ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN KHOA HỌC MÁY TÍNH

BÁO CÁO KẾT QUẢ NGHIÊN CỨU VÀ THỰC NGHIỆM

Cảnh báo về việc huấn luyện cây quyết định trên dữ liệu từ các mô hình cộng tính: các cận dưới về khả năng khái quát hóa

Nhóm 04 - Lớp 23CLC03

23127004 - Lê Nhật Khôi

23127165 - Nguyễn Hải Đăng

23127271 - Võ Ngọc Bích Trâm

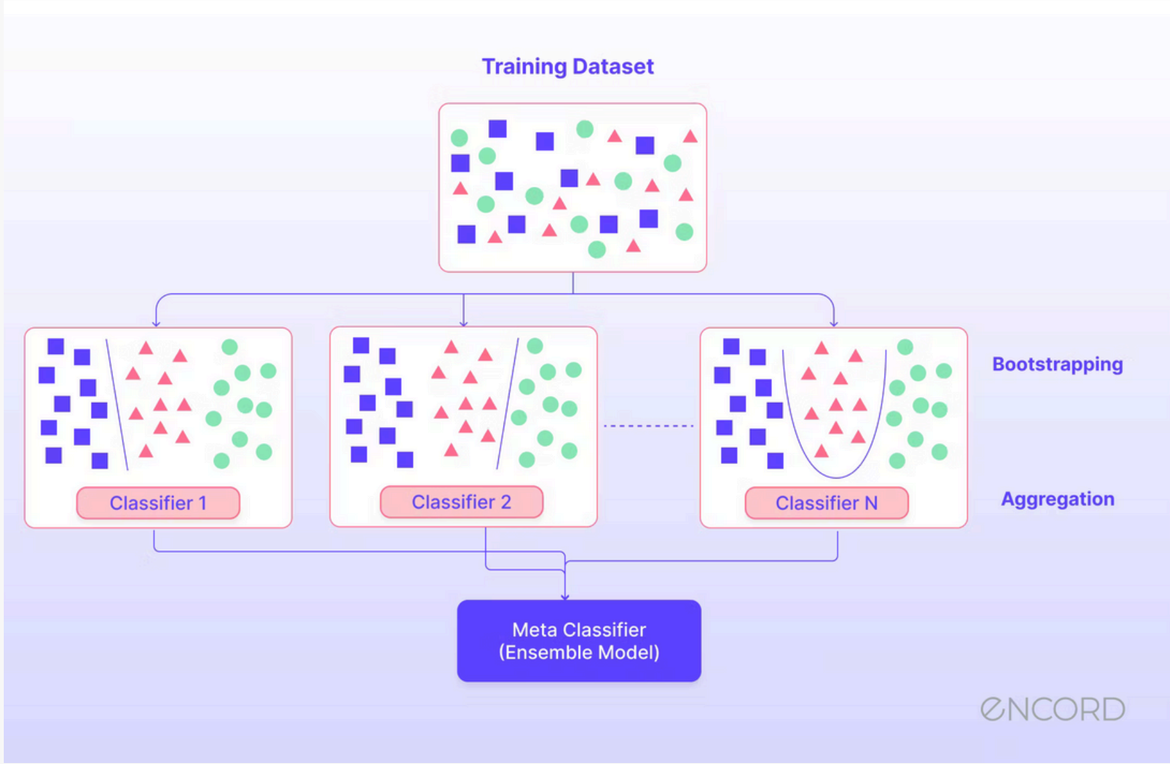
23127486 - Phan Quốc Thịnh

Nhóm 4

Tóm lược

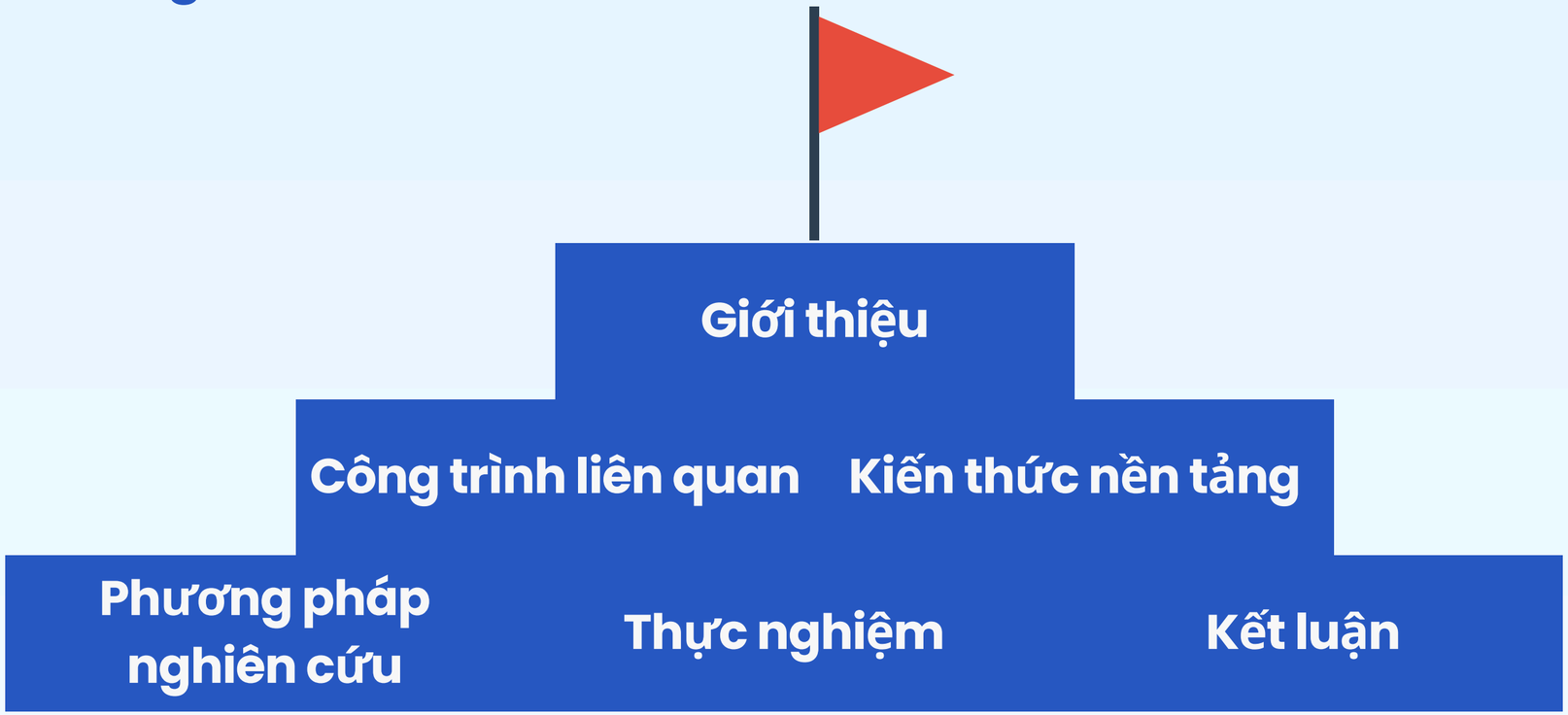
Cây quyết định là thuật toán machine learning quan trọng, có **tính diễn giải cao**, và là nền tảng của các phương pháp **ensemble** như **random forests**. Tuy nhiên, các thuộc tính thống kê của chúng vẫn chưa được hiểu rõ, với các nghiên cứu trước đây chỉ tập trung vào tính nhất quán điểm.

Bài báo này nghiên cứu hiệu suất của cây quyết định trên các **mô hình cộng tính thưa** để làm sáng tỏ thiên kiến quy nạp của thuật toán. Kết quả cho thấy **hiệu suất tổng quát hóa** của chúng **kém hơn tốc độ tối ưu** minimax. Điều này **không do tính tham lam** của thuật toán, mà do việc lấy trung bình trên mỗi lá làm mất khả năng phát hiện cấu trúc toàn cục, một vấn đề được làm rõ bằng cách liên kết với **lý thuyết tốc độ-méo.**



Nhóm 4

Mục lục

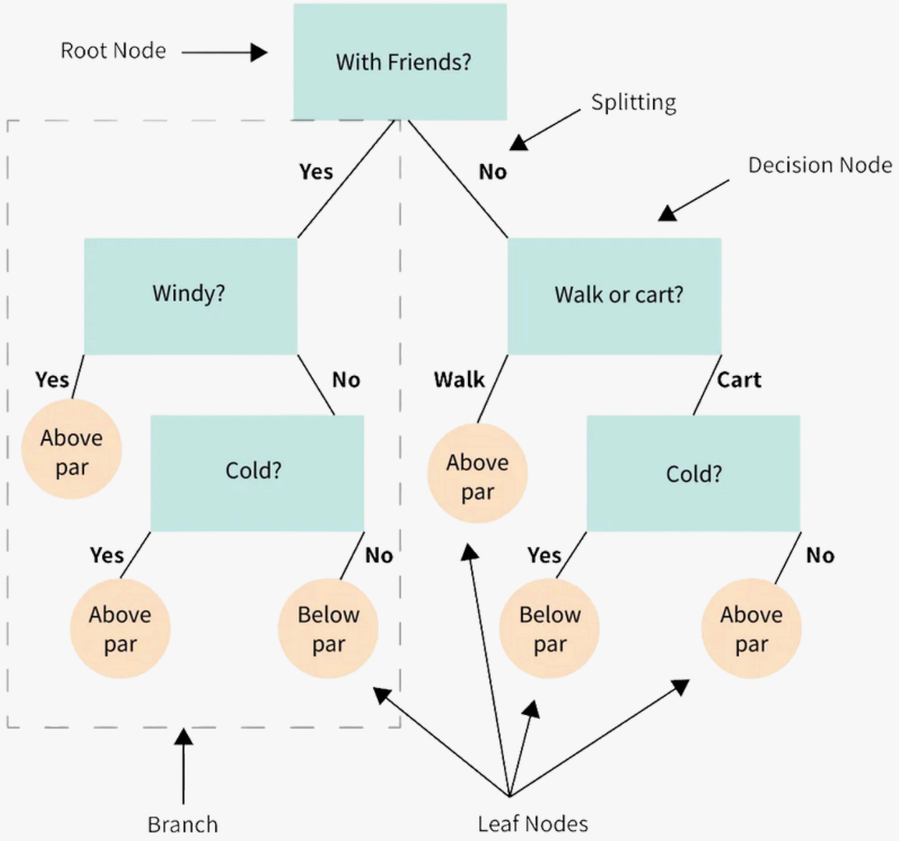


Nhóm 4

Giới thiệu

Nghiên cứu này tập trung vào việc **phân tích hiệu suất** của decision trees khi được áp dụng cho **sparse additive models** (mô hình cộng tính thưa). Đây là những mô hình có độ phức tạp thống kê **thấp** nhưng vẫn duy trì **tính linh hoạt phi tham số** cần thiết để mô tả tốt các tập dữ liệu thực tế.

Mục tiêu chính của nghiên cứu là chứng minh **các cận dưới về generalization error** cho một lớp rộng các thuật toán decision tree, được gọi là **ALA** (axis-aligned partition với leaf-only averaging). Điều này cho phép so sánh hiệu suất của chúng với các thuật toán được thiết kế đặc biệt như **backfitting**, từ đó hiểu được liệu **inductive bias** của decision trees có thể khai thác hoàn toàn cấu trúc có trong additive models hay không



Các nghiên cứu liên quan

Công trình & Hướng nghiên cứu	Đóng góp chính	Ưu điểm	Khoảng trống nghiên cứu	Hướng phát triển
Nghiên cứu Tính nhất quán điểm (Biau, 2012; Wager & Athey, 2018)	Chứng minh rằng thuật toán CART có thể hội tụ về hàm mục tiêu (tính nhất quán điểm) trong hồi quy tổng quát.	- Đặt nền móng lý thuyết đầu tiên cho Cây Quyết Định. - Cung cấp đảm bảo toán học về khả năng học.	- Cần thay đổi thuật toán gốc (lá co về 0). - Không phân tích tốc độ hội tụ.	Cần phân tích cho CART nguyên bản và định lượng hiệu quả.
Tính nhất quán trên Mô hình Cộng tính (Scornet et al., 2015)	Chứng minh được tính nhất quán cho CART nguyên bản.	- Thực tiễn cao hơn. - Giả định mô hình cộng tính giúp chứng minh khả thi.	- Vẫn chỉ dừng ở tính nhất quán. - Giả định mô hình còn đơn giản.	Mở rộng cho mô hình cộng tính thưa và nghiên cứu tốc độ hội tụ.
Tính nhất quán trong bối cảnh thưa (Klusowski 2020, 2021)	Mở rộng kết quả Scornet chứng minh CART vẫn nhất quán trong không gian nhiều chiều thưa.	- Tăng tính liên quan thực tiễn. - Cho thấy CART thích ứng tự nhiên với tính thưa.	- Vẫn chỉ dừng ở tính nhất quán. - Chưa phân tích hiệu quả thống kê.	Cần định lượng sai số tổng quát hóa và so với cận dưới lý thuyết
Nghiên cứu về sự Bất nhất quán (Tang et al., 2018)	Chỉ ra các trường hợp Honest RF không nhất quán.	- Hiếm có kết quả phủ định. - Chỉ ra giới hạn tồn tại	- Điều kiện đặc thù, không tổng quát. - Không định lượng mức độ lỗi.	Cần phân tích tổng quát hơn để tìm nguyên nhân gốc rễ.
Bài báo nghiên cứu (Tan et al., 2021)	Thiết lập cận dưới lý thuyết cho sai số tổng quát hóa của CART trên mô hình cộng tính.	- Lần đầu định lượng tốc độ hội tụ. - Xác định nguyên nhân là "leaf-only averaging". - Kết nối với lý thuyết tốc độ-biến dạng.	- Chứng minh toán học mới cho Honest Trees. - Tập trung vào mô hình cộng tính.	Mở rộng cho Random Forest, Gradient Boosting

Kiến thức nền tảng

Trên mỗi phân vùng, sẽ có một hàm ước lượng đánh giá thông qua trung bình lá(leaf-only averaging):

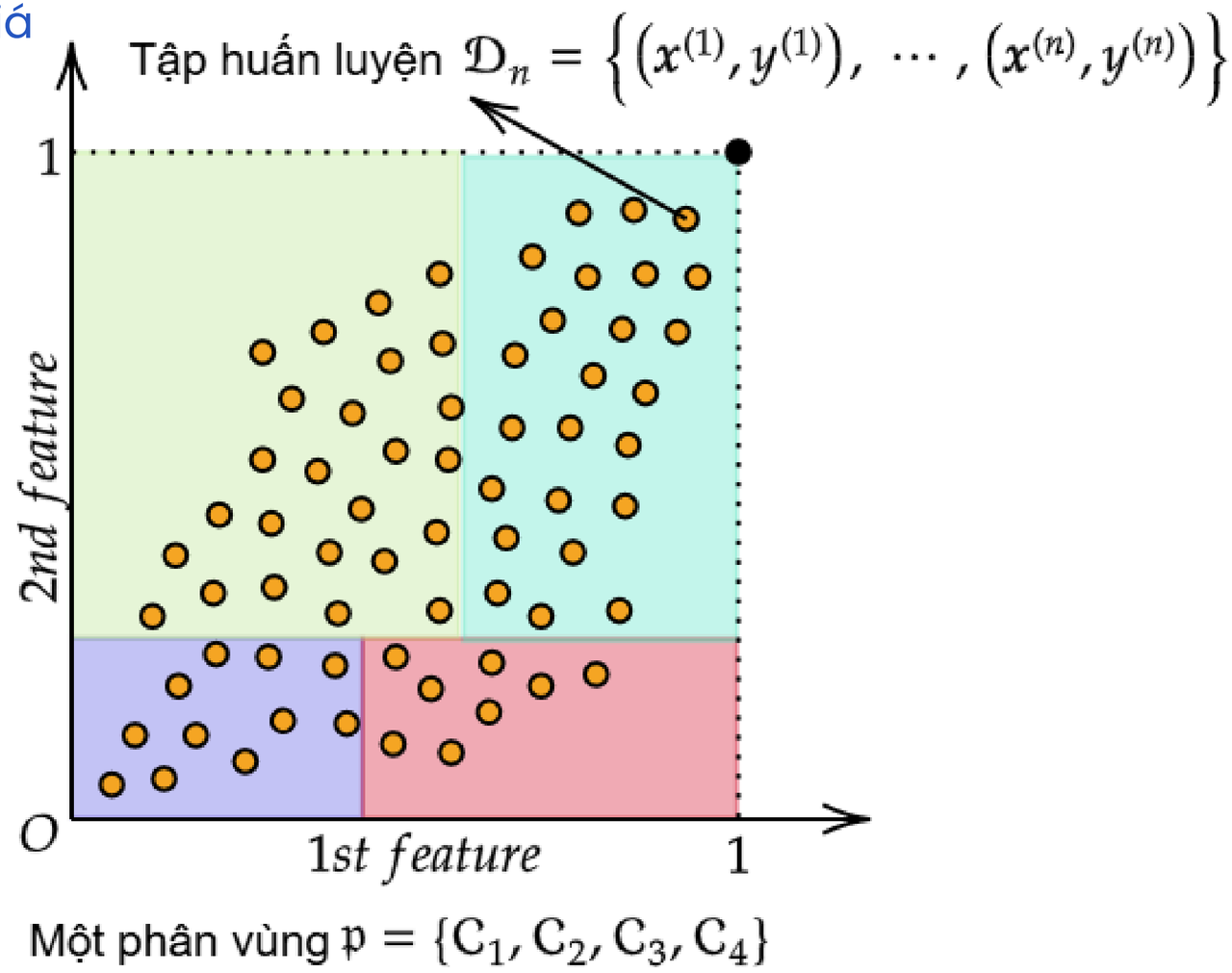
$$\hat{f}(\mathbf{x}; \mathfrak{p}, \mathcal{D}_n) := \sum_{\mathcal{C} \in \mathfrak{p}} \left(\frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} y^{(i)} \right) \mathbf{1}\{\mathbf{x} \in \mathcal{C}\}$$

Rủi ro bình phương/Rủi ro tổng quát:

$$\mathcal{R}(\hat{f}) := \mathbb{E}_{\mathbf{x} \sim \nu} \left\{ \left(\hat{f}_n(\mathbf{x}) - f(\mathbf{x}) \right)^2 \right\}$$

Rủi ro kỳ vọng tối ưu::

$$\mathcal{R}^*(f, \nu, n) := \inf_{\mathfrak{p}} \mathbb{E} \left\{ R(\hat{f}(-; \mathfrak{p}, \mathcal{D}_n) \right\}$$



Phương pháp nghiên cứu 01

Với một phép phân hoạch $\mathfrak{p}$ hợp lệ và một tập huấn luyện $\mathcal{D}_n$ rủi ro kỳ vọng thỏa mãn cận dưới:

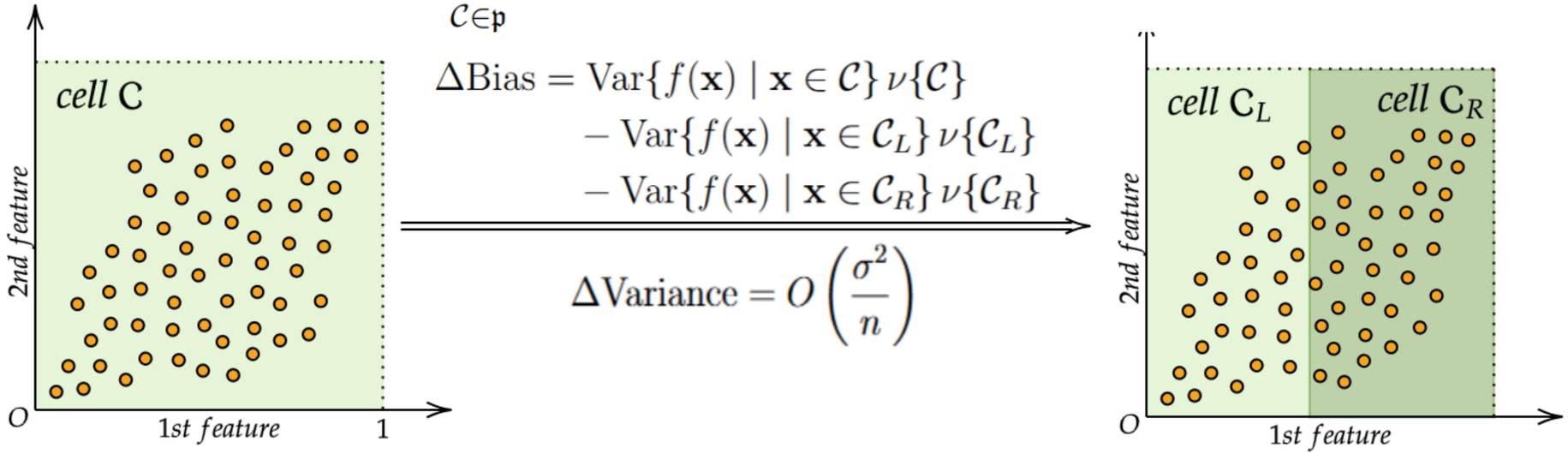
$$\mathbb{E}\mathcal{R}(\hat{f}(-;\mathfrak{p},\mathcal{D}_n)) \geq \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} + \frac{|\mathfrak{p}|\sigma^2}{2n}$$

Và cận trên:

$$\mathbb{E}\mathcal{R}(\hat{f}(-;\mathfrak{p},\mathcal{D}_n)) \leq 7\sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} + \frac{6|\mathfrak{p}|\sigma^2}{n} + E(\mathfrak{p})$$

Với:

$$E(\mathfrak{p}) = \sum_{\mathcal{C}\in\mathfrak{p}} \mathbb{E}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}^2 (1-\nu\{\mathcal{C}\})^n \nu\{\mathcal{C}\}$$



Phương pháp nghiên cứu 02

Trong lý thuyết thông tin, khi bạn muốn mã hóa một nguồn dữ liệu X thành một dạng rút gọn, bạn thường phải chấp nhận một mức sai số nhất định giữa dữ liệu gốc và dữ liệu nén. Từ đó tác giả đã xây dựng một cầu nối giữa phân rã Bias-Variance với lý thuyết tốc độ biến dạng:

- **Độ biến dạng:** $\delta(p;\beta) := \mathbb{E}_{(\mathbf{x},\hat{\mathbf{x}})\sim p}\{\|\mathbf{x}-\hat{\mathbf{x}}\|_\beta^2\}$ đo lường sai số trung bình bình phương giữa một điểm dữ liệu gốc và dữ liệu tái tạo liên hệ với **Bias** thông qua bất đẳng thức:

$$\sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} \geq \frac{\delta(p;\boldsymbol{\beta})}{2}$$

- **Tốc độ nén:** số bit trung bình cần để mã hóa dữ liệu sao cho độ biến dạng không vượt quá một ngưỡng D liên hệ với **Số lá** thông qua bất đẳng thức với hàm trade-off tối ưu $R(D;p_{\mathbf{x}},\boldsymbol{\beta}) := \inf_{p_{\hat{\mathbf{x}}|\mathbf{x}}} I(\mathbf{x};\hat{\mathbf{x}})$

$$\log|\mathfrak{p}| \geq R(\delta(p;\beta);\nu,\beta)$$

Từ đó, thay vào bất đẳng thức trên và lấy infinium lên toàn bộ phân vùng thì ta thu được kết quả(bổ đề):

$$\mathcal{R}^*(f,\nu,n) \geq \frac{1}{2}\inf_{D>0}\left\{D+\frac{\sigma^2 2^{R(D;\nu,\boldsymbol{\beta})}}{n}\right\}$$

Phương pháp nghiên cứu 02

Nếu đặt thêm điều kiện mỗi biến hiệp phương sai tuân theo một phân phối biên ν_0
Giả sử tồn tại một tập hợp con các tọa độ $S \subset [d]$ với kích thước s sao cho $|\beta_j| \geq \beta_0 \forall j \in S$
Tác giả chứng minh được:

$$D + \frac{\sigma^2 2^{R(D;\nu,\beta)}}{n} \geq D + \frac{\sigma^2 2^{h(\nu_0)}}{n} \left(\frac{s\beta_0^2}{2\pi e D} \right)^{s/2}$$

Bài toán đã chuyển đổi bài toán tổ hợp trên các phân vùng thành một bài toán tối ưu hóa với biến liên tục:

$$\varphi(D) = D + kD^{-s/2}$$

Từ đó thu được kết quả cho trường hợp f tuyến tính:

$$\mathcal{R}^*(f, \nu, n) \geq s 2^{\frac{2s}{s+2} h(\nu_0) - 2} \left(\frac{\beta_0^2}{\pi e} \right)^{s/(s+2)} \left(\frac{\sigma^2}{n} \right)^{2/(s+2)}$$

Tốc độ hội tụ của mô hình là $\Omega(n^{-2/s+2})$ cũng là tốc độ hội tụ minimax tối ưu cho việc ước lượng một hàm trơn tổng quát trong không gian s chiều

Phương pháp nghiên cứu 03

Chặn dưới cho đặc trưng [0,1]^d

Với mô hình hồi quy và hàm mục tiêu f là mô hình cộng tính, giả sử rằng không gian hiệp phương sai là khối lập phương đơn vị $[0,1]^d \forall j = 1, \cdots, d$. Cho $I_1, I_2, \cdots, I_d \subset [0,1]$ là các đoạn con và giả sử tồn tại tập con các chỉ số $S \subset [d]$ với kích thước s sao cho $\min_{t \in I_j} |\phi_j'(t)| \geq \beta_0 > 0$

Đặt $\mathcal{K} = \{\mathbf{x} : x_j \in I_j, j = 1, \cdots, d\}$. Giả sử rằng ν là phân phối liên tục với mật độ q và ký hiệu $q_{min} = \min_{\mathbf{x} \in \mathcal{K}} q(\mathbf{x})$
Khi đó rủi ro kỳ vọng tối ưu bị chặn dưới bởi:

$$R^*(f, \nu, n) \geq s\mu(\mathcal{K}) \left(\frac{\beta_0^2 q_{min}}{12} \right)^{s/(s+2)} \left(\frac{\sigma^2}{4n} \right)^{2/(s+2)}$$

Tốc độ hội tụ tổng quát theo bất đẳng thức này vẫn là $\Omega(n^{-2/s+2})$
Trong khi đó, tốc độ minimax tối ưu cho lớp mô hình cộng tính thưa là nhanh hơn đáng kể $\max\left\{ \frac{s \log(d/s)}{n}, \frac{s}{n^{2/3}} \right\}$

Điều này cho ta thấy cây quyết định đã hoàn toàn thất bại trong việc khai thác cấu trúc cộng tính đơn giản hơn của dữ liệu

Phương pháp nghiên cứu 03

Chặn dưới cho đặc trưng Boolean {0,1}^d

Giả sử các hiệp phương sai độc lập với $x_j \sim Ber(\pi)$ thỏa $0 \leq \pi \leq 1/2$. Giả sử tồn tại tập con các chỉ số $S \subset [d]$ với số lượng là s sao cho $\beta_j \geq \beta_0 > 0$. Khi đó rủi ro kỳ vọng tới ưu bị chặn dưới bởi:

$$R^*(f, \nu, n) \geq \frac{s\beta_0^2}{2}\left(1 - \left(\frac{2e^s n\beta_0^2}{2^{sH(\pi)}\sigma^2}\right)^{\frac{1}{s-1}}\right)$$

Về mặt hình thái, chặn dưới cho mô hình cộng tính Bool hoàn toàn khác với chặn dưới cho mô hình cộng tính khối lập phương đơn vị. Nguyên nhân là do trong không gian Boolean, một cây quyết định có thể, về mặt lý thuyết, phân chia không gian cho đến khi mỗi lá chỉ chứa một điểm dữ liệu duy nhất tức là Bias bằng không!

Điều này có nghĩa sự đánh đổi Bias-Variance (Distortion-Rate) có bản chất khác đi: không còn là sự đánh đổi giữa một số hạng tiến về 0 và một số hạng bùng nổ ra vô cùng

Thiết lập thực nghiệm

- Thông số CPU được sử dụng cho thực nghiệm

Thông số	Giá trị
Kiến trúc	9 (x64)
Bộ nhớ đệm L2	5120 KB
Bộ nhớ đệm L3	8192 KB
Tốc độ tối đa (Max Clock Speed)	2701 MHz
Tên CPU	2419 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz
Số Core	4
Số luồng (Number of Logical Processors)	8

Thiết lập thực nghiệm

- Danh sách các hyperparameters (siêu tham số)

Siêu tham số	Ý nghĩa
n_train	Danh sách kích thước của các tập huấn luyện, lần lượt là: 100, 250, 500, 750, 1000, 1500, 2000, 2500.
n_test	Kích thước tập kiểm thử, cố định là 500.
d	Số lượng các đặc trưng, mặc định là 50.
beta	Hệ số cho các đặc trưng của mô hình tuyến tính, mặc định là 1.
sigma	Mức độ nhiễu của mô hình huấn luyện, mặc định là 0.1.
sparsity	Các mức độ thưa của mô hình, hay số lượng các đặc trưng quan trọng, mặc định 10 và 20.
n_avg	Số lần lặp của mỗi tập huấn luyện để lấy trung bình kết quả.

Thiết lập thực nghiệm

- Dữ liệu được mô phỏng thông qua 3 loại mô hình với đặc trưng thưa

Mô hình tuyến tính với đặc trưng nhị phân:

$$y = \sum_{j=1}^{s} \beta_j x_j + \epsilon, \quad \mathbf{x} = (x_1, \ldots, x_s) \in \{0,1\}^s, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Mô hình tuyến tính với đặc trưng liên tục:

$$y = \sum_{j=1}^{s} \beta_j x_j + \epsilon, \quad \mathbf{x} = (x_1, \ldots, x_s) \in Unif[0,1)^s, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Mô hình tổng bình phương với đặc trưng liên tục:

$$y = \sum_{j=1}^{s} \beta_j x_j^2 + \epsilon, \quad \mathbf{x} = (x_1, \ldots, x_s) \in Unif[0,1)^s, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Phương pháp đánh giá, so sánh

Chỉ số đánh giá: Ta sử dụng Lỗi bình phương trung bình (Mean Squared Error) trên tập kiểm tra.

Các phương pháp so sánh:

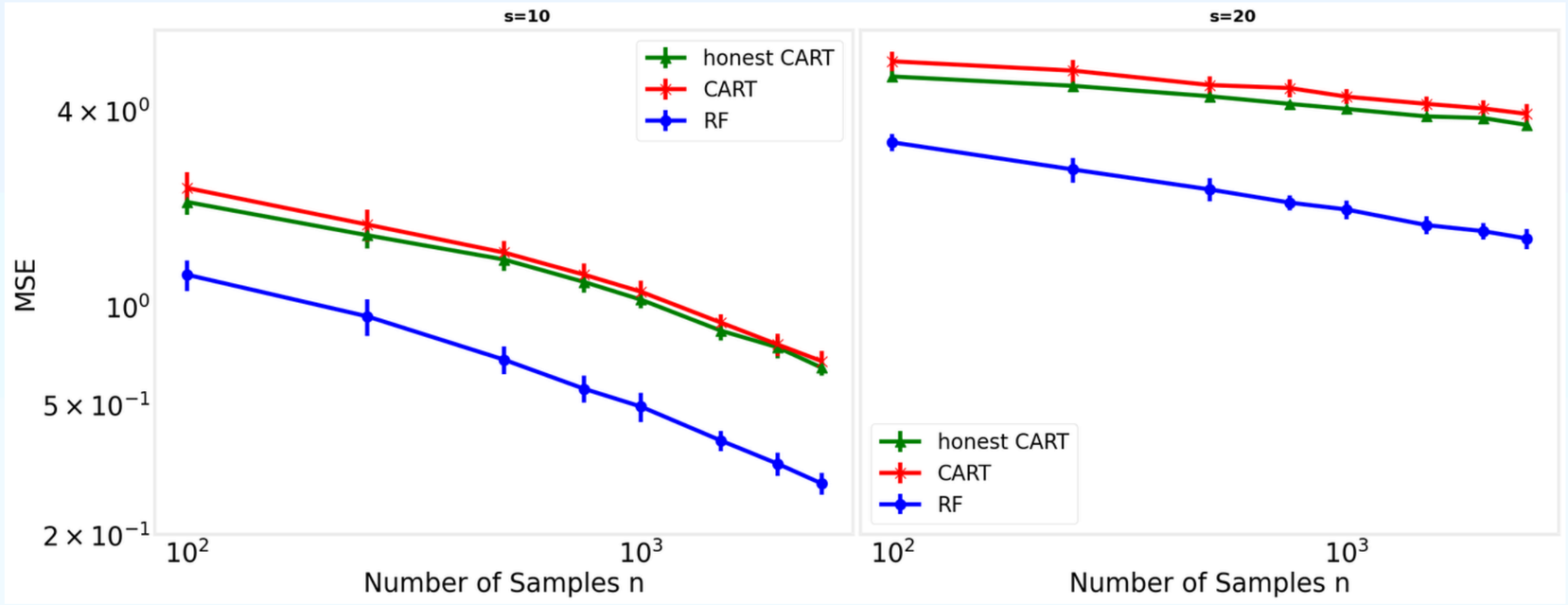
– Honest CART: chia dữ liệu thành hai phần, một phần để xây cây và một phần để gán giá trị lá cho bước dự đoán.

– Dishonest CART: dùng toàn bộ dữ liệu để xây cây và gán lá.

– Random Forest (RF): rừng gồm nhiều CART.

Nhóm 4

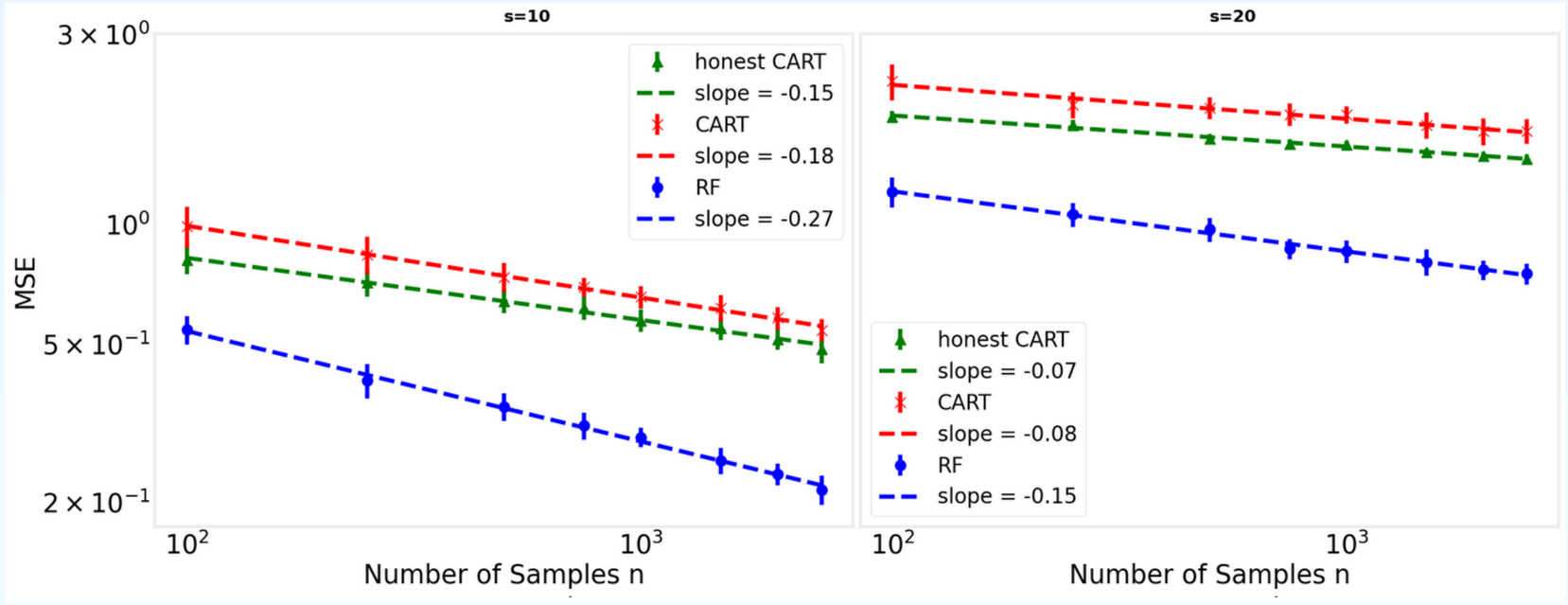
Thực nghiệm & Kết quả



Biểu đồ so sánh MSE của các mô hình trên dữ liệu tuyến tính thưa với đặc trưng Boolean.

Nhóm 4

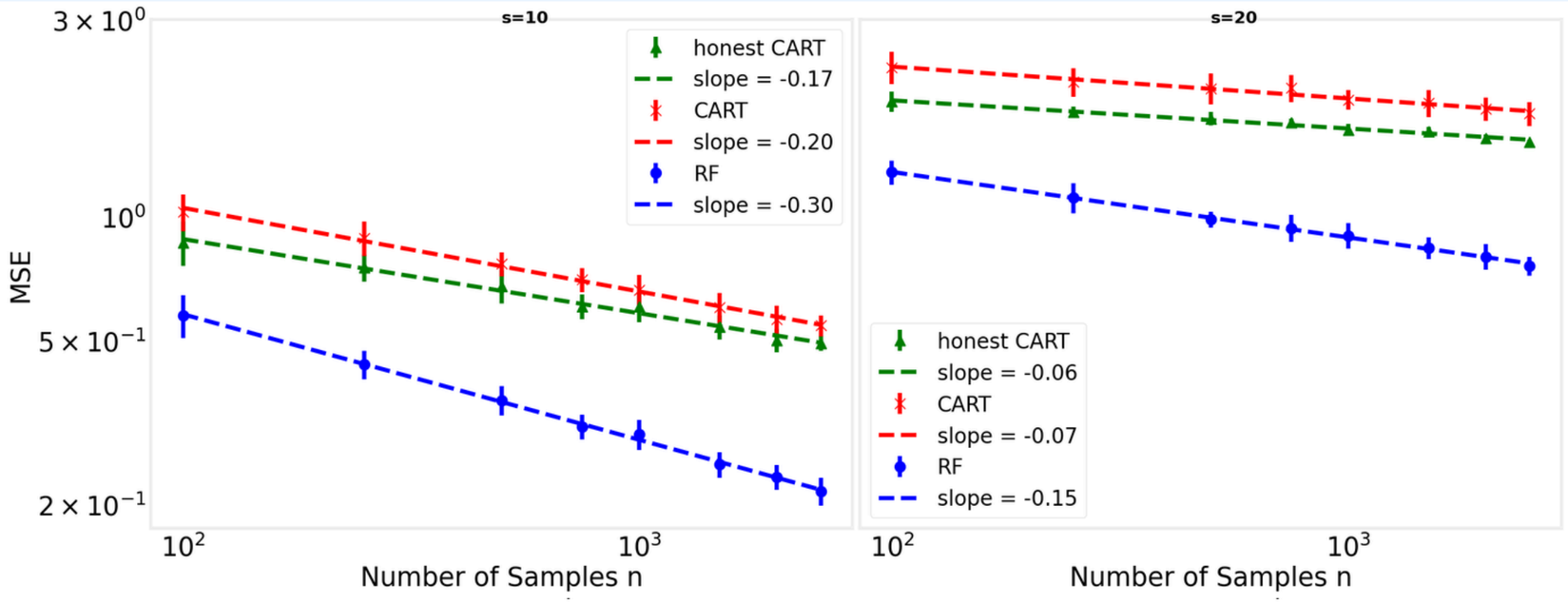
Thực nghiệm & Kết quả



Biểu đồ log-log so sánh MSE của các mô hình trên dữ liệu tuyến tính thưa với đặc trưng liên tục (Uniform).

Nhóm 4

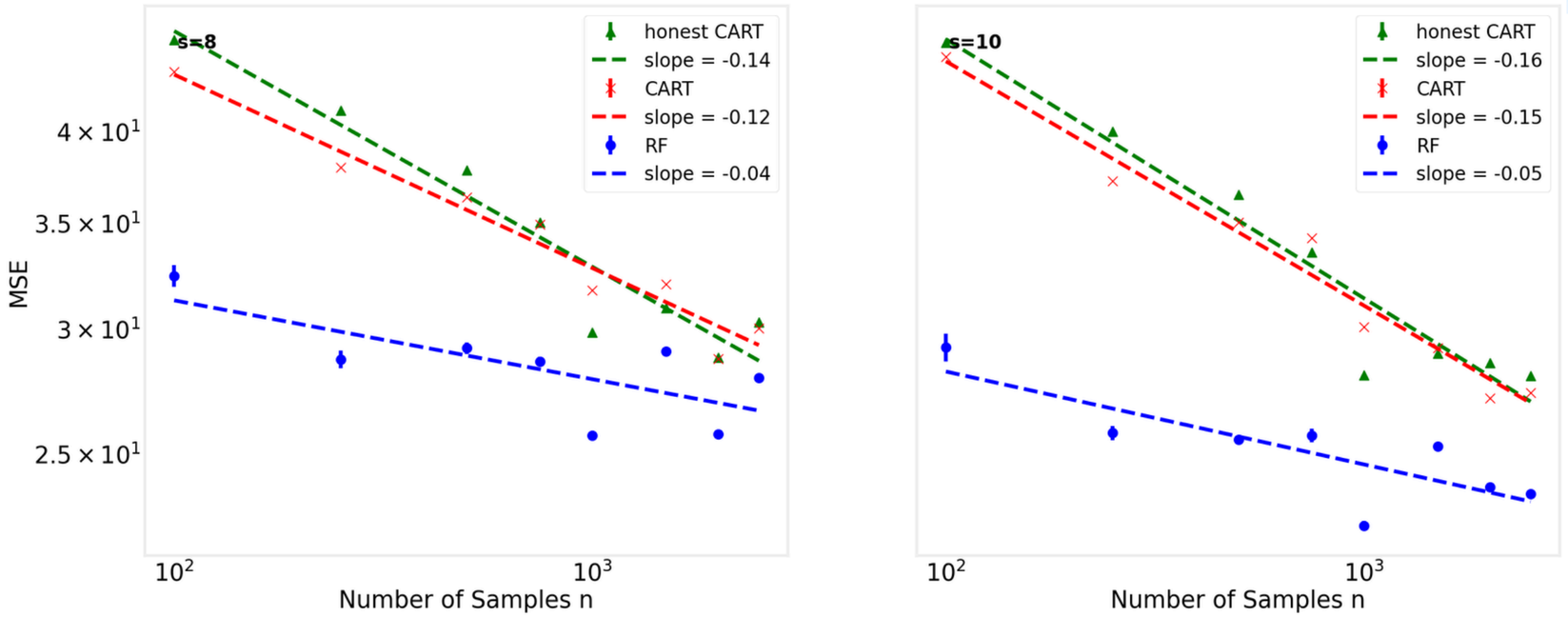
Thực nghiệm & Kết quả



Biểu đồ log-log so sánh MSE của các mô hình trên dữ liệu tổng bình phương thưa với đặc trưng liên tục.

Nhóm 4

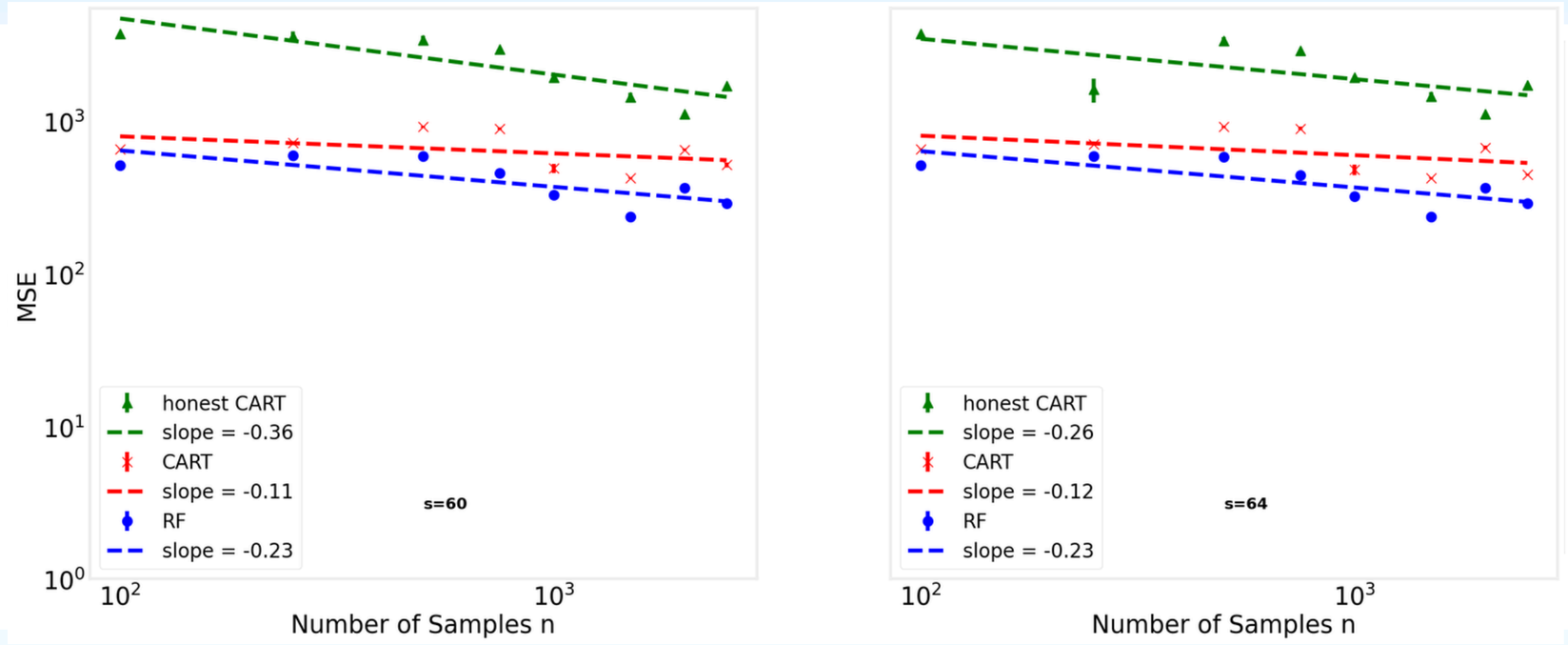
Thực nghiệm & Kết quả



Biểu đồ log-log so sánh MSE của các mô hình trên dữ liệu thực tế (15 đặc trưng)

Nhóm 4

Thực nghiệm & Kết quả



Biểu đồ log-log so sánh MSE của các mô hình trên dữ liệu thực tế (81 đặc trưng).

Các yếu tố ảnh hưởng

Ảnh hưởng của kích thước mẫu: Khi số lượng mẫu tăng lên, MSE của các mô hình đều giảm.

Ảnh hưởng của số lượng đặc trưng quan trọng: MSE tăng khi độ thưa (sparsity) tăng.

So sánh hiệu năng

Random Forest (RF): RF là mô hình có hiệu suất tốt nhất (MSE thấp nhất).

Honest CART và Dishonest CART: Trong các trường hợp mô phỏng, honest CART cho kết quả tốt hơn một chút (MSE thấp hơn) so với CART truyền thống.

Thực nghiệm & Kết quả

Dữ liệu mô phỏng

- RF là mô hình có hiệu suất tốt nhất (MSE thấp nhất).

- Honest CART có hiệu suất tốt hơn CART thường.

- RF có độ dốc âm nhất.

Dữ liệu thực tế

- RF là mô hình có hiệu suất tốt nhất (MSE thấp nhất).

- CART thường có hiệu suất tốt hơn honest CART.

- Honest CART có độ dốc âm nhất.

⇒ **Nhận xét:** Điểm khác biệt trên có thể được lý giải là do bộ dữ liệu thực tế có độ nhiễu lớn hơn so với dữ liệu mô phỏng và chưa đảm bảo tính chất cộng tính mạnh. Tuy nhiên, tốc độ hội tụ của các mô hình vẫn hợp lý với giới hạn lý thuyết đã được nêu ra trước đó.

Thảo luận & Kết luận

Điểm Mạnh và Hạn chế của Phương pháp Tiếp cận

Theo nhóm tác giả

- **Điểm mạnh:**
  - Tiếp cận này **phân tích một thuật toán như một đối tượng nghiên cứu sơ cấp, giúp làm rõ thiên kiến quy nạp** của nó dưới các mô hình dữ liệu khác nhau.
  - **Thích hợp với dữ liệu hiện đại** (nhiều đặc trưng)
- **Hạn chế:**
  - Chỉ mới chạm đến bề nổi của việc điều tra thiên kiến quy nạp của các thuật toán cây, RF và gradient boosting.
  - Việc tập trung vào "rừng cây" cũng làm **giảm khả năng giữ được tính giải thích** của mô hình.

Phát hiện và Đóng góp Chính

- **Giới hạn lý thuyết cho Cây ALA**
- **CART** có **thiên kiến quy nạp** chống lại cấu trúc toàn cục do đặc tính chỉ sử dụng giá trị trung bình ở lá
- **Random Forest (RF)** không bị ảnh hưởng bởi những giới hạn trên phù hợp với luận điểm về việc giảm phương sai nhờ sự đa dạng của các cây
- **Tốc độ của RF** vẫn chậm hơn đáng kể so với tốc độ tối ưu (minimax rates) cho các mô hình cộng tính thưa

Thảo luận & Kết luận

Kết quả Thực nghiệm dữ liệu thực tế

Theo thực nghiệm quan sát

- **RF**: hiệu suất **ổn định, vượt trội** hơn CART → nhất quán với dữ liệu mô phỏng.
- **Honest CART < Dishonest CART** → khác biệt so với kết quả mô phỏng.
- **Honest CART** có **độ dốc âm lớn nhất** → mô hình thay đổi khi áp dụng dữ liệu thực.
- **Nguyên nhân:** dữ liệu thực nhiều nhiễu, không tuân thủ chặt chẽ giả định cộng tính.
- **Kết luận:**
  - Các mô hình vẫn có tốc độ hội tụ hợp lý, phù hợp giới hạn lý thuyết.
  - Thực nghiệm trên dữ liệu thực tế là cần thiết để kiểm chứng khả năng ứng dụng và tổng quát hóa.

Đề xuất cho dữ liệu thực nghiệm

- **Đa dạng dữ liệu:** chọn bộ dữ liệu có đặc trưng thưa, chứa nhiễu, và mức độ cộng tính khác nhau.
- **Thiết kế thí nghiệm:** kiểm tra ảnh hưởng của kích thước mẫu, mức độ thưa, và tỉ lệ “honesty”.
- **Đánh giá:** so sánh MSE, tốc độ hội tụ, và tính ổn định giữa Honest CART, Dishonest CART, và RF.
- **Kết luận:** cần kiểm chứng trên dữ liệu thực để hiểu rõ giới hạn lý thuyết và khả năng ứng dụng.

Cảm ơn
&
Hỏi đáp

Nhóm 04