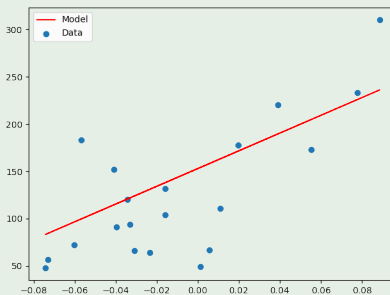
Thành phần

- Cơ chế đưa ra dự đoán
- Cách thức có được cơ chế (hoàn chỉnh)
 - Thông thường: tối ưu một hàm mục tiêu



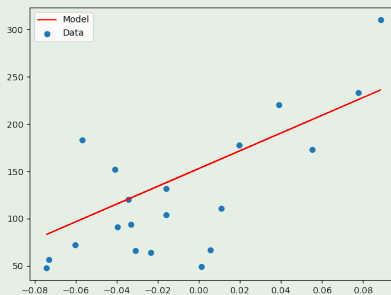
Mô hình - Ví dụ

Hồi quy tuyến tính



Mô hình - Ví dụ

Hồi quy tuyến tính

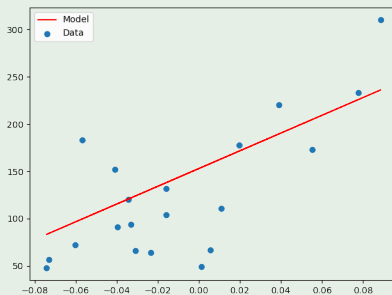


- Cơ chế ra dự đoán: $f_{a,b}(x) = ax + b$



Mô hình - Ví dụ

Hồi quy tuyến tính

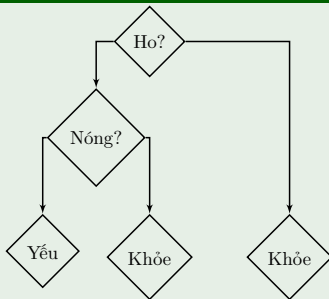


- Cơ chế ra dự đoán: $f_{a,b}(x) = ax + b$
- Cách hoàn chỉnh cơ chế: tìm a và b sao cho độ lỗi tối thiểu



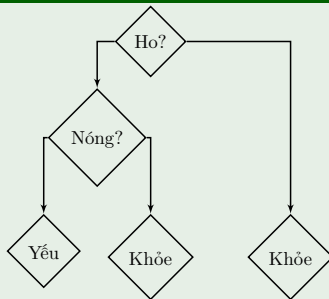
Mô hình - Ví dụ

Cây quyết định



Mô hình - Ví dụ

Cây quyết định

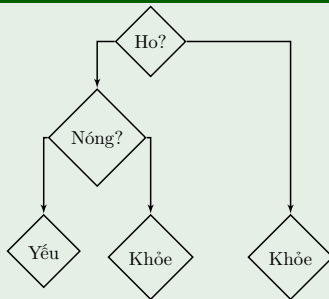


- Cơ chế ra dự đoán: theo hướng dẫn của cây phân loại



Mô hình - Ví dụ

Cây quyết định



- Cơ chế ra dự đoán: theo hướng dẫn của cây phân loại
- Cách hoàn chỉnh cơ chế: tìm cấu trúc cây



Đánh giá

- Kiểm tra chất lượng của một mô hình



Đánh giá

- Kiểm tra chất lượng của một mô hình
- Một số cách:
 - Dùng hàm mục tiêu (khi xây dựng mô hình)
 - Dùng cách đánh giá khác



Đánh giá

- Kiểm tra chất lượng của một mô hình
- Một số cách:
 - Dùng hàm mục tiêu (khi xây dựng mô hình)
 - Dùng cách đánh giá khác

Đánh giá mô hình

Đánh giá phải được thực hiện trên dữ liệu **mới** và nên **gần thực tế** nhất có thể.



Các pha

- 1 Tiền xử lý và khai phá dữ liệu
- 2 Xây dựng mô hình và kiểm thử sơ bộ
- 3 Đánh giá mô hình và ứng dụng thực tế



Tiền xử lý và khai phá dữ liệu

Khai phá dữ liệu

- Thống kê
- Trực quan hóa

Tiền xử lý:

- Làm sạch
- Chuẩn hóa
- Phân tách (cho bước mô hình hóa)



Xây dựng mô hình và kiểm thử sơ bộ

- Xây dựng mô hình: trên tập huấn luyện (train set)
- Kiểm thử sơ bộ: trên tập kiểm thử (validation set)



Đánh giá mô hình và ứng dụng thực tế

- Đưa mô hình vào sử dụng thực tế
- Nhận phản hồi thực tế để đánh giá



Giới thiệu dự án

- Generalized Linear Model
 - Tổng quát hóa của mô hình Hồi quy Tuyến tính
 - Có thể dùng cho hồi quy và phân loại
 - Kiến thức: thống kê, tối ưu hóa



Giới thiệu dự án

- Generalized Linear Model
 - Tổng quát hóa của mô hình Hồi quy Tuyến tính
 - Có thể dùng cho hồi quy và phân loại
 - Kiến trúc: thống kê, tối ưu hóa
- Support Vector Machine
 - Chủ yếu dùng cho phân loại
 - Kiến trúc: đại số tuyến tính, tối ưu hóa

