Sơ lược Tại sao dùng ML ML là gi? Hồi quy tuyến tính (Linear Regression) K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off) Q&A và Extra

Giới thiệu Máy học (Machine Learning - ML)

Thiên Lê

UIUC

Sơ lước Tai sao dùng ML ML là gì? Hồi quy tuyến tính (Linear Regression) K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off)

Q&A và Extra

Sơ lược

- 1 Tại sao dùng ML
 - Trí tuê nhân tao (Artificial Inteligence AI)
 - Rule-based: Truyền tải thông minh đơn giản
- 2 ML là gì?
 - Bài toán ML là bài toán tìm hàm số
 - Bài toán ML là bài toán tối ưu hoá (optimization)
- 3 Hồi quy tuyến tính (Linear Regression)
 - Bài toán
 - Giải ■ Hoc có giám sát (Supervised Learning)
 - Hồi quy (Regression) vs. Phân loại (classification)
- 4 K-means
 - Bài toán
 - Giải
 - Hoc không giám sát
- 5 Cân bằng đô 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
 - Uớc lương (estimator)
 - Cân bằng độ lệch phương sai
- 6 Q&A và Extra ■ Q&A

 - Extra
 - (Extra) Giải Kmeans chính xác khó thể nào



Q&A và Extra

Trí tuệ nhân tạo (Artificial Inteligence - AI) Rule-based: Truyền tải thông minh đơn giản

Sơ lược

- Tại sao dùng ML
 - Trí tuệ nhân tạo (Artificial Inteligence AI)
 - Rule-based: Truyền tải thông minh đơn giản
- 2 ML là gì
- Bài toán MI là bài toán tìm hàm số
- Bài toán MI là bài toán tối ưu hoá (ontimization

Cân bằng độ lệch - phương sai (Bias-variance trade-off)

- 🖪 Hồi quy tuyến tính (Linear Regression
- Bài toán
- Già
 - 11-- -4 -:4-- -4+ (C
- Học co giam sat (Supervised Learning
- Hối quy (Regression) vs. Phân loại (classification
- 4 K-mean
 - Bài toár
- Giải
- Hoc không giám sá
- 5 Cân bằng độ lệch phương sai (Bias-variance trade-of
 - I ro lại bai toan noi quy tuyên tinh
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical mode
 - Uốc lượng (estimator)
 - Cân bằng độ lệch phương sai
 - 0&A
 - Evtr
 - (Extra) Giải Kmeans chính xác khó thế nào



Trí tuệ nhân tạo (Artificial Inteligence - AI)

John McCarthy (1955)

... making a machine behave in ways that would be called intelligent if a human were so behaving ...

Marvin Minsky (1968)

... making machines do things that would require intelligence if done by men ...





Trí tuệ nhân tạo (Artificial Inteligence - AI)

■ Thuật toán truyền tải trí thông minh của con người cho máy tính

Ví dụ truyền tải trí thông minh

Bài toán

Viết thuật toán dự đoán email có phải spam hay không (spam filtering)

Giải đơn giản bài toán spam filtering

Nhân xét

- Spam email thường chứa những từ 'quảng cáo', 'nhanh tay', 'miễn phí', v.v.
- Spam email thường chứa link độc hại

Giải đơn giản bài toán spam filtering

Nhân xét

- Spam email thường chứa những từ 'quảng cáo', 'nhanh tay', 'miễn phí', v.v.
- Spam email thường chứa link độc hại

Ví du thuật toán đơn giản

nếu nội dung chứa 'quảng cáo' hay 'nhanh tay' hay 'miễn phí' thì dư đoán 'spam!'

còn không:

nếu nội dung có chứa link độc hại, thì dự đoán 'spam!' còn không, dự đoán 'không spam'

Thuật toán rule-based

- Bắt đầu từ những nhận xét căn bản (rule)
- Xét mọi trường hợp, dùng câu lệnh if-else

Nhận xét thuật toán rule-based

Dễ viết, chỉ phụ thuộc vào rule

- Cần rule cụ thể
- Có thể phải xét rất nhiều trường hợp
- ...

...

Tóm tắt

- Dễ làm, nhưng khó để làm tốt!
- Cách nào đỡ "cưc" hơn? Thuật toán tư tìm rule?



Q&A và Extra

Sơ lược

- Tại sao dùng ML
- Trí tuệ nhân tạo (Artificial Inteligence Al
- Rule-based: Truyền tải thông minh đơn giản
- 2 ML là gì?
 - Bài toán ML là bài toán tìm hàm số
 - Bài toán ML là bài toán tối ưu hoá (optimization)
- Hồi quy tuyến tính (Linear Regression
- Bài toán
- 0:3
- Hoc có giám sát (Supervised Learnin
- Hồi quy (Regression) vs. Phân loại (classification
- 4 K-mean
 - Rài toái
 - (13)
 - Hockhông giám sá
- 5 Cân bằng độ lệch phương sai (Bias-variance trade-of
 - I ro iại bai toàn noi quy tuyên tinn
 - Mô hình thống kê (Statistical model
 - Bài toán ML là bài toán fit statistical mode
 - Uớc lương (estimator)
 - Cân bằng độ lệch phương sa
- O Q A
 - E de de la
 - (Extra) Giải Kmeans chính xác khó thế nào

Liên quan tới bài toán Al

- Bài toán ML ⊂ bài toán Al
- Truyền tải trí thông về *dữ liệu* (data)
- Thuật toán ML cố gắng giải quyết bài toán ML

Bài toán ML là bài toán tìm hàm số Bài toán ML là bài toán tổi ưu hoá (optimization)

Định nghĩa vắn tắt ML

Định nghĩa vắn tắt l

Định nghĩa vắn tắt ML

Định nghĩa vắn tắt l

- Tìm ở đâu?
- 2 Giải thích dữ liêu như thế nào?

Định nghĩa vắn tắt ML

Định nghĩa vắn tắt l

- Tìm ở đâu?
- 2 Giải thích dữ liêu như thế nào?

Tìm ở đâu? - Gia đình hàm (function family)

- Hàm số (function)
 - f: miền xác định (domain) \rightarrow miền giá trị (range)
 - f: biến \mapsto giá trị của biến
- Gia đình hàm: tập hợp nhiều hàm số có cùng tính chất

Bài toán ML là bài toán tìm hàm số Bài toán ML là bài toán tối ưu hoá (optimization)

Ví dụ về hàm số

Ví dụ hàm số, gia đình hàm

■ bình phương : $\mathbb{R} \to \mathbb{R}$: $x \mapsto x^2$

Ví dụ về hàm số

Ví dụ hàm số, gia đình hàm

- bình phương : $\mathbb{R} \to \mathbb{R}$: $x \mapsto x^2$
- phân loại email : {email} → {'spam', 'không spam'}

Ví dụ về hàm số

Ví dụ hàm số, gia đình hàm

- bình phương : $\mathbb{R} \to \mathbb{R}$: $x \mapsto x^2$
- lacktriangle phân loại email : $\{email\} \rightarrow \{\text{'spam'}, \text{'không spam'}\}$
- gia đình hàm tuyến tính (linear) $\{f: x \mapsto ax + b | a, b \in \mathbb{R}\}$

Ví dụ ML

Ví dụ bài toán ML khái quát cho spam filtering

- Cho: gia đình hàm $\mathcal{H} = \{f : \{\text{email}\} \rightarrow \{\text{'spam'}, \text{'không spam'}\}\}$
- Dữ liệu: tập hợp nhiều email nhận được trong năm vừa rồi
- $Bài toán \ ML$: tìm hàm số $f^* \in \mathcal{H}$ giải thích tốt dữ liệu có được

Ví dụ ML

Ví dụ bài toán ML khái quát cho spam filtering

- Cho: gia đình hàm $\mathcal{H} = \{f : \{\text{email}\} \rightarrow \{\text{'spam'}, \text{'không spam'}\}\}$
- Dữ liệu: tập hợp nhiều email nhận được trong năm vừa rồi
- Bài toán ML: tìm hàm số $f^* \in \mathcal{H}$ giải thích tốt dữ liệu có được

Ứng dụng ML cho spam filtering

- Tìm f* bằng cách giải bài toán ML
- Khi có email x mới, phân loại theo $f^*(x)$



Định nghĩa vắn tắt ML

Định nghĩa vắn tắt l

- Tìm ở đâu?
- 2 Giải thích dữ liêu như thế nào?

Định nghĩa vắn tắt ML

Định nghĩa vắn tắt l

- 1 Tìm ở đâu?
- 2 Giải thích dữ liêu như thế nào?

Giải thích dữ liệu thế nào là tốt (trong bài toán spam filtering)?

- Thông tin di chuyển như thế nào trong bài toán ML?
 - Thông tin hiện tại: dữ liệu hiện tại đang có
 - Thông tin mới: thông tin về dữ liệu chưa quan sát
 - Mục đích sau cùng: thông tin về dữ liệu tương lai đúng càng nhiều càng tốt

Giải thích dữ liệu thế nào là tốt (trong bài toán spam filtering)?

- Thông tin di chuyển như thế nào trong bài toán ML?
 - Thông tin hiện tại: dữ liệu hiện tại đang có
 - Thông tin mới: thông tin về dữ liệu chưa quan sát
 - Mục đích sau cùng: thông tin về dữ liệu tương lai đúng càng nhiều càng tốt
- lacksquare Thiết kế hàm mất mát (loss) $L:\mathcal{H} o\mathbb{R}$
 - $\forall f \in \mathcal{H}, L(f)$ đánh giá xem hàm f có đạt được *mục đích sau cùng* hay không.
 - Thường L(f) càng nhỏ thì f càng làm tốt nhiệm vụ



Bài toán ML là bài toán tìm hàm số Bài toán ML là bài toán tối ưu hoá (optimization)

Tối ưu hoá (optimization)

Ví dụ bài toán optimization

Cho: hàm số $f: x \mapsto x^2 - 4x + 4$

 $Tim: x^* \in \mathbb{R}$ sao cho giá trị của $f(x^*)$ nhỏ nhất

Tối ưu hoá (optimization)

Ví dụ bài toán optimization

Cho: hàm số $f: x \mapsto x^2 - 4x + 4$

Tìm: $x^* \in \mathbb{R}$ sao cho giá trị của $f(x^*)$ nhỏ nhất

Bài toán optimization khái quát

Cho: tập hợp A, hàm số $f:A\to\mathbb{R}$

Tìm: $x^* \in A$ sao cho giá trị của $f(x^*)$ nhỏ nhất

Tối ưu hoá (optimization)

Ví dụ bài toán optimization

Cho: hàm số $f: x \mapsto x^2 - 4x + 4$

Tìm: $x^* \in \mathbb{R}$ sao cho giá trị của $f(x^*)$ nhỏ nhất

Bài toán optimization khái quát

Cho: tập hợp A, hàm số $f:A\to\mathbb{R}$

Tìm: $x^* \in A$ sao cho giá trị của $f(x^*)$ nhỏ nhất

Bài toán ML khái quát

Cho: gia đình hàm \mathcal{H} , hàm mất mát (loss) $L: \mathcal{H} \to \mathbb{R}$ *Tìm:* hàm số $f^* \in \mathcal{H}$ sao cho giá trị của $L(f^*)$ nhỏ nhất



ML ⊂ Optimization

Định nghĩa vắn tắt II

Bài toán ML là bài toán tìm hàm số giải thích dữ liệu trong một gia đình hàm nào đó bằng optimization.

Q&A và Extra

Bài toán Giải Học có giám sát (Supervised Learning) Hồi quy (Regression) vs. Phân loại (classification

Sơ lược

- | Tại sao dùng ML
- Trí tuệ nhân tạo (Artificial Inteligence Al
- Rule-based: Truyền tải thông minh đơn giản
- 2 ML là gì?
 - Bài toán MI là bài toán tìm hàm số
 - Bài toán ML là bài toán tối ưu hoá (optimization

Cân bằng độ lệch - phương sai (Bias-variance trade-off)

- Hồi quy tuyến tính (Linear Regression)
 - Bài toán ■ Giải
 - Hoc có giám sát (Supervised Learning)
 - Hồi quy (Regression) vs. Phân loại (classification)
- 4 K-means
 - Bài toái
 - Giải
- Học không giám sát
- 5 Cân bằng độ lệch phương saj (Bias-variance trade-off
 - I ro lai bai toan noi duy tuyen tinn
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical mode
 - Uốc lượng (estimator)
 - Cân bằng độ lệch phương sa
- Eve
- (Extra) Giải Kmeans chính xác khó thế nà

Bài toán Giải Học có giám sát (Supervised Learning) Hồi quy (Regression) vs. Phân loại (classification)

Bài toán nhỏ

- Hoàn cảnh
 - Cho biến *x*, *y*
 - Cho biết hàm số f liên hệ x với y có bậc = 1

Q&A và Extra

- Dữ liêu
 - Nếu x = 1 thì y = 1
 - Nếu x = 2 thì y = 5
- \blacksquare Tim f

Q&A và Extra

Bài toán Giải Học có giám sát (Supervised Learning) Hồi quy (Regression) vs. Phân loại (classification)

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*



Sơ lược Tại sao dùng ML ML là gi? **Hổi quy tuyến tính (Linear Regression)** K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off)

Q&A và Extra

Bài toán

ai tuaii

Học có giám sát (Supervised Learning)

Bài toán 1 chiều

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

Q&A và Extra

Bài toán

ai toan

Học co giam sat (Supervised Learning) Hồi quy (Regression) vs. Phân loại (classification)

Bài toán 1 chiều

- Dữ liêu
 - $[x_i]_{i=1}^n \in \mathbb{R}$
 - $(y_i)_{i=1}^n \in \mathbb{R}$
- Cho
 - $\blacksquare \mathcal{H} := \{f : x \mapsto ax + b | a, b \in \mathbb{R}\}$
 - $L(f) := \sum_{i=1}^{n} (f(x_i) y_i)^2$

Q&A và Extra

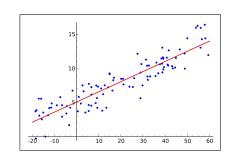
Bài toán

Học có giám sát (Supervised Learning)

Hồi quy (Regression) vs. Phân loại (classification)

Bài toán 1 chiều

- Dữ liêu
 - $[x_i]_{i=1}^n \in \mathbb{R}$
 - $(y_i)_{i=1}^n \in \mathbb{R}$
- Cho
 - $\blacksquare \mathcal{H} := \{f : x \mapsto ax + b | 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 Giải Học có giám sát (Supervised Learning) Hổi quy (Regression) vs. Phân loại (classification

Bài toán *d-*chiều

- Dữ liêu
 - $[x_i]_{i=1}^n \in \mathbb{R}^d$
 - $(y_i)_{i=1}^n \in \mathbb{R}$
- Cho
 - $\blacksquare \mathcal{H} \subseteq \{f : \mathbf{x} \mapsto \langle \mathbf{a}, \mathbf{x} \rangle + b | \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}\}$

Q&A và Extra

- $L(f) = \sum_{i=1}^{n} (f(x_i) y_i)^2$
- 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)



 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)

- 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

- 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
- $Nh\hat{q}n \times \acute{e}t \ 2$: \mathcal{H} đơn giản. L là hàm liên tục, khả vi (theo tham số của \mathcal{H})

- 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)

Bài toán Giải **Học có giám sát (Supervised Learning)** Hồi quy (Regression) vs. Phân loại (classification)

Học có giám sát (Supervised Learning)

- Linear regression là một ví dụ về học có giám sát
- Dữ liệu được cung cấp có "mác", hàm số đầu ra đoán mác cho dữ liệu tương lai / chưa nhìn thấy

Hồi quy (Regression) vs. Phân loại (classification)

- Linear regression là một ví dụ về bài toán hồi quy
 - Tóm tắt thông tin của dữ liệu cũ để làm thông tin dữ liệu mới
- Spam filtering là một ví dụ về bài toán phân loại
 - Dữ liệu cũ được chia thành 2 hay nhiều loại, tìm cách đoán xem dữ liêu mới thuôc loai nào

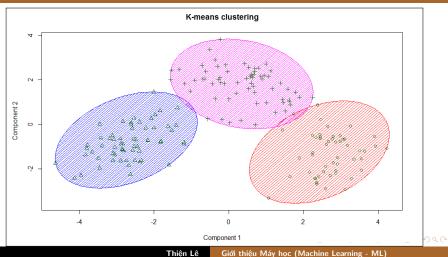
Bài toán Giải Học không giám sá

Sơ lược

- | Tại sao dùng ML
- Trí tuệ nhân tạo (Artificial Inteligence Al
- Rule-based: Truyền tải thông minh đơn giản
- 2 ML là gì?
 - Bài toán ML là bài toán tìm hàm số
 - Bài toán ML là bài toán tối ưu hoá (optimization)
- B Hồi quy tuyến tính (Linear Regression
 - Bài toán
- C:3
 - 11----
- Hoc co glam sat (Supervised Learning)
- Hối quy (Regression) vs. Phân loại (classification
- 4 K-means
 - Bài toán
 - Giải
- Học không giám sát
- 5 Cân bằng độ lệch phương sai (Bias-variance trade-off
 - I ro lại bai toan noi quy tuyen tinn
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical mode
 - Uốc lương (estimator)
 - Cân bằng độ lệch phương sai
- Q&A Va E
 - Evtr
 - (Extra) Giải Kmeans chính xác khó thế nào



Minh hoa



Bài toán Giải Học không giám sát

Bài toán

- Dữ liêu
 - $X = (x_i)_{i=1}^n \in \mathbb{R}^d$
 - 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 $S_{f(x)}$
 - $L(f) = \sum_{i=1}^{k} |S_i| \operatorname{Var}[S_i]$
 - $Var[S_i]$ 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 $S_{f(x)}$
 - $L(f) = \sum_{i=1}^{k} |S_i| \operatorname{Var}[S_i]$
 - $Var[S_i]$ đo phương sai của các điểm x có f(x) = i
- Tìm
 - $f^* \in \mathcal{H}$ sao cho $L(f^*)$ nhỏ nhất



Bài toán **Giải** Học không giám sát

Giải

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

- 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)

- 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.

- 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.



Học không giám sát

- K-means là ví dụ cho học không giám sát
 - Không có khái niệm dữ kiệu mới, cũ; chỉ đi tìm thông tin trong dữ liệu đang có
 - Dữ liệu đang có không được gán 'mác'

Q&A và Extra

Trở lại bài toán hồi quy tuyến tính Mô hình thống kế (Statistical model) Bài toán ML là bài toán fit statistical model Ước lượng (estimator) Cân bằng độ lệch - phương sai

Sơ lược

- Tại sao dùng Ml
- Trí tuệ nhân tạo (Artificial Inteligence Al
- Rule-based: Truyền tải thông minh đơn giản
- 2 ML là gì
 - Bài toán MI là bài toán tìm hàm số
 - Bài toán ML là bài toán tối ưu hoá (optimization
- Hồi quy tuyến tính (Linear Regression)
- Bài toán
- 0:3
 - 11-- -4 -14-- -4+ /
- Hoc co giam sat (Supervised Learning
- Höi quy (Regression) vs. Phän loại (classification
- 4 K-mean
 - Bài toán
 - Dai toai
- Hoc không giám sá
- 5 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
- 0 Q&A va E:
 - Q (C.
 - (Extra) Giải Kmeans chính xác khó thế nà



Q&A và Extra

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
- \blacksquare Tim f

Sơ lược Tại sao dùng ML ML là gi? Hồi quy tuyến tính (Linear Regression) K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off)

Q&A và Extra

Trở lại bài toán hỗi quy tuyến tính Mô hinh thống kê (Statistical model) Bài toán ML là bài toán fit statistical mode Ước lượng (estimator) Cân bằng độ lệch - phương sai

Gia đình hàm trong hồi quy tuyến tính

$$f(x) \neq y$$

Sơ lước Tai sao dùng ML ML là gì? Hồi quy tuyến tính (Linear Regression) K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off)

Trở lai bài toán hồi quy tuyến tính

Gia đình hàm trong hồi quy tuyến tính

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

Gia đình hàm trong hồi quy tuyến tính

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

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

Mô hình thống kê (Statistical model)

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

Trở lại bài toán hồi quy tuyền 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

Mô hình thống kê (Statistical model)

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

Q&A và Extra

■ 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
- $m{\mathbb{H}}=\{x
 ightarrow ax+b+\epsilon|a,b\in\mathbb{R}\}$ là 1 mô hình thống kê

Q&A và Extra

Trở lại bài toán hồi quy tuyến tính Mô hình thống kế (Statistical model) Bài toán ML là bài toán fit statistical model Ước lượng (estimator) Cân bằng độ lệch - phương sai

Tại sao cần biến ngẫu nhiên

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

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

Mô hình mẫu

■ Bài toán ML = optimization + thống kê

Q&A và Extra

Dùng xác suất thống kê dựng mô hình mẫu cho dữ liệu

Định nghĩa vắn tắt III

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

Trở lại bài toán hồi quy tuyến tính Mô hình thống kê (Statistical model) Bài toán ML là bài toán fit statistical mode Ước lượng (estimator) Cân bằng độ lệch - phương sai

Ước lượng tham số (Parameter Fitting)

Q&A và Extra

Định nghĩa vắn tắt III

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?

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

Ước lượng (Estimator)

- Hoàn cảnh
 - Cho biết $y = ax + b + \epsilon$ với tham số a, b nào đó, ϵ là biến ngẫu nhiên
- Dữ liêu
 - Nếu x = 1 thì y = 1
 - Nếu x = 2 thì y = 5
 - Nếu x = 3 thì y = 6
- Tìm f (hoặc tìm a, b)
 - lacksquare 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.

Trở lại bài toán hối quy tuyền 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

Ước lượng không chệch (unbiased estimator)

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

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

Ước lượng không chệch (unbiased estimator)

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

Trở lại bài toán hồi quy tuyến tính Mô hình thống kế (Statistical model) Bài toán ML là bài toán fit statistical mode Ước lượng (estimator) Cân bằng độ lệch - phương sai

Ước lượng hiệu quả (Efficient estimator)

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

Trở lại bài toán hồi quy tuyến tính Mô hình thống kê (Statistical model) Bài toán ML là bài toán fit statistical mode Ước lượng (estimator) Cân bằng độ lệch - phương sai

Ước lượng hiệu quả (Efficient estimator)

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

Q&A và Extra

■ 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

Sơ lược Tại sao dùng ML ML là gi? Hổi quy tuyến tính (Linear Regression) K-means Cân bằng đô lệch - phương sai (Bias-variance trade-off)

Q&A và Extra

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

Cân bằng độ lệch - phương sai

■ $MSE = \text{phương sai } + bias^2$

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

Cân bằng độ lệch - phương sai

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

Q&A và Extra

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

Cân bằng độ lệch - phương sai

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

Q&A và Extra

Q&A (Extra) Giải Kmeans chính xác khó thế n

Sơ lược

- | Tại sao dùng MI
- Trí tuệ nhân tạo (Artificial Inteligence Al
- Rule-based: Truyền tải thông minh đơn giải
- 2 ML là gì
 - Bài toán MI là bài toán tìm hàm số
 - Bài toán ML là bài toán tối ưu hoá (optimization
- Hồi quy tuyến tính (Linear Regression
 - Bài toái
- Gi
- Hoc co gram sat (Supervised Learning)
- Hối quy (Regression) vs. Phân loại (classification
- 4 K-mear
 - Bài toár
 - C:2:
- Hoc không giám sá
- 5 Cân bằng độ lệch phương sai (Bias-variance trade-of
 - I ro iai bai toan noi quy tuyen tinn
 - Mô hình thống kê (Statistical model)
 - Bài toán ML là bài toán fit statistical mode

 - Cân bằng độ lệch phương sa
- 6 Q&A và Extra ■ Q&A
 - Extra
 - (Extra) Giải Kmeans chính xác khó thế nào

Sơ lược Tại sao dùing ML ML là gì? Hồi quy tuyến tính (Linear Regression) K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off) Q&A và Extra

Q&A

Q&A

Q&A (Extra) Giải Kmeans chính xác khó thế nà Sơ lược Tại sao dùng ML ML là gì? Hồi quy tuyến tính (Linear Regression) K-means Cân bằng độ lệch - phương sai (Bias-variance trade-off) Q&A và Extra

Q&A (Extra) Giải Kmeans chính xác khó thế nào

(Extra) Giải Kmeans chính xác khó thế nào