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

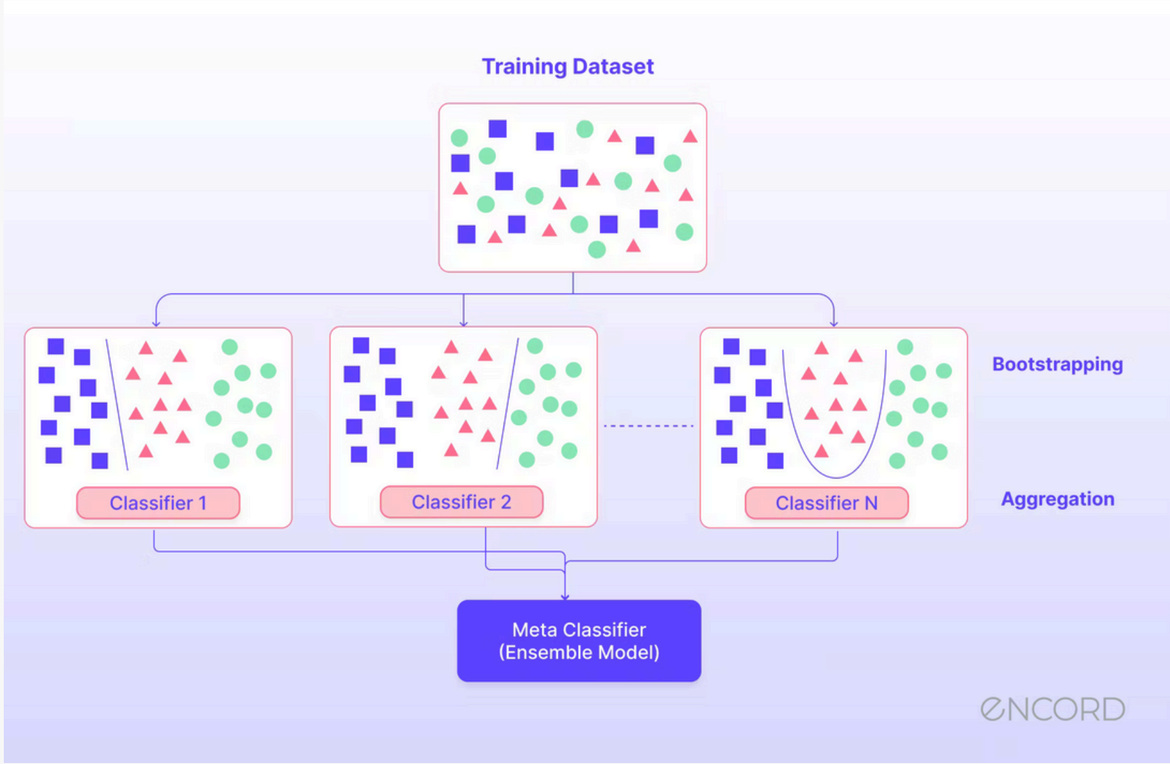
23127486 - Phan Quốc Thịnh

Nhóm 4

Tóm lược

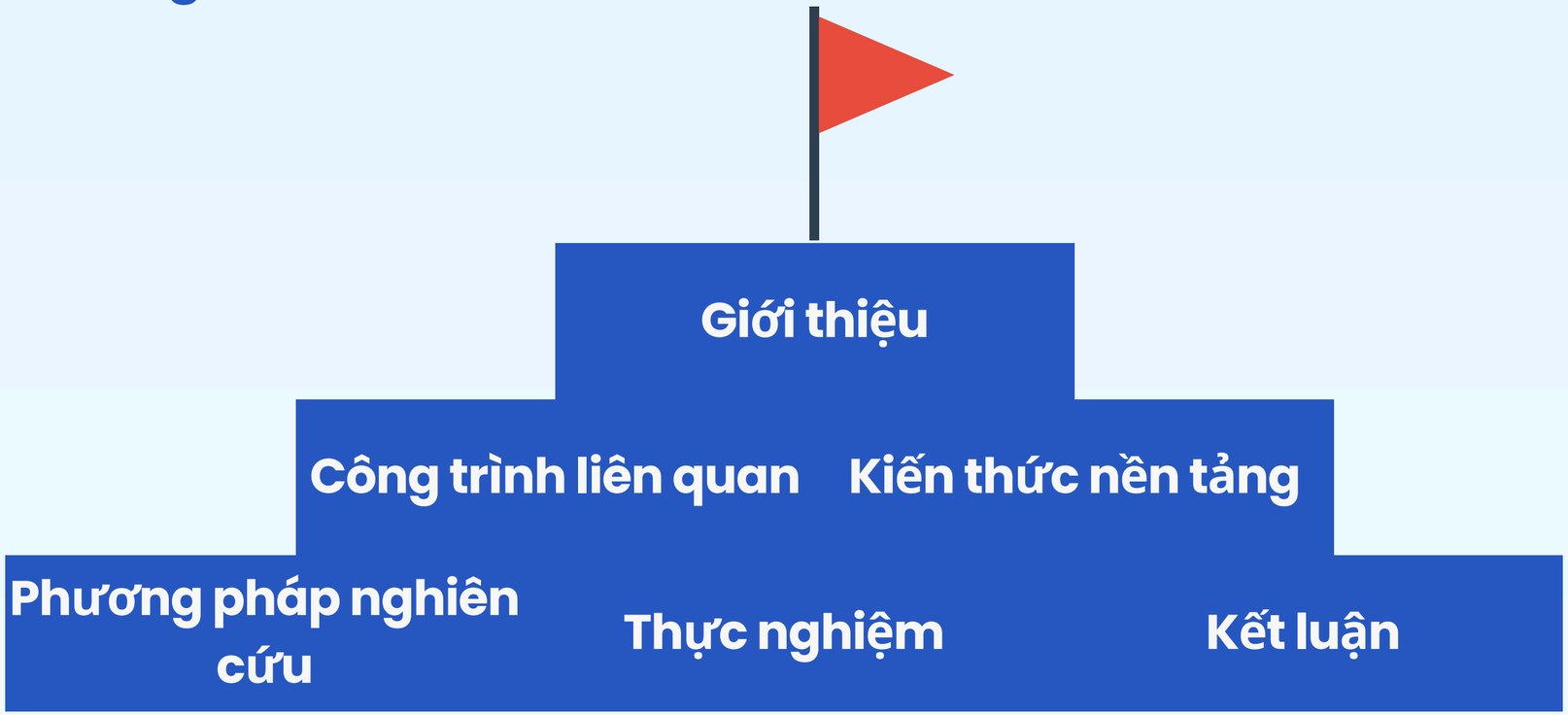
Cây quyết định là thuật toán machine learning quan trọng, có **tính diễn giải cao**, và là nền tảng của các phương pháp **ensemble** như **random forests**. Tuy nhiên, các thuộc tính thống kê của chúng vẫn chưa được hiểu rõ, với các nghiên cứu trước đây chỉ tập trung vào tính nhất quán điểm.

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

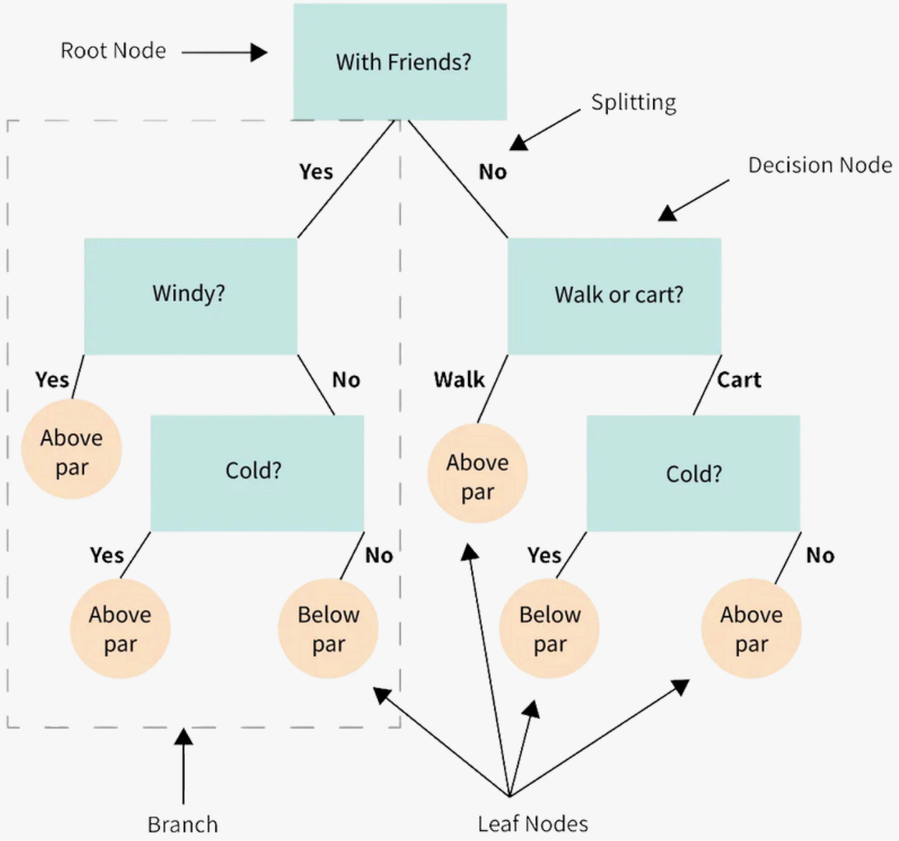


Nhóm 4

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



Nhóm 4

Các nghiên cứu liên quan

Minimax Rates #1

Raskutti et al. (2012) đã thiết lập minimax rate cho **sparse additive models**, tỉ lệ như

$$\max\left\{\frac{s\log(d/s)}{n}, \frac{s}{n^{2/3}}\right\}$$

Stone (1982) đã thiết lập l2 minimax rate cho **nonparametric estimation** của các hàm C^1 trong s chiều là

$$\Omega\left(n^{-\frac{2}{s+2}}\right)$$

Boolean features #2

Syrgkanis & Zampetakis (2020) đã chứng minh **generalization upper bounds** cho CART trong setting khác nhau, xem xét Boolean features và áp đặt giả định submodularity trên conditional mean function.

Feature Importance #3

Scornet (2020) đã quay lại thiết lập additive model và tính toán các công thức bất biến tiệm cận (asymptotic explicit formulas) cho **mean impurity decrease (MDI)** feature importance score.

Kiến thức nền tảng

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

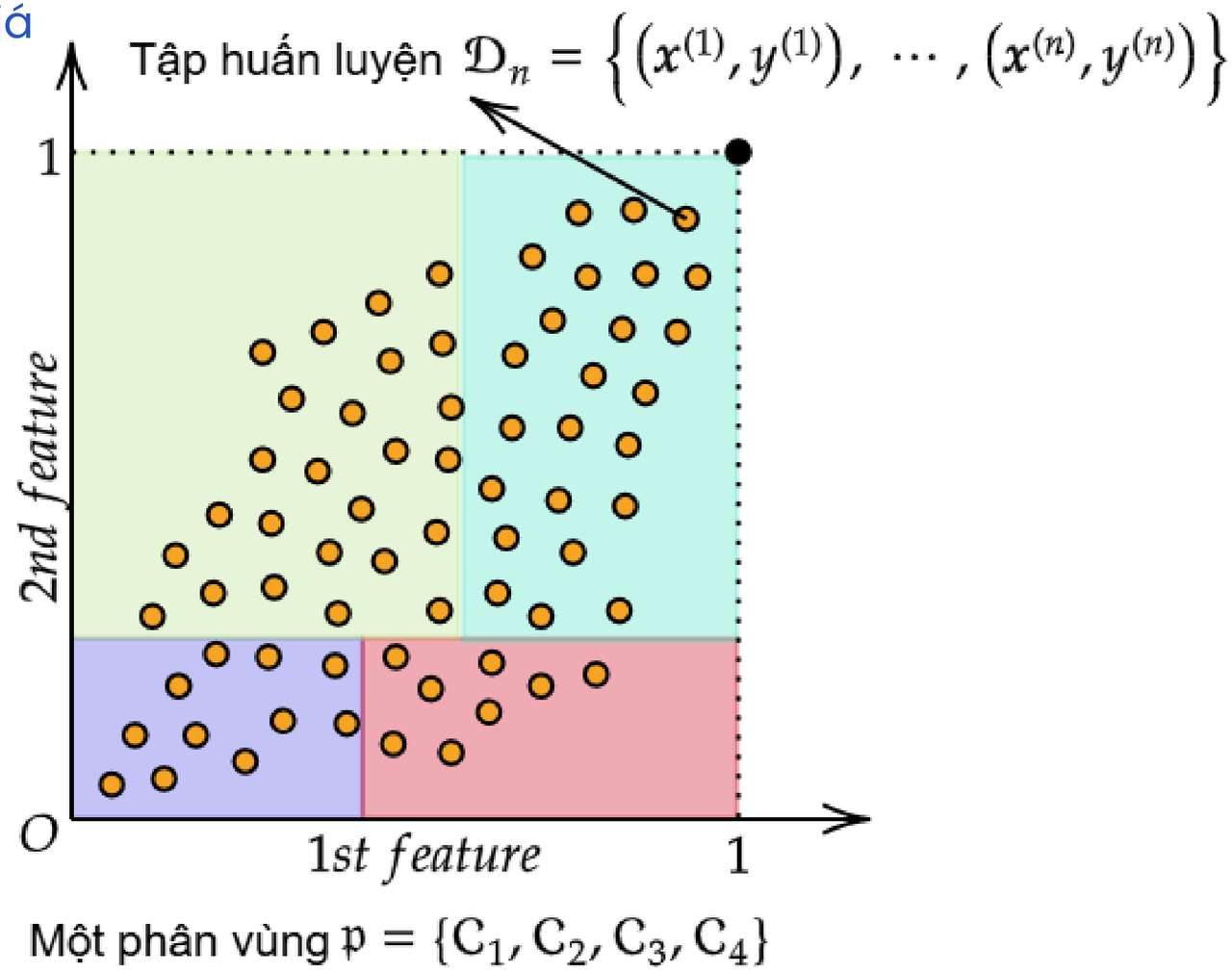
$$\hat{f}(\mathbf{x}; \mathfrak{p}, \mathcal{D}_n) := \sum_{\mathcal{C} \in \mathfrak{p}} \left(\frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} y^{(i)} \right) \mathbf{1}\{\mathbf{x} \in \mathcal{C}\}$$

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:

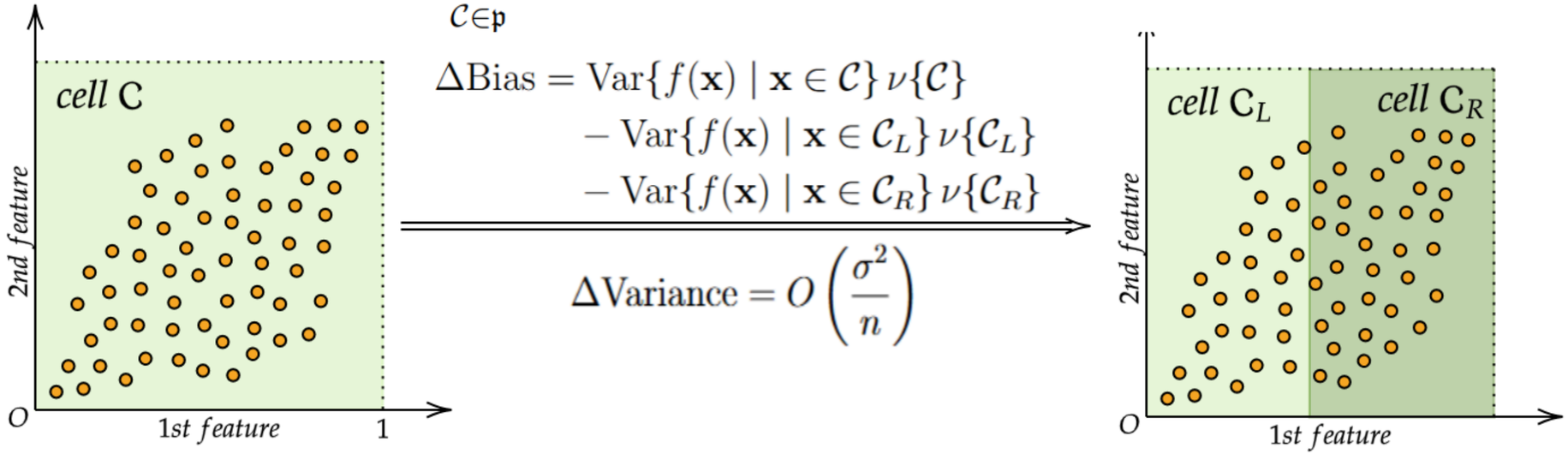
$$\mathbb{E}\mathcal{R}(\hat{f}(-;\mathfrak{p},\mathcal{D}_n)) \geq \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} + \frac{|\mathfrak{p}|\sigma^2}{2n}$$

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:

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

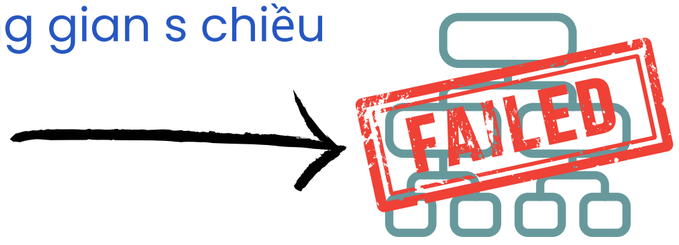
Bài toán đã chuyển đổi bài toán tổ hợp trên các phân vùng thành một bài toán tối ưu hóa với biến liên tục:

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

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 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

- 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:

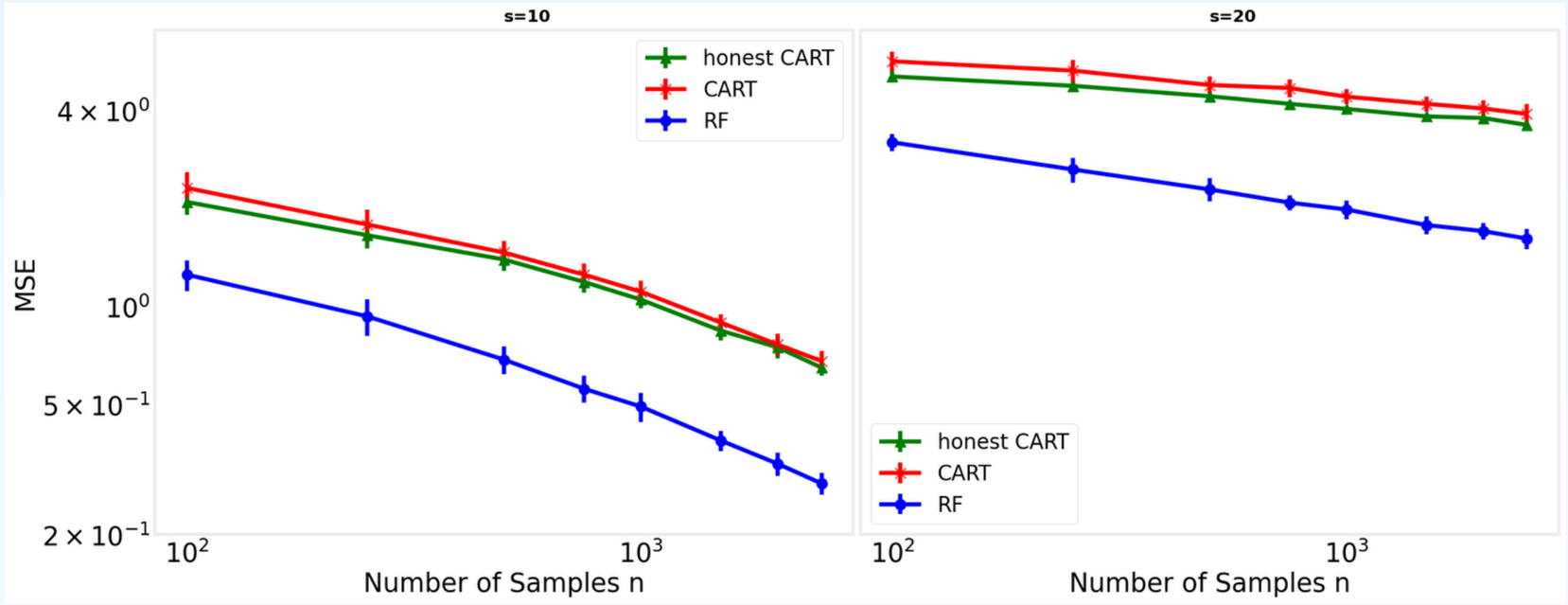
– Honest CART: chia dữ liệu thành hai phần, một phần để xây cây và một phần để gán giá trị lá cho bước dự đoán.

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

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

Nhóm 4

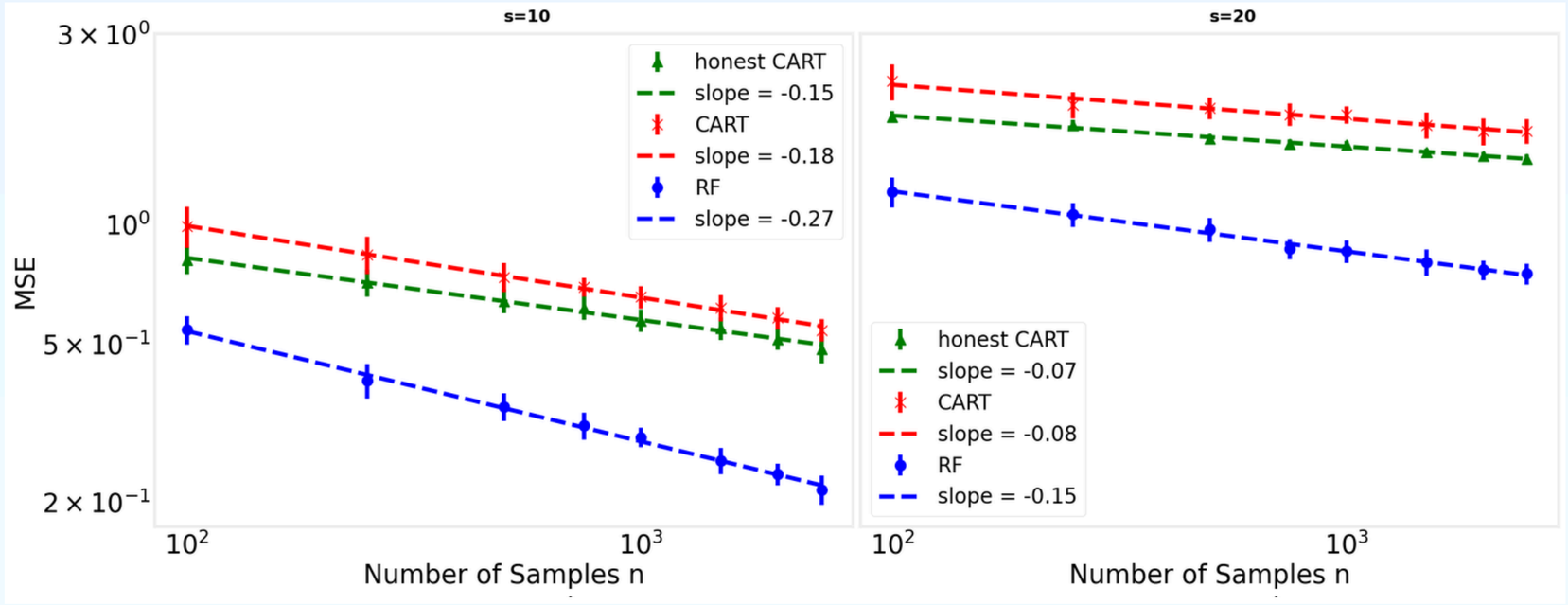
Thực nghiệm & Kết quả



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

Nhóm 4

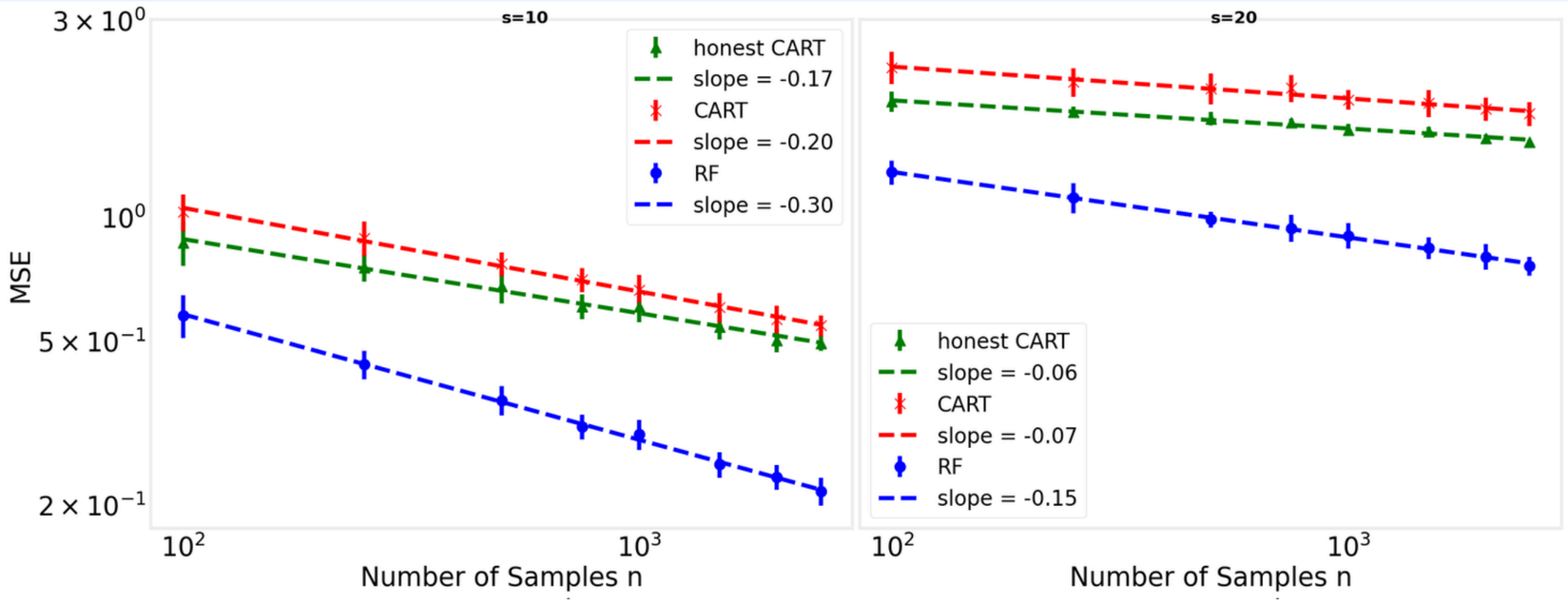
Thực nghiệm & Kết quả



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.

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. Song, đây là mô hình có độ dốc âm nhất → tổng hợp thông tin từ các mẫu huấn luyện một cách hiệu quả hơn.

Honest CART và Dishonest CART: Trong hầu hết các trường hợp, honest CART cho kết quả tốt hơn một chút (MSE thấp hơn) so với CART truyền thống. Sử dụng tập honest để gán giá trị cho lá giúp giảm độ chệch (bias) so với mô hình Dishonest CART.

Thảo luận & Kết luận

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

Điểm Mạnh và Hạn chế của Phương pháp Tiếp cận

- **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**
- **Hạn chế:**
  - Nghiên cứu 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.

Ý nghĩa và Định hướng Nghiên cứu Tương lai

- **Cải tiến thuật toán:** Sửa đổi các thuật toán cây để dễ dàng học được cấu trúc toàn cục hơn
- **Xác định thuật toán phù hợp** dựa trên phân tích thiên kiến quy nạp
- **Định hướng tương lai:**
  - Tiếp tục **phân tích thiên kiến quy nạp** cho các mô hình hồi quy khác.
  - Áp dụng cho các thuật toán học máy khác.
  - Khám phá phương pháp kết hợp **cây** với **phương pháp tuyến tính** hoặc **cộng tính.**

Cảm ơn
&
Hỏi đáp

Nhóm 04