

Tổng quan về ML/AI trong Khoa học Dữ Liệu

Lê Thiện

MIT/CSAIL

Ngày 25 tháng 7 năm 2021

Sơ lược

- 1 Khoa học Dữ liệu
 - Dữ liệu
 - Khoa học Dữ liệu
 - Big data
- 2 Máy học (ML) trong Khoa học Dữ Liệu
 - Giải thích dữ liệu - Sơ lược về máy học
 - Hồi quy tuyến tính (Linear Regression)
 - K-means
- 3 Cân bằng độ lệch - phương sai (Bias-variance trade-off)
 - Trở lại bài toán hồi quy tuyến tính
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical model
 - Ước lượng (estimator)
 - Cân bằng độ lệch - phương sai
- 4 Q&A
 - Q&A

Sơ lược

- 1 Khoa học Dữ liệu
 - Dữ liệu
 - Khoa học Dữ liệu
 - Big data
- 2 Máy học (ML) trong Khoa học Dữ Liệu
 - Giải thích dữ liệu - Sơ lược về máy học
 - Hồi quy tuyến tính (Linear Regression)
 - K-means
- 3 Cân bằng độ lệch - phương sai (Bias-variance trade-off)
 - Trở lại bài toán hồi quy tuyến tính
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical model
 - Ước lượng (estimator)
 - Cân bằng độ lệch - phương sai
- 4 Q&A
 - Q&A

Dữ liệu (data) là gì?

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Ví dụ 1

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**
 - ① điểm kiểm tra 15', 1 tiết, học kì, ... ,

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**
 - 1 điểm kiểm tra 15', 1 tiết, học kì, ... ,
 - 2 điểm hạnh kiểm,

Dữ liệu là gì? Bài toán nào, dữ liệu gì?

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**
 - 1 điểm kiểm tra 15', 1 tiết, học kì, ... ,
 - 2 điểm hạnh kiểm,
 - 3 nhận xét của giáo viên.

Phân loại sơ lược dữ liệu

Phân loại sơ lược dữ liệu

Ví dụ 1

Phân loại sơ lược dữ liệu

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh

Phân loại sơ lược dữ liệu

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**

Phân loại sơ lược dữ liệu

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**
 - ① điểm kiểm tra 15', 1 tiết, học kì, . . . , - dữ liệu số (numerical), để có cấu trúc (structure)

Phân loại sơ lược dữ liệu

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**
 - 1 điểm kiểm tra 15', 1 tiết, học kì, . . . , - dữ liệu số (numerical), để có cấu trúc (structure)
 - 2 điểm hạnh kiểm, - dữ liệu số (numerical)

Phân loại sơ lược dữ liệu

Ví dụ 1

- **Bài toán sơ lược:** nhà trường phân loại học sinh
- **Dữ liệu:**
 - 1 điểm kiểm tra 15', 1 tiết, học kì, . . . , - dữ liệu số (numerical), để có cấu trúc (structure)
 - 2 điểm hạnh kiểm, - dữ liệu số (numerical)
 - 3 nhận xét của giáo viên - dữ liệu văn bản, ký tự (textual).

Phân loại sơ lược dữ liệu

Phân loại sơ lược dữ liệu

Ví dụ 2

Phân loại sơ lược dữ liệu

Ví dụ 2

- **Bài toán sơ lược:** doanh nghiệp bán lẻ (retail) muốn tối ưu hoá lợi nhuận (profit)

Phân loại sơ lược dữ liệu

Ví dụ 2

- **Bài toán sơ lược:** doanh nghiệp bán lẻ (retail) muốn tối ưu hoá lợi nhuận (profit)
- **Dữ liệu:**

Phân loại sơ lược dữ liệu

Ví dụ 2

- **Bài toán sơ lược:** doanh nghiệp bán lẻ (retail) muốn tối ưu hoá lợi nhuận (profit)
- **Dữ liệu:**
 - ① lợi nhuận những năm (quý/tháng) trước, - dữ liệu phức tạp: dãy số thời gian (time series)

Phân loại sơ lược dữ liệu

Ví dụ 2

- **Bài toán sơ lược:** doanh nghiệp bán lẻ (retail) muốn tối ưu hoá lợi nhuận (profit)
- **Dữ liệu:**
 - 1 lợi nhuận những năm (quý/tháng) trước, - dữ liệu phức tạp: dãy số thời gian (time series)
 - 2 nhận xét của khách hàng, - dữ liệu văn bản

Phân loại sơ lược dữ liệu

Ví dụ 2

- **Bài toán sơ lược:** doanh nghiệp bán lẻ (retail) muốn tối ưu hoá lợi nhuận (profit)
- **Dữ liệu:**
 - ① lợi nhuận những năm (quý/tháng) trước, - dữ liệu phức tạp: dãy số thời gian (time series)
 - ② nhận xét của khách hàng, - dữ liệu văn bản
 - ③ chiến lược của đối thủ cạnh tranh, - dữ liệu trừu tượng (abstract).
 - ④ ...

Tại sao phải phân loại dữ liệu?

Tại sao phải phân loại dữ liệu?

Thuật toán, phần mềm cụ thể để thu thập, xử lý, phân tích mỗi loại dữ liệu.

Tại sao phải phân loại dữ liệu?

Thuật toán, phần mềm cụ thể để thu thập, xử lý, phân tích mỗi loại dữ liệu.

- Dữ liệu số: đa số thuật toán/phần mềm cổ điển

Tại sao phải phân loại dữ liệu?

Thuật toán, phần mềm cụ thể để thu thập, xử lý, phân tích mỗi loại dữ liệu.

- Dữ liệu số: đa số thuật toán/phần mềm cổ điển
- Dữ liệu văn bản: natural language processing (NLP), vv

Tại sao phải phân loại dữ liệu?

Thuật toán, phần mềm cụ thể để thu thập, xử lý, phân tích mỗi loại dữ liệu.

- Dữ liệu số: đa số thuật toán/phần mềm cổ điển
- Dữ liệu văn bản: natural language processing (NLP), vv
- Dữ liệu hình ảnh: machine vision, vv

Khoa học trong tự nhiên

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.
- **Giải:**

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.
- **Giải:**
 - 1 Thử với nhiều L và thu thập T tương ứng (dữ liệu),

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.
- **Giải:**
 - 1 Thử với nhiều L và thu thập T tương ứng (dữ liệu),
 - 2 Vẽ lên hệ Oxy đồ thị $L^{\frac{1}{2}}$ và T ,

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.
- **Giải:**
 - 1 Thử với nhiều L và thu thập T tương ứng (dữ liệu),
 - 2 Vẽ lên hệ Oxy đồ thị $L^{\frac{1}{2}}$ và T ,
 - 3 Vẽ đường thẳng đi (gần) qua mọi điểm,

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.
- **Giải:**
 - 1 Thử với nhiều L và thu thập T tương ứng (dữ liệu),
 - 2 Vẽ lên hệ Oxy đồ thị $L^{\frac{1}{2}}$ và T ,
 - 3 Vẽ đường thẳng đi (gần) qua mọi điểm,
 - 4 Tính hệ số góc (slope) của đường thẳng.

Khoa học trong tự nhiên

- Chuyển động con lắc $T = 2\pi\sqrt{\frac{L}{g}}$
- Một nhà khoa học không tin vào công thức tính chu kỳ T trong chuyển động con lắc với độ dài L và muốn xác minh lại bằng cách tìm m sao cho $T = mL^{\frac{1}{2}}$.
- **Giải:**
 - 1 Thử với nhiều L và thu thập T tương ứng (dữ liệu),
 - 2 Vẽ lên hệ Oxy đồ thị $L^{\frac{1}{2}}$ và T ,
 - 3 Vẽ đường thẳng đi (gần) qua mọi điểm,
 - 4 Tính hệ số góc (slope) của đường thẳng.
- Tự nhiên là hộp đen để nhà khoa học thu thập dữ liệu và dùng nó để tính mọi chu kỳ T khác chưa từng được thử.

Khoa học Dữ liệu

Khoa học Dữ liệu

Tổng quát hoá hai bài toán sơ lược và khoa học trong tự nhiên:

Khoa học Dữ liệu

Tổng quát hoá hai bài toán sơ lược và khoa học trong tự nhiên:

- Môi trường \mathcal{M} phức tạp, rất nhiều biến số, không liên tục, ...

Khoa học Dữ liệu

Tổng quát hoá hai bài toán sơ lược và khoa học trong tự nhiên:

- Môi trường \mathcal{M} phức tạp, rất nhiều biến số, không liên tục, ...
- Dữ liệu \mathcal{D} thu thập được về môi trường \mathcal{M} .

Khoa học Dữ liệu

Tổng quát hoá hai bài toán sơ lược và khoa học trong tự nhiên:

- Môi trường \mathcal{M} phức tạp, rất nhiều biến số, không liên tục, ...
- Dữ liệu \mathcal{D} thu thập được về môi trường \mathcal{M} .
- Lựa chọn mô hình A phụ thuộc vào dữ liệu \mathcal{D} .

Mục đích của khoa học dữ liệu

Mục đích của khoa học dữ liệu

Môi trường \mathcal{M} , dữ liệu \mathcal{D} , mô hình $A(\mathcal{D})$.

Mục đích của khoa học dữ liệu

Môi trường \mathcal{M} , dữ liệu \mathcal{D} , mô hình $A(\mathcal{D})$.

- Tìm phương pháp thu thập \mathcal{D} để thu tóm nhiều nhất thông tin về \mathcal{M} (liên quan đến việc chọn A).

Mục đích của khoa học dữ liệu

Môi trường \mathcal{M} , dữ liệu \mathcal{D} , mô hình $A(\mathcal{D})$.

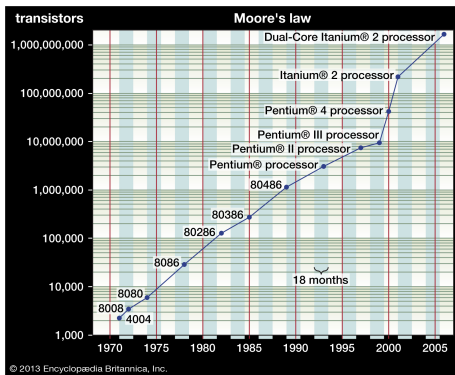
- Tìm phương pháp thu thập \mathcal{D} để thu tóm nhiều nhất thông tin về \mathcal{M} (liên quan đến việc chọn A).
- Xử lý thông tin (transform) \mathcal{D} cho gọn nhẹ mà không mất nhiều thông tin.

Mục đích của khoa học dữ liệu

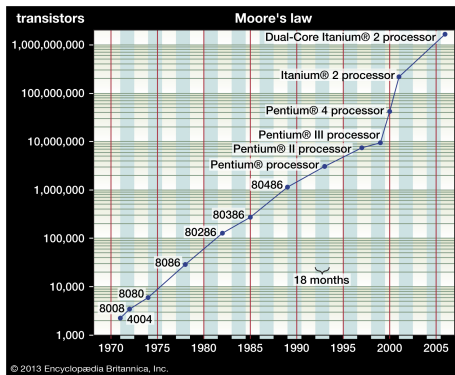
Môi trường \mathcal{M} , dữ liệu \mathcal{D} , mô hình $A(\mathcal{D})$.

- Tìm phương pháp thu thập \mathcal{D} để thu tóm nhiều nhất thông tin về \mathcal{M} (liên quan đến việc chọn A).
- Xử lý thông tin (transform) \mathcal{D} cho gọn nhẹ mà không mất nhiều thông tin.
- Hệ thống hoá việc lựa chọn mô hình A dựa vào dữ liệu \mathcal{D} .

Định luật Moore



Định luật Moore



Dữ liệu \mathcal{D} thu thập được tăng về khối lượng lẫn chiều.

Mục đích của khoa học dữ liệu hiện đại

Thuật toán mới, mạnh hơn, chặt chẽ hơn để

- Tìm phương pháp thu thập \mathcal{D} để thu tóm nhiều nhất thông tin về \mathcal{M} (liên quan đến việc chọn A).
- Xử lý thông tin (transform) \mathcal{D} cho gọn nhẹ mà không mất nhiều thông tin.
- Hệ thống việc lựa chọn mô hình A dựa vào dữ liệu \mathcal{D} .

Sơ lược

- 1 Khoa học Dữ liệu
 - Dữ liệu
 - Khoa học Dữ liệu
 - Big data
- 2 Máy học (ML) trong Khoa học Dữ Liệu
 - Giải thích dữ liệu - Sơ lược về máy học
 - Hồi quy tuyến tính (Linear Regression)
 - K-means
- 3 Cân bằng độ lệch - phương sai (Bias-variance trade-off)
 - Trở lại bài toán hồi quy tuyến tính
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical model
 - Ước lượng (estimator)
 - Cân bằng độ lệch - phương sai
- 4 Q&A
 - Q&A

Bài toán máy học là bài toán tối ưu hoá (optimization)

Bài toán máy học là bài toán tối ưu hoá (optimization)

- Cho dữ liệu \mathcal{D} .

Bài toán máy học là bài toán tối ưu hoá (optimization)

- Cho dữ liệu \mathcal{D} .
- Cho tập hợp hàm số \mathcal{H} . (Giả định: mô hình A đúng nằm trong \mathcal{H})

Bài toán máy học là bài toán tối ưu hoá (optimization)

- Cho dữ liệu \mathcal{D} .
- Cho tập hợp hàm số \mathcal{H} . (Giả định: mô hình A đúng nằm trong \mathcal{H})
- Cho hàm mất mát $L_{\mathcal{D}} : \mathcal{H} \rightarrow \mathbb{R}$ để đánh giá hàm số trong \mathcal{H} trong việc giải thích dữ liệu \mathcal{D} .

Bài toán máy học là bài toán tối ưu hoá (optimization)

- Cho dữ liệu \mathcal{D} .
- Cho tập hợp hàm số \mathcal{H} . (Giả định: mô hình A đúng nằm trong \mathcal{H})
- Cho hàm mất mát $L_{\mathcal{D}} : \mathcal{H} \rightarrow \mathbb{R}$ để đánh giá hàm số trong \mathcal{H} trong việc giải thích dữ liệu \mathcal{D} .
- Tìm hàm số $f^* \in \mathcal{H}$ giải thích dữ liệu tốt nhất:

$$L_{\mathcal{D}}(f^*) \leq L_{\mathcal{D}}(f), \forall f \in \mathcal{H}.$$

và dùng f^* làm mô hình cho các bước tiếp theo.

Bài toán nhỏ

- Hoàn cảnh
 - Cho biến x, y
 - Cho biết hàm số f liên hệ x với y có bậc $= 1$
- Dữ liệu
 - Nếu $x = 1$ thì $y = 1$
 - Nếu $x = 2$ thì $y = 5$
- Tìm f

Bài toán lớn hơn

- Hoàn cảnh
 - Cho biến x, y
 - Cho biết tồn tại hàm f bậc 1 sao cho $f(x)$ 'gần' với y
- Dữ liệu
 - Nếu $x = 1$ thì $y = 1$
 - Nếu $x = 2$ thì $y = 5$
 - Nếu $x = 3$ thì $y = 6$
- Tìm f

Bài toán 1 chiều

- Dữ liệu

- $(x_i)_{i=1}^n \in \mathbb{R}$
- $(y_i)_{i=1}^n \in \mathbb{R}$

Bài toán 1 chiều

- Dữ liệu

- $(x_i)_{i=1}^n \in \mathbb{R}$
- $(y_i)_{i=1}^n \in \mathbb{R}$

- Cho

- $\mathcal{H} := \{f : x \mapsto ax + b \mid a, b \in \mathbb{R}\}$
- $L(f) := \sum_{i=1}^n (f(x_i) - y_i)^2$

Bài toán 1 chiều

- Dữ liệu

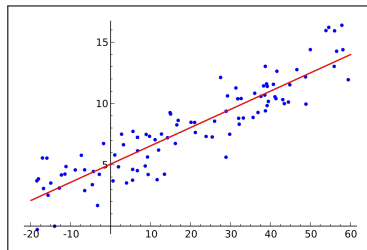
- $(x_i)_{i=1}^n \in \mathbb{R}$
- $(y_i)_{i=1}^n \in \mathbb{R}$

- Cho

- $\mathcal{H} := \{f : x \mapsto ax + b \mid a, b \in \mathbb{R}\}$
- $L(f) := \sum_{i=1}^n (f(x_i) - y_i)^2$

- Tìm

- $f^* \in \mathcal{H}$ sao cho $L(f^*)$ nhỏ nhất



Bài toán d -chiều

- Dữ liệu

- $(\mathbf{x}_i)_{i=1}^n \in \mathbb{R}^d$
- $(y_i)_{i=1}^n \in \mathbb{R}$

- Cho

- $\mathcal{H} \subseteq \{f : \mathbf{x} \mapsto \langle \mathbf{a}, \mathbf{x} \rangle + b \mid \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}\}$
- $L(f) = \sum_{i=1}^n (f(x_i) - y_i)^2$

- Tìm

- $f^* \in \mathcal{H}$ sao cho $L(f^*)$ nhỏ nhất
(hoặc $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$ sao cho $L(\mathbf{x} \mapsto \langle \mathbf{a}, \mathbf{x} \rangle + b)$ nhỏ nhất)

Giải

- Bài toán optimization này có nghiệm giải tích hoàn toàn (analytical solution), muốn biết kết quả chỉ cần bỏ dữ liệu vào công thức chuẩn (normal equation)

Giải

- Bài toán optimization này có nghiệm giải tích hoàn toàn (analytical solution), muốn biết kết quả chỉ cần bỏ dữ liệu vào công thức chuẩn (normal equation)
- Nhận xét 1: không phải bài toán ML nào cũng phức tạp

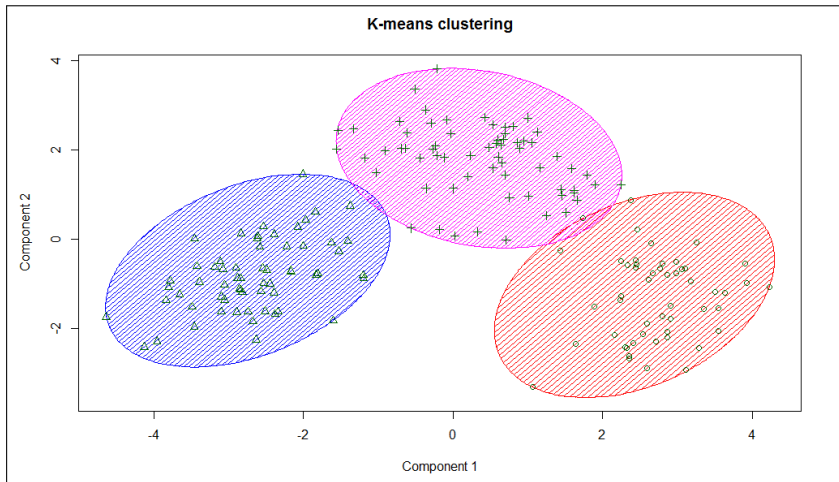
Giải

- Bài toán optimization này có nghiệm giải tích hoàn toàn (analytical solution), muốn biết kết quả chỉ cần bỏ dữ liệu vào công thức chuẩn (normal equation)
- Nhận xét 1: không phải bài toán ML nào cũng phức tạp
- Nhận xét 2: \mathcal{H} đơn giản. L là hàm liên tục, khả vi (theo tham số của \mathcal{H})

Giải

- Bài toán optimization này có nghiệm giải tích hoàn toàn (analytical solution), muốn biết kết quả chỉ cần bỏ dữ liệu vào công thức chuẩn (normal equation)
- Nhận xét 1: không phải bài toán ML nào cũng phức tạp
- Nhận xét 2: \mathcal{H} đơn giản. L là hàm liên tục, khả vi (theo tham số của \mathcal{H})
- Trong thực tế, ít xài công thức chuẩn vì nó không ổn định số học (numerically stable)

Minh họa



Bài toán

- Dữ liệu

- $X = (x_i)_{i=1}^n \in \mathbb{R}^d$
- Không có mác!

Bài toán

- Dữ liệu

- $X = (x_i)_{i=1}^n \in \mathbb{R}^d$
- Không có mác!

- Cho

- $\mathcal{H} \subseteq \{f : X \mapsto \{1, 2, \dots, k\}\}$ tập hợp các hàm gán điểm trong $x \in X$ vào cluster $S_{f(x)}$
- $L(f) = \sum_{i=1}^k |S_i| \text{Var}[S_i]$
 - $\text{Var}[S_i]$ đo phương sai của các điểm x có $f(x) = i$

Bài toán

- Dữ liệu

- $X = (x_i)_{i=1}^n \in \mathbb{R}^d$
- Không có mác!

- Cho

- $\mathcal{H} \subseteq \{f : X \mapsto \{1, 2, \dots, k\}\}$ tập hợp các hàm gán điểm trong $x \in X$ vào cluster $S_{f(x)}$
- $L(f) = \sum_{i=1}^k |S_i| \text{Var}[S_i]$
 - $\text{Var}[S_i]$ đo phương sai của các điểm x có $f(x) = i$

- Tìm

- $f^* \in \mathcal{H}$ sao cho $L(f^*)$ nhỏ nhất

Giải

- Không có nghiệm giải tích hoàn toàn. Dùng thuật toán dự đoán nghiệm (heuristics). Không có định lý về tính đúng sai.

Giải

- Không có nghiệm giải tích hoàn toàn. Dùng thuật toán dự đoán nghiệm (heuristics). Không có định lý về tính đúng sai.
- Thuật toán EM (Expectation - Maximization) (không trong scope)

Giải

- Không có nghiệm giải tích hoàn toàn. Dùng thuật toán dự đoán nghiệm (heuristics). Không có định lý về tính đúng sai.
- Thuật toán EM (Expectation - Maximization) (không trong scope)
- Nhận xét 1: Phần lớn bài toán ML đòi hỏi giải một bài optimization khó như kmeans, cần dùng những thuật toán numerical methods phức tạp, khó kiểm soát hơn.

Giải

- Không có nghiệm giải tích hoàn toàn. Dùng thuật toán dự đoán nghiệm (heuristics). Không có định lý về tính đúng sai.
- Thuật toán EM (Expectation - Maximization) (không trong scope)
- Nhận xét 1: Phần lớn bài toán ML đòi hỏi giải một bài optimization khó như kmeans, cần dùng những thuật toán numerical methods phức tạp, khó kiểm soát hơn.
- Nhận xét 2: Lựa chọn \mathcal{H} và L cân bằng giữa độ khó của bài toán và ý nghĩa của bài toán.

Sơ lược

- 1 Khoa học Dữ liệu
 - Dữ liệu
 - Khoa học Dữ liệu
 - Big data
- 2 Máy học (ML) trong Khoa học Dữ Liệu
 - Giải thích dữ liệu - Sơ lược về máy học
 - Hồi quy tuyến tính (Linear Regression)
 - K-means
- 3 Cân bằng độ lệch - phương sai (Bias-variance trade-off)
 - Trở lại bài toán hồi quy tuyến tính
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical model
 - Ước lượng (estimator)
 - Cân bằng độ lệch - phương sai
- 4 Q&A
 - Q&A

Bài toán lớn hơn

- Hoàn cảnh
 - Cho biến x, y
 - Cho biết tồn tại hàm f bậc 1 sao cho $f(x)$ ‘gần’ với y
- Dữ liệu
 - Nếu $x = 1$ thì $y = 1$
 - Nếu $x = 2$ thì $y = 5$
 - Nếu $x = 3$ thì $y = 6$
- Tìm f

Gia đình hàm trong hồi quy tuyến tính

- $f(x) \neq y$

Gia đình hàm trong hồi quy tuyến tính

- $f(x) \neq y$
- $f(x) \approx y$

Gia đình hàm trong hồi quy tuyến tính

- $f(x) \neq y$
- $f(x) \approx y$
- $f(x) = ax + b + \epsilon$, trong đó $a, b \in \mathbb{R}$ còn ϵ là sai số ngẫu nhiên

Mô hình thống kê (Statistical model)

- ϵ trong $f(x) = ax + b + \epsilon$ là 1 biến ngẫu nhiên

Mô hình thống kê (Statistical model)

- ϵ trong $f(x) = ax + b + \epsilon$ là 1 biến ngẫu nhiên
- f là 1 hàm số ngẫu nhiên

Mô hình thống kê (Statistical model)

- ϵ trong $f(x) = ax + b + \epsilon$ là 1 biến ngẫu nhiên
- f là 1 hàm số ngẫu nhiên
- $\mathcal{H} = \{x \rightarrow ax + b + \epsilon | a, b \in \mathbb{R}\}$ là 1 mô hình thống kê

Tại sao cần biến ngẫu nhiên

- Mô hình mẫu của dữ liệu
- Lý thuyết ngoại suy, nội suy
- Mục đích của máy học

Mô hình mẫu

- Bài toán ML = optimization + thống kê
- Dùng xác suất thống kê dựng mô hình mẫu cho dữ liệu

Định nghĩa vắn tắt

Bài toán ML là bài toán tìm tham số của mô hình thống kê giải thích dữ liệu (data)

Ước lượng tham số (Parameter Fitting)

Định nghĩa vắn tắt

Bài toán ML là bài toán tìm tham số của mô hình thống kê giải thích dữ liệu (data)

- Tìm tham số như thế nào?

Ước lượng (Estimator)

- Hoàn cảnh
 - Cho biết $y = ax + b + \epsilon$ với tham số a, b nào đó, ϵ là biến ngẫu nhiên
- Dữ liệu
 - Nếu $x = 1$ thì $y = 1$
 - Nếu $x = 2$ thì $y = 5$
 - Nếu $x = 3$ thì $y = 6$
- Tìm f (hoặc tìm a, b)
 - Dùng dữ liệu hữu hạn để đoán tham số $\hat{a}(\epsilon), \hat{b}(\epsilon)$

Ước lượng

Ước lượng của một tham số là một cách dựa vào dữ liệu để đoán tham số.

Ước lượng không chệch (unbiased estimator)

- Cho tham số a , ước lượng \hat{a} phụ thuộc vào dữ liệu
- Chệch (bias) $:= E[\hat{a}] - a$

Ước lượng không chệch (unbiased estimator)

- Cho tham số a , ước lượng \hat{a} phụ thuộc vào dữ liệu
- Chệch (bias) $:= E[\hat{a}] - a$
- Ước lượng không chệch $\iff \text{bias} = 0$

Ước lượng hiệu quả (Efficient estimator)

- Phương sai của ước lượng $Var[\hat{a}] = E[(E[\hat{a}] - \hat{a})^2]$

Ước lượng hiệu quả (Efficient estimator)

- Phương sai của ước lượng $Var[\hat{a}] = E[(E[\hat{a}] - \hat{a})^2]$
- Mean squared error (MSE) $:= E[(a - \hat{a})^2]$

Ước lượng hiệu quả (Efficient estimator)

- Phương sai của ước lượng $Var[\hat{a}] = E[(E[\hat{a}] - \hat{a})^2]$
- Mean squared error (MSE) $:= E[(a - \hat{a})^2]$
- Ước lượng không chệch với phương sai nhỏ nhất $:=$ ước lượng hiệu quả

Cân bằng độ lệch - phương sai

- $MSE = \text{phương sai} + \text{bias}^2$

Cân bằng độ lệch - phương sai

- $MSE = \text{phương sai} + \text{bias}^2$
- Phương sai lớn biểu diễn overfit
- Chênh lệch lớn biểu diễn underfit

Cân bằng độ lệch - phương sai

- $MSE = \text{phương sai} + \text{bias}^2$
- Phương sai lớn biểu diễn overfit
- Chệch lớn biểu diễn underfit
- Để tìm ước lượng tốt cần giảm cả 2.

Sơ lược

- 1 Khoa học Dữ liệu
 - Dữ liệu
 - Khoa học Dữ liệu
 - Big data
- 2 Máy học (ML) trong Khoa học Dữ Liệu
 - Giải thích dữ liệu - Sơ lược về máy học
 - Hồi quy tuyến tính (Linear Regression)
 - K-means
- 3 Cân bằng độ lệch - phương sai (Bias-variance trade-off)
 - Trở lại bài toán hồi quy tuyến tính
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical model
 - Ước lượng (estimator)
 - Cân bằng độ lệch - phương sai
- 4 Q&A
 - Q&A

Q&A

Q&A