ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

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

Báo cáo Bài báo Nghiên cứu khoa học
CSC14003 - Cơ sở Trí tuệ nhân tạo

Sinh viên thực hiện:

23127004 - Lê Nhật Khôi

23127165 - Nguyễn Hải Đăng

23127271 - Võ Ngọc Bích Trâm

23127486 - Phan Quốc Thịnh

Giảng viên hướng dẫn:

thầy Bùi Tiến Lên

thầy Lê Nhựt Nam

thầy Võ Nhật Tân

Thành phố Hồ Chí Minh, Ngày 3 tháng 9 năm 2025

THÔNG TIN NHÓM

Danh sách thành viên và phân công

STT	MSSV	Họ tên	Vai trò	Phân công	Phần trăm
01	23127004	Lê Nhật Khôi	Thành viên	Thư ký Viết Báo cáo Chương 3, 4 Thiết kế slide	100%
02	23127165	Nguyễn Hải Đăng	Nhóm trưởng	Tổ chức framework làm việc Viết Báo cáo Chương 1, 2 Thiết kế slide	100%
03	23127271	Võ Ngọc Bích Trâm	Thành viên	Viết Báo cáo Chương 6 Thực nghiệm đề tài Chỉnh sửa video	100%
04	23127486	Phan Quốc Thịnh	Thành viên	Viết Báo cáo Chương 5 Thực nghiệm đề tài Quay video	100%

Đánh giá mức độ hoàn thành

Yêu cầu đồ án	**Chi tiết**	**Phần trăm**
Báo cáo nghiên cứu	Nội dung chính xác Cấu trúc và tổ chức Phân tích phản biện Ngôn ngữ và trình bày	100%
Kiểm tra thực nghiệm	Thiết kế thực nghiệm Phân tích và ghi chép	100%
Thuyết trình nhóm	Nội dung kỹ thuật Kỹ năng giao tiếp	100%

MỤC LỤC

CHƯƠNG 1

GIỚI THIỆU

1.1. Bối cảnh và Động lực nghiên cứu

Trong lĩnh vực học máy, **Cây quyết định (Decision Trees)** giữ một vị thế trung tâm và có vai trò kép. Thứ nhất, chúng là một trong số ít các mô hình có khả năng diễn giải cao, cho phép con người hiểu được logic đằng sau các dự đoán. Đặc tính này làm cho chúng trở nên vô giá trong các lĩnh vực yêu cầu sự minh bạch và trách nhiệm giải trình cao như chẩn đoán y khoa hay cấu trúc hình sự. Thứ hai, chúng chính là nền tảng để xây dựng nên các thuật toán *ensemble* mạnh mẽ bậc nhất hiện nay như Random Forests và Gradient Boosting, những phương pháp thường xuyên đạt hiệu suất hàng đầu trên một loạt các bài toán dự đoán phức tạp.

Tuy nhiên, có một nghịch lý đáng chú ý: mặc dù được ứng dụng rộng rãi, sự hiểu biết về các thuộc tính thống kê cốt lõi của cây quyết định vẫn còn rất hạn chế. Các công trình lý thuyết trước đây chủ yếu tập trung vào việc chứng minh **tính nhất quán (consistency)**—một tiêu chuẩn đảm bảo rằng thuật toán sẽ hội tụ về mô hình thực sự nếu được cung cấp vô hạn dữ liệu. Dù quan trọng nhưng tính nhất quán chỉ là một "ngưỡng" cơ bản và không cho chúng ta biết thuật toán hoạt động hiệu quả ra sao với lượng dữ liệu hữu hạn trong thực tế, tức là nó không trả lời được câu hỏi về **tốc độ hội tụ (rate of convergence)**.

Nghiên cứu này đề xuất một hướng tiếp cận khác, sâu sắc hơn: thay vì phân tích trong một bối cảnh tổng quát, chúng em chủ trương nghiên cứu hiệu suất của cây quyết định trên các **mô hình sinh dữ liệu (generative models)** có cấu trúc cụ thể. Phương pháp này cho phép chúng em thăm dò và làm sáng tỏ **thiên kiến quy nạp (inductive bias)** của thuật toán—những giả định ngầm mà nó áp đặt lên dữ liệu để có thể tổng quát hóa. Cụ thể, chúng em chọn **mô hình cộng tính thưa (sparse additive models)** làm đối tượng nghiên cứu. Đây là một lớp mô hình

$$f(x) = \sum_j \varphi_j(x_j),$$

có cấu trúc tương đối đơn giản (là sự mở rộng của mô hình tuyến tính) nhưng vẫn đủ linh hoạt để mô tả các mối quan hệ phi tuyến. Việc chọn lớp mô hình này như một phép kiểm tra: nếu một thuật toán mạnh như cây quyết định lại hoạt động kém hiệu quả trên một mô hình có cấu trúc rõ ràng như vậy, điều đó sẽ tiết lộ một yếu điểm cơ bản và sâu sắc của chính thuật toán đó.

1.2. Mục tiêu và ý nghĩa nghiên cứu

Mục tiêu cốt lõi của nghiên cứu này là tiến hành một phân tích định lượng chặt chẽ nhằm **chứng minh một cách toán học về sự kém hiệu quả về mặt thống kê của cây quyết định** khi áp dụng cho dữ liệu có cấu trúc cộng tính. Để thực hiện điều này, nghiên cứu đặt ra các mục tiêu cụ thể sau:

1. **Thiết lập một cận dưới lý thuyết (theoretical lower bound)** cho lỗi tổng quát hóa bình phương (squared generalization error) của một lớp thuật toán cây quyết định rộng, được gọi là cây **ALA (Axis-Aligned partition with Leaf-only Averaging)**. Lớp này bao hàm hầu hết các thuật toán cây hồi quy phổ biến, bao gồm cả CART.

2. **So sánh cận dưới này với tốc độ tối ưu (optimal minimax rate)** mà bất kỳ thuật toán nào cũng có thể đạt được cho lớp mô hình cộng tính thưa. Sự so sánh này sẽ định lượng hóa mức độ kém hiệu quả của cây quyết định.

Ý nghĩa của nghiên cứu này mang tính đa chiều. Về mặt **thực tiễn**, nó đưa ra một "cảnh báo" quan trọng cho các nhà khoa học dữ liệu: cần thận trọng khi áp dụng các mô hình dựa trên cây cho những bài toán mà dữ liệu có thể ẩn chứa cấu trúc cộng tính. Về mặt **lý thuyết**, nghiên cứu này cung cấp một cái nhìn sâu sắc chưa từng có về bản chất của cây quyết định. Bằng cách chỉ ra chính xác điểm yếu của chúng, nghiên cứu mở ra những hướng đi mới để cải tiến các thuật toán dựa trên cây, giúp chúng khai thác cấu trúc dữ liệu một cách hiệu quả hơn.

1.3. Tổng quan về đóng góp của bài báo

Nghiên cứu này mang lại những đóng góp khoa học quan trọng, làm thay đổi hiểu biết của chúng ta về cây quyết định:

1. **Chứng minh một Cận dưới Lý thuyết Mới:** Đóng góp chính là việc chứng minh rằng lỗi

tổng quát hóa của cây ALA có tốc độ hội tụ tệ hơn đáng kể so với tốc độ minimax tối ưu. Cụ thể, kết quả cho thấy cây quyết định không thể thoát khỏi **"lời nguyền của số chiều" (curse of dimensionality)** ngay cả khi mô hình cơ bản là thưa, một phát hiện trái với trực giác thông thường.

2. **Phân tích sâu sắc về Thiên kiến Quy nạp:** Bài báo chỉ ra rằng nguyên nhân của sự kém hiệu quả này **không phải do tính "tham lam" (greediness)** của thuật toán—một chỉ trích phổ biến nhưng có phần chưa chính xác. Thay vào đó, "thủ phạm" thực sự là một thuộc tính cố hữu và cơ bản hơn: cơ chế **"chỉ lấy trung bình trên lá" (leaf-only averaging)**. Cơ chế mang tính cục bộ này làm mất khả năng của cây trong việc phát hiện các cấu trúc và xu hướng toàn cục của dữ liệu.

3. **Xây dựng Cầu nối với Lý thuyết Thông tin:** Để đạt được các kết quả trên, các tác giả đã phát triển một bộ công cụ kỹ thuật hoàn toàn mới, thiết lập một mối liên hệ độc đáo giữa bài toán ước lượng bằng cây quyết định và **Lý thuyết Tốc độ–Méo (Rate-Distortion Theory)**. Họ đã diễn giải thành công thiên vị (bias) của mô hình như là độ méo (distortion), và phương sai (variance) như là tốc độ (rate), một phương pháp luận đầy sáng tạo.

4. **Gợi ý các Hướng cải tiến Thực tiễn:** Từ việc xác định được nguyên nhân cốt lõi, bài báo đề xuất các hướng đi tiềm năng để cải thiện thuật toán cây, chẳng hạn như tích hợp các cơ chế học cấu trúc toàn cục như **co cụm phân cấp (hierarchical shrinkage)** hoặc **tổng hợp toàn cục (global pooling)**.

1.4. Cấu trúc báo cáo

Báo cáo này được tổ chức theo cấu trúc sau:

- Chương 2 sẽ trình bày tổng quan về các công trình liên quan;
- Chương 3 sẽ giới thiệu kiến thức nền tảng cần thiết;
- Chương 4 sẽ mô tả chi tiết phương pháp nghiên cứu;
- Chương 5 sẽ phân tích các thí nghiệm và kết quả;
- Chương 6 sẽ đưa ra kết luận và định hướng nghiên cứu tương lai.

CHƯƠNG 2

CÁC CÔNG TRÌNH LIÊN QUAN

2.1. Tổng quan về lý thuyết Decision Trees

Nghiên cứu về tính chất thống kê của decision trees có một lịch sử tương đối ngắn so với tầm quan trọng của chúng trong machine learning. Trong bối cảnh hồi quy, một số nghiên cứu được trích dẫn nhiều nhất đã tập trung vào việc chứng minh các đảm bảo **tính nhất quán điểm** cho CART.

Các nghiên cứu về tính nhất quán điểm:

- Biau (2012) [1] và Wager & Athey (2018) [17] đã chứng minh tính nhất quán điểm cho CART, nhưng buộc phải thay đổi tiêu chuẩn tách để đảm bảo kích thước của các ô phân hoạch học được sẽ co về 0.
- Các nghiên cứu này gặp hạn chế về tính thực tiễn do cần *điều chỉnh* thuật toán gốc.

Tiến bộ với mô hình cộng tính:

- Scornet et al. (2015) [3] đã chứng minh kết quả tính nhất quán đầu tiên cho thuật toán CART không bị thay đổi bằng cách thay thế mô hình hồi quy phi tham số hoàn toàn bằng *mô hình cộng tính*.
- Giả định sinh này đơn giản hóa tính toán bằng cách tránh một số phụ thuộc phức tạp giữa các phép tách có thể tích lũy trong quá trình quá trình tách đệ quy.

2.2. Nghiên cứu về mô hình cộng tính thưa

Mở rộng cho mô hình cộng tính thưa:

- Klusowski (2020, 2021) [9][10] đã mở rộng phân tích cho mô hình cộng tính thưa, chỉ ra rằng khi hàm conditional mean thực sự chỉ phụ thuộc vào một tập con cố định s covariates, CART vẫn nhất quán ngay cả khi tổng số covariates được phép tăng theo hàm mũ của sample size.
- Tính thích ứng với sparsity này phần nào giảm thiểu *curse of dimensionality* và giải thích một phần tại sao CART và Random Forests thường được ưa chuộng trong thực tế so với k-nearest neighbors.

Tầm quan trọng của additive models:

- Additive models, như những tổng quát tự nhiên của linear models, đồng thời có độ phức tạp thống kê thấp và tính linh hoạt phi tham số đủ để mô tả tốt một số tập dữ liệu thực tế.
- Nếu component functions không quá phức tạp, additive models có các khía cạnh của interpretability.
- Chúng đã tích lũy một văn liệu thống kê phong phú (Hastie & Tibshirani, 1986; Sadhanala & Tibshirani, 2019)[7][11].

2.3. Các nghiên cứu về cận và tỉ lệ hội tụ

Gap trong nghiên cứu hiện tại:

- Trong khi các nghiên cứu trước đây đã chứng minh tính nhất quán cho CART trên *additive regression models*, việc tính toán *rate upper* và *lower bounds* cho *generalization error* của CART và các thuật toán decision tree khác vẫn còn là một vấn đề quan trọng.
- Điều này cho phép so sánh hiệu suất của chúng với các thuật toán được thiết kế đặc biệt như *backfitting*.

Nghiên cứu về minimax rates:

- Raskutti et al. (2012) [5] 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) [13] đã thiết lập ℓ_2 minimax rate cho *nonparametric estimation* của các hàm C^1 trong s chiều là $\Omega\left(n^{-\frac{2}{s+2}}\right)$.

2.4. Các nghiên cứu liên quan khác

Nghiên cứu về Boolean features:

- Syrgkanis & Zampetakis (2020) [14] đã chứng minh *generalization upper bounds* cho CART trong các thiết lập khác nhau, xem xét Boolean features và áp đặt giả định *submodularity* trên conditional mean function.
- Mặc dù điều này bao gồm additive models, các tác giả không đưa ra ví dụ cụ thể về các models khác thỏa mãn giả định này.

Nghiên cứu về feature importance:

- Scornet (2020) [12] 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.

Các ứng dụng sinh học:

- Behr et al. (2021) đã công thức hóa một *discontinuous nonlinear regression model* lấy cảm hứng từ sinh học và chỉ ra rằng CART trees có thể được sử dụng để thực hiện inference cho model này.

2.5. Định vị nghiên cứu hiện tại

Sự khác biệt với các nghiên cứu trước:

- Nghiên cứu này là công trình đầu tiên thiết lập *algorithm-specific lower bounds* cho CART hoặc bất kỳ thuật toán decision tree nào khác.
- Algorithm-specific lower bounds đặc biệt khó trong literature machine learning vì chúng yêu cầu các kỹ thuật chuyên biệt thay vì dựa vào công thức chung (như trường hợp với minimax lower bounds).

So sánh với Tang et al. (2018):

- Tang et al. (2018) đã chứng minh các điều kiện đủ để honest random forest estimators không nhất quán cho một số regression functions đặc biệt, sử dụng *Stone (1977)'s adversarial construction*.

- Đây là công trình duy nhất khác mà chúng tôi biết cung cấp kết quả *negative* cho tree-based estimators.

- Tuy nhiên, họ không tính toán lower bounds, và các điều kiện của họ hoặc liên quan đến lựa chọn hyperparameters không thực tế, hoặc liên quan đến các thuộc tính của trees sau khi chúng được grown.

Gap nghiên cứu được địa chỉ:

- Hiểu rõ hơn về *inductive bias* của decision trees—những giả định mà các thuật toán thực hiện để tổng quát hóa sang dữ liệu mới.

- Cung cấp guidance cho practitioners về khi nào và làm thế nào để áp dụng các phương pháp này.

- Khám phá hiệu suất tổng quát hóa của decision trees đối với các *generative regression models* khác nhau.

2.6. Bảng so sánh các công trình nghiên cứu liên quan

Để hệ thống hóa và làm rõ vị trí của bài báo nghiên cứu này, chúng tôi trình bày bảng so sánh các hướng tiếp cận lý thuyết chính về Cây Quyết Định dưới đây. Bảng này tóm tắt đóng góp, ưu điểm, nhược điểm và định hướng phát triển từ mỗi công trình, cho thấy một sự tiến triển logic dẫn đến những câu hỏi mà bài báo của **Tan et al. (2021)** đã giải quyết.

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

Kết luận từ Bảng so sánh

Bảng trên cho thấy một lộ trình nghiên cứu rõ ràng: từ việc khẳng định Cây Quyết Định có thể hoạt động (tính nhất quán), đến việc tìm hiểu xem nó hoạt động tốt như thế nào (tốc độ hội tụ và hiệu quả). Các công trình trước đây đã thành công trong việc chứng minh tính nhất quán trong các bối cảnh ngày càng thực tế hơn, nhưng đều bỏ ngỏ câu hỏi quan trọng nhất về hiệu quả thống kê.

Bài báo của Tan et al. (2021) chính là công trình đầu tiên lấp đầy khoảng trống này. Nó không chỉ trả lời câu hỏi "CART có hiệu quả không?" bằng một câu trả lời định lượng "Không, nó không hiệu quả bằng mức tối ưu", mà còn chỉ ra một cách thuyết phục **lý do tại sao**, mở ra một hướng đi hoàn toàn mới để cải thiện các thuật toán dựa trên cây trong tương lai.

CHƯƠNG 3

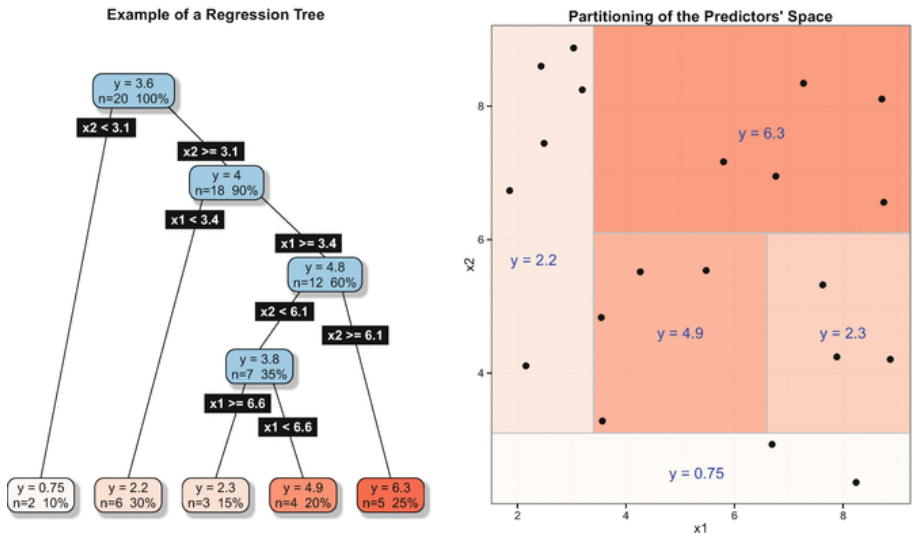
KIẾN THỨC NỀN TẢNG

Trong chương này, nhóm trình bày những kiến thức cơ bản và các khái niệm quan trọng làm nền tảng cho hướng nghiên cứu và chứng minh các kết quả trong bài báo. Kiến thức bao gồm những khái niệm cơ bản về cây quyết định, cơ sở toán học và thuật toán, kỹ thuật liên quan.

3.1. Khái niệm cơ bản về cây quyết định

Cây quyết định là một trong những mô hình quan trọng và được sử dụng rộng rãi trong học máy có giám sát. Ý tưởng chính của cây quyết định là chia nhỏ không gian đặc trưng thành nhiều vùng con (cells) thông qua các phép tách theo trục (axis-aligned splits), sau đó xây dựng mô hình dự đoán đơn giản trong từng vùng.

Thuật toán điển hình nhất là CART (Classification and Regression Trees) do Breiman et al. (1984) đề xuất. Trong hồi quy, CART dự đoán giá trị đầu ra bằng **trung bình của các quan sát trong cùng một lá**.



Hình 3.1: Hình ảnh minh họa về cách hoạt động của quyết định hồi quy [2]

Một trong những điểm mạnh của cây quyết định là: **tính diễn giải** cao khi mà các cây có kích thước nhỏ hay vừa phải dễ đọc, dễ trực quan hóa và giải thích. Bên cạnh đó, cây quyết định còn là thành phần quan trọng trong các phương pháp mạnh như Random Forest hay Gradient Boosting

Trong bài báo này, ta tập trung vào cây ALA(Averaging over Leaf-Only Partitions) khi mà không gian đặc trưng $\mathcal{X}$ được chia thành các ô, và dự đoán cho mọi điểm x trong một ô được tính bằng trung bình các giá trị y trong ô đó. Nói một cách khác cây ALA là một tổng quát của CART. Chi tiết công thức sẽ được trình bày trong phần kế tiếp.

3.2. Cơ sở toán học và kỹ thuật

3.2.1. Lý thuyết độ đo và xác suất

Trong phần chứng minh các kết quả thu được, để đọc hiểu các chứng minh cần trang bị các kiến thức liên quan đến lý thuyết độ đo và xác suất nền tảng:

- **Không gian đo (Measure Space):** Không gian đo là bộ ba $(\Omega, \mathcal{F}, \mu)$ trong đó Ω là không gian mẫu, $\mathcal{F}$ là σ-đại số trên Ω và μ là độ đo là ánh xạ từ $\mathcal{F} \longrightarrow [0, +\infty]$ với phép tính cộng vô hạn [15]. Trong bài báo ta thấy được không gian các vector đặc trưng $\mathcal{X} \subset \mathbb{R}^d$.

- **Kỳ vọng (Expectation):** Cho $f : X \to \mathbb{R}$ là một hàm đo được và độ đo ν trên $\mathcal{X}$ thì

$$\mathbb{E}_\nu[f(x)] := \int_X f(x)\, d\nu(x).$$

  Ngoài ra trong paper còn đề cập đến kì vọng có điều kiện (Conditional Expectation) trên một ô $C \subset X$ là:

$$\mathbb{E}_\nu[f(x) \mid x \in C] := \frac{1}{\nu(C)} \int_C f(x)\, d\nu(x).$$

  chính là cách tính trung bình trong một lá của cây.

- **Phương sai (Variance):** Tương tự như kỳ vọng thì trong bài báo sử dụng đến phương sai có điều kiện:

$$\mathrm{Var}_\nu[f(x) \mid x \in C] := \frac{1}{\nu(C)} \int_C \Big(f(x) - \mathbb{E}_\nu[f(x) \mid x \in C]\Big)^2 d\nu(x).$$

- **Bất đẳng thức cơ bản [8]:** Sử dụng chủ yếu hai bất đẳng thức xác suất là bất đẳng thức Chebyshev:

$$\mathbb{P}(\mid X - \mathbb{E}[X] \mid \geq k\sigma) \leq \frac{1}{k^2}$$

và bất đẳng thức Cauchy-Schwarz:

$$\mid \mathbb{E}[XY] \mid \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

Hai bất đẳng thức này được sử dụng trong chặn trên và dưới trong bias-variance decomposition

3.2.2. Hồi quy giám sát và mô hình cộng tính

Bài báo làm việc trong khung hồi quy giám sát chuẩn:

$$y = f(\mathbf{x}) + \epsilon,$$

trong đó $x \in \mathbb{R}^d$ là vector đặc trưng, $y \in \mathbb{R}$ là output, và ϵ là biến nhiễu thỏa $\mathrm{Var}[\epsilon \mid x] = \sigma^2$

Mô hình cộng tính biểu diễn hàm trung bình có điều kiện dưới dạng tổng các hàm một biến:

$$f(\mathbf{x}) = \sum_{j=1}^{d} \phi_j(x_j),$$

với ϕ_j là hàm phụ thuộc vào tọa độ j của vector x.

3.2.3. Phân vùng không gian của cây ALA

Không gian đặc trưng $\mathcal{X}$ được chia thành các ô(cell):

- Nếu $X = [0,1]^d$, thì $C = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$.
- Nếu $X = \{0,1\}^d$, thì $C(S, z) = \{x \in \{0,1\}^d : x_j = z_j, \forall j \in S\}$, với $S \subset [d]$.

Cho tập huấn luyện $D_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, định nghĩa số mẫu trong cell C:

$$N(C) := \#\{i : x^{(i)} \in C\}.$$

Một phân vùng $\mathfrak{p} = \{C_1, \ldots, C_m\}$ là tập hợp các ô có phần trong rỗng tách rời và phủ toàn bộ không gian X. Trên mỗi phân vùng, ALA tree được định nghĩa bằng *leaf-only averaging*:

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

với quy ước nếu $N(C) = 0$ thì trung bình bằng 0.

3.2.4. Rủi ro bình phương và Rủi ro kỳ vọng tối ưu

- **Rủi ro bình phương/Rủi ro tổng quát**:

$$R(\hat{f}) := \mathbb{E}_{x\sim\nu}\big[(\hat{f}(x) - f(x))^2\big],$$

thể hiện sai số trung bình bình phương của bộ ước lượng $\hat{f}$ so với hàm thực f trên phân phối dữ liệu.

- **Rủi ro kỳ vọng tối ưu**:

$$R^*(f, \nu, n) := \inf_{\mathfrak{p}} \mathbb{E}[R(\hat{f}(-; \mathfrak{p}, D_n))],$$

là giá trị rủi ro trung bình nhỏ nhất có thể đạt được bởi ALA tree khi chọn phân vùng hợp lệ tốt nhất.

3.3. Lý thuyết và khái niệm liên quan

- **Lời nguyền số chiều (Curse of dimensionality):** hiện tượng phát sinh khi phân tích dữ liệu trong không gian nhiều chiều. Khi số lượng đặc trưng (chiều) tăng lên, không gian dữ liệu trở nên cực kỳ thưa thớt. Khoảng cách giữa các điểm dữ liệu bất kỳ có xu hướng trở nên gần bằng nhau. Để duy trì một mật độ dữ liệu nhất định, số lượng mẫu cần thiết sẽ tăng theo hàm mũ với số chiều.

- **Phân rã bias và variance (Bias-variance decomposition):** Cho một mô hình ước lượng $\hat{f}$ cho hàm trung bình có điều kiện f, lỗi bình phương kỳ vọng được phân rã thành:

$$\mathbb{E}\big[(\hat{f}(x) - f(x))^2\big] = \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{\text{Bias}^2(x)} + \underbrace{\mathbb{E}\big[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\big]}_{\text{Variance}(x)} + \underbrace{\sigma^2}_{\text{Lỗi không thể giảm thiểu}},$$

Luôn tồn tại sự đánh đổi (trade-off) giữa Bias và Variance, tức là giảm bias thường làm tăng variance và ngược lại.

- **Lý thuyết tốc độ biến dạng (Rate-Distortion Theory):** Là một nhánh của lý thuyết thông tin, lý thuyết tốc độ biến dạng cung cấp các giới hạn cơ bản về việc nén dữ liệu có tổn hao (lossy compression). Nó trả lời câu hỏi: Cần tối thiểu bao nhiêu bit (tốc độ - rate, R) để mã hóa một tín hiệu sao cho khi giải nén, sai số (biến dạng - distortion, D) so với tín hiệu gốc không vượt quá một ngưỡng nhất định.

CHƯƠNG 4

PHƯƠNG PHÁP NGHIÊN CỨU

4.1. Phân tích rủi ro dưới dạng bias–variance đối với cây ALA

Như đã đề cập ở phần trước, luôn tồn tại sự đánh đổi giữa Bias(Độ chệch) và Variance(Phương sai), nghĩa là khi giảm đi Bias(Độ chệch) thường sẽ làm tăng Variance(Phương sai) và ngược lại. Chính vì điều này, mục tiêu của học máy là tìm điểm cân bằng để sai số tổng nhỏ nhất.

Tương tự như vậy với cây ALA, về mặt cảm tính, khi ta càng tăng số lượng lá (tức càng tăng số lượng ô trong một phân vùng) thì tất nhiên cây sẽ học tốt hơn khi độ chệch càng giảm nhưng ngược lại phương sai tăng.

Để làm sáng tỏ hơn nhận định này, bài báo đã chứng minh một cách chặt chẽ mối quan hệ giữa Bias-Variance bằng công cụ toán học thông qua định lý sau đây:

Định lý 4.1: Cận trên và cận dưới rủi ro kỳ vọng

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

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

và cận trên:

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

trong đó:

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

- Trước hết về mặt tổng quát hóa, bất đẳng thức đúng với bất kỳ hàm mục tiêu $f(\mathbf{x})$ và bất kỳ phân phối dữ liệu nào. Đặc biệt hơn nữa là không nhất thiết các đường chia phải song song với trục tọa độ, nghĩa là nó bao quát một lớp thuật toán phân vùng còn rộng hơn cả cây quyết định thông thường.

- Cả hai chặn trên và chặn dưới đều có hai thành phần chính: thành phần Bias(Độ chệch):

$$\sum_{\mathcal{C}\in p} \text{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} = \mathbb{E}\{\text{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\} = \mathbb{E}\{(f(\mathbf{x}) - \bar{f}_p(\mathbf{x})\}$$

  và thành phần Variance(Phương sai) là $\dfrac{|\mathfrak{p}|\sigma^2}{n}$. Chúng chỉ khác nhau bởi các hằng số và một số hạng lỗi $\mathbb{E}(\mathfrak{p})$ có thể kiểm soát được

- Điều quan trọng là bất đẳng thức đã chứng minh được sự diễn ra trade-off của Bias và Variance từ đó cung cấp một cái nhìn sâu sắc và giải thích được cơ chế hoạt động của các thuật toán thực tế như CART. Thậm chí biết được "giá trị" cho sự trao đổi này: Khi mà chia một ô $\mathcal{C}$ thành hai ô con $\mathcal{C}_L$ và $\mathcal{C}_R$ thì Bias sẽ giảm đi một lượng là:

$$\Delta\text{Bias} = \text{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} - \text{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}_L\}\nu\{\mathcal{C}_L\} - \text{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}_R\}\nu\{\mathcal{C}_R\}$$

  trong khi đó thì Variance sẽ tăng lên theo $\Delta\text{Variance} = O\left(\dfrac{\sigma^2}{n}\right)$

- Kết quả nhận xét trên có ý nghĩa lớn khi cho thấy các phương pháp chống overfitting kinh điển trong CART, như dừng sớm dựa trên độ giảm tạp chất tối thiểu (minimum impurity decrease) hay cắt tỉa dựa trên độ phức tạp chi phí (cost-complexity pruning) thực chất là đi tối ưu hóa trade-off giữa Bias và Variance.

Nhận xét 1

+Cái hay trong kết quả này là góp phần cho thấy rằng các phương pháp tránh overfitting truyền thống là hoàn toàn hợp lý. Chẳng hạn như (minimum impurity decrease) là một dạng phương pháp dừng sớm hay là (cost-complexity pruning) là một dạng regularization.
+Kết quả chặng trên và dưới đều mang tính tổng quát cao với bất kỳ cách phân hoạch nào, thậm chí trong phần phụ lục bất đẳng thức còn được đánh giá chặt chẽ hơn.
+Phần chứng minh và mọi thứ thuần lý thuyết khiến cho việc đọc hiểu chứng minh cần nhiều kiến thức nền tảng và nhiều thời gian

4.2. Mối liên hệ giữa ước lượng cây quyết định và lý thuyết tốc độ–biến dạng trong trường hợp f tuyến tính

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 (có trọng số β) giữa một điểm dữ liệu gốc và dữ liệu tái tạo, liên hệ mật thiết đến thành phần Bias thông qua bất đẳng thức $\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} \geq \dfrac{\delta(p;\beta)}{2}$ khi mà thay thế tất cả các điểm trong một lá $\mathcal{C}$ bằng một điểm đại diện duy nhất (tâm $z(\mathcal{C})$)
- **Tốc độ nén:** số bit trung bình cần để mã hóa dữ liệu sao cho độ biến dạng không vượt quá D và thêm cả hàm trade-off tối ưu $R(D;p_{\mathbf{x}},\beta) := \inf_{p_{\hat{x}|x}} I(\mathbf{x};\hat{\mathbf{x}})$ thì quan hệ với số lá thông qua bất đẳng thức $\log|\mathfrak{p}| \geq R(\delta(p;\beta);\nu,\beta)$

Định lý 4.2: Chặn dưới cho kỳ vọng tối ưu trong trường hợp f tuyến tính

Giả sử rằng các biến hiệp phương sai là độc lập, và hàm trung bình có điều kiện là tuyến tính: $f(\mathbf{x}) = \beta^T\mathbf{x}$. Khi đó, rủi ro kỳ vọng tối ưu bị chặn dưới bởi:

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

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 số lượng s sao cho $|\beta_j|\geq\beta_0 \forall j\in S$. Khi đó, với bất đẳng thức được tác giả chứng minh:

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

Nó đã chuyển đổi bài toán tổ hợp trên các phân vùng thành một bài toán tối ưu hóa $\varphi(D) = D + kD^{-s/2}$ một biến liên tục D và thu kết quả $D^* = s2^{\frac{2s}{s+2}h(\nu_0)-1}\left(\frac{\beta_0^2}{\pi e}\right)^{s/(s+2)}\left(\frac{\sigma^2}{n}\right)^{2/(s+2)}$ dẫn tới kết quả chặn dưới cuối cùng cho f tuyến tính:

$$R^*(f,\nu,n) \geq s2^{\frac{2s}{s+2}h(\nu_0)-2}\left(\frac{\beta_0^2}{\pi e}\right)^{s/(s+2)}\left(\frac{\sigma^2}{n}\right)^{2/(s+2)}$$

- Phần này đã thể hiện một kỹ thuật xử lý sáng tạo khi liên hệ giữa lý thuyết độ chệch-phương sai với lý thuyết tốc độ-biến dạng. Từ một bài toán làm việc với biến cần tối ưu là $\mathfrak{p}$-một biến tổ hợp phức tạp chuyển sang biến số D-một biến vô hướng liên tục.
- Từ kết quả thu được từ chặn dưới thì $R^*(f, \nu, n) \geq O(n^{-2/s+2})$ ta có thể thấy tốc độ hội tụ $\Omega(n^{-2/s+2})$ được biết đến là tốc độ hội tụ minimax tối ưu cho việc ước lượng một hàm trơn tổng quát trong không gian s chiều [16].
- Chính vì điều này đã cho thấy cây quyết định hoàn toàn thất bại trong việc khai thác cấu trúc đơn giản của dữ liệu (mô hình tuyến tính-một cấu trúc đơn giản hơn nhiều so với một hàm trơn tổng quát)

4.3. Kết quả chặn dưới cho mô hình cộng tính

Trong phần này, với sự tổng quát lên mô hình cộng tính khiến cho mối liên hệ giữa $\text{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}$ và $\|x - x'\|^2$ không còn nữa, các tác giả đã tiếp cận hướng khác và thu được hai kết quả mấu chốt cuối cùng của bài báo:

Định lý 4.3: Chặn dưới cho mô hình cộng tính trên khối lập phương đơn vị

Với mô hình hồi quy và hàm mục tiêu f là mô hình cộng tính với ϕ_j như trên, 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 $S \subset [d]$ có số lượng s sao cho $\min_{t \in I_j} |\phi_j'(t)| \geq \beta_0 > 0$ với mọi $j \in S$.
Đặt $\mathcal{K} = \{\mathbf{x} : x_j \in I_j vij = 1, \cdots, d\}$. Giả sử rằng ν là phân phối liên tục với mật độ q và kí hiệu $q_{min} = \min_{\mathbf{x} \in \mathcal{K}} q(\mathbf{x})$. Khi đó rủi ro kỳ vọng tối ưu bị chặn dưới bởi:

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

- Về mặt tổng quát, kết quả này là mở rộng của kết quả trong phần phía trước khi cho phép hàm ϕ_j phi tuyến tính và không đòi hỏi hiệp phương sai độc lập.
- Tuy nhiên, tốc độ hội tụ tổng quát theo bất đẳng thức này vẫn là $\Omega(n^{-2/s+2})$ như đã tìm trong trường hợp tuyến tính cho thấy đây là một thuộc tính cơ bản của cây quyết định khi đối mặt với bất kỳ cấu trúc cộng tính trơn nào.

- Tốc độ $\Omega(n^{-2/s+2})$ là tốc độ minimax cho việc ước lượng một hàm trơn C^1 trong không gian s bất kỳ. 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

Kết quả tiếp theo sẽ liên quan đến không gian đặc trưng rời rạc và hữu hạn $\{0,1\}^d$. Khi đó điều đặc biệt xảy ra, bất kỳ mô hình cộng tính nào cũng là mô hình tuyến tính từ đó có thể tận dụng lại hướng tiếp cận cũ mà thu được cận dưới:

Định lý 4.4: Chặn dưới cho mô hình cộng tính Bool

Với giả thiết mô hình hồi quy như cũ. Giả sử rằng không gian hiệp phương sai là $\{0,1\}^d$ và giả sử các hiệp phương sai độc lập với $x_j \sim Ber(\pi)$ với $0 \leq \pi \leq 1/2$ và $j = 1, \cdots, d$. Giả sử tồn tại tập hợp con các chỉ số $S \subset [d]$ có số lượng s sao cho $\beta_j \geq \beta_0 > 0$ với mọi $j \in S$. Khi đó rủi ro kỳ vọng tối 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

CHƯƠNG 5

THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

5.1. Thiết lập thực nghiệm và tập dữ liệu

5.1.1. Thiết lập thực nghiệm

- **Thông số CPU**:

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

Bảng 5.1: Thông tin CPU sử dụng cho việc huấn luyện

- **Siêu tham số (hyperparameters):**

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.

Bảng 5.2: Các siêu tham số và ý nghĩa

5.1.2. Thiết lập dữ liệu

- Trong nghiên cứu này, dữ liệu được mô phỏng thông qua 3 loại mô hình với đặc trưng thưa (sparse), s là số đặc trưng quan trọng trong d đặc trưng:
  - **Mô hình tuyến tính với đặc trưng nhị phân**

    Với mỗi điểm dữ liệu x:

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

  - **Mô hình tuyến tính với đặc trưng liên tục**

    Với mỗi điểm dữ liệu x:

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

  - **Mô hình tổng bình phương với đặc trưng liên tục**

    Với mỗi điểm dữ liệu x:

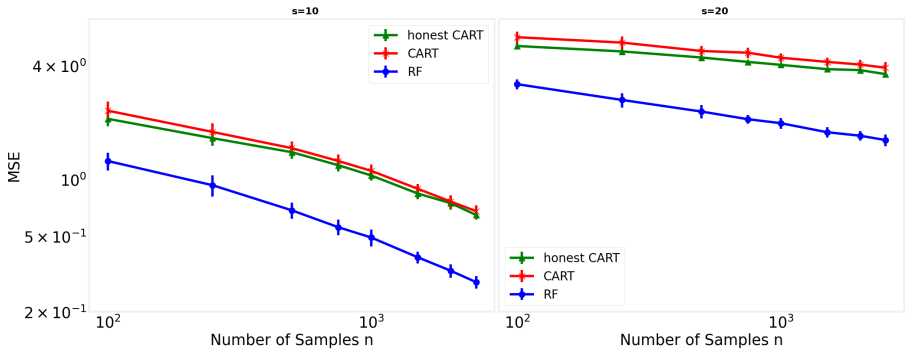
    $$y = \sum_{j=1}^{s} \beta_j x_j^2 + \epsilon, \quad \mathbf{x} = (x_1, \ldots, x_s) \in Unif[0,1)^s, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Bên cạnh đó, nhóm cũng lấy dữ liệu thực tế (không giả định phân phối).[6] [4]

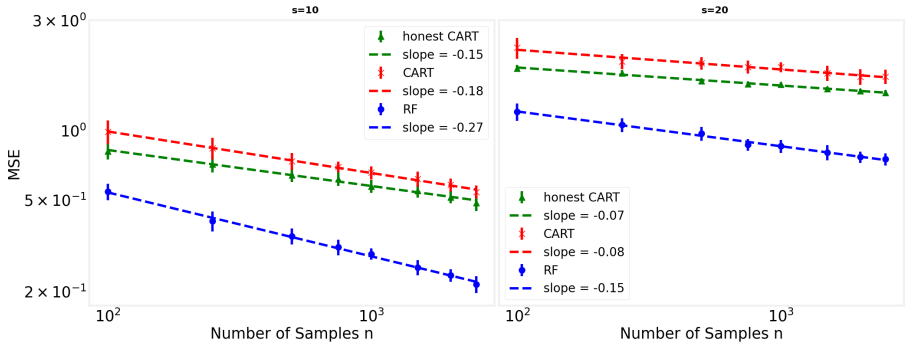
5.2. Chỉ số đánh giá và các phương pháp so sánh

- **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.

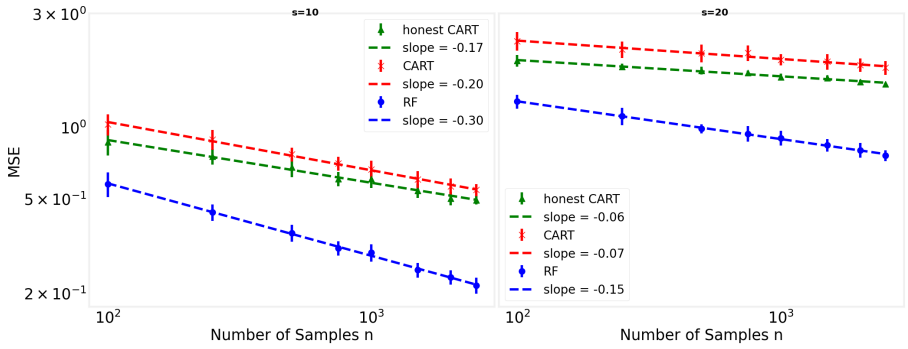
5.3. Phân tích kết quả



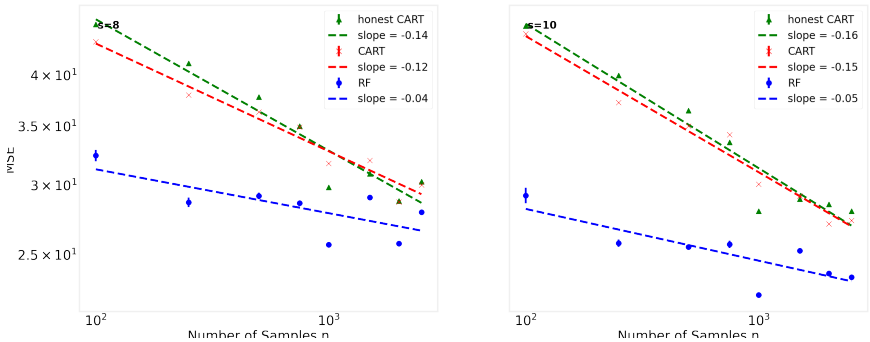
Hình 5.1: Biểu đồ so sánh MSE của các mô hình trên dữ liệu tuyến tính thưa với đặc trưng boolean.



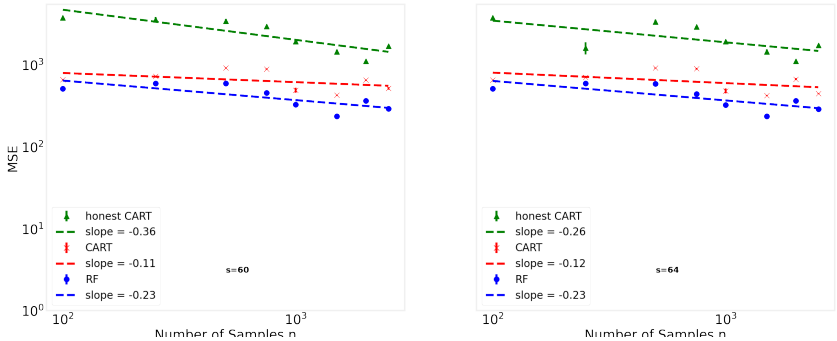
Hình 5.2: Biểu đồ so sánh MSE của các mô hình trên dữ liệu tuyến tính thưa với đặc trưng liên tục (Uniform).



Hình 5.3: Biểu đồ so sánh MSE của các mô hình trên dữ liệu tổng bình phương thưa với đặc trưng liên tục (Uniform).



Hình 5.4: Biểu đồ so sánh MSE của các mô hình trên dữ liệu thực tế (15 đặc trưng).[4]



Hình 5.5: Biểu đồ so sánh MSE của các mô hình trên dữ liệu thực tế (81 đặc trưng).[6]

- **Ảnh hưởng của kích thước mẫu:** Khi số lượng mẫu tăng lên, MSE của các mô hình đều giảm. Điều này là hợp lý vì dữ liệu càng nhiều, mô hình có thể học tốt hơn và đưa dự đoán chính xác hơn.

- **Ảnh hưởng của số lượng đặc trưng quan trọng (sparsity):** Quan sát thấy, với cùng số lượng mẫu, MSE tăng khi độ thưa tăng. Nghĩa là bài toán trở nên khó hơn khi số lượng đặc trưng quan trọng tăng. Khi $s = 20$, mô hình cần phải học một mối quan hệ phức tạp hơn, do đó cần nhiều dữ liệu hơn để đạt được cùng một mức độ chính xác so với khi $s = 10$.

5.4. So sánh hiệu năng

Kiểu dữ liệu	Thời gian chạy (phút)	Thời gian trung bình (s/it)
Mô hình tuyến tính với đặc trưng Boolean	~ 20	~ 85
Mô hình tuyến tính với đặc trưng liên tục	~ 25	~ 93.75
Mô hình tổng bình phương với đặc trưng liên tục	~ 25	~ 93.75
Dữ liệu thực tế (gồm 15 đặc trưng)	~ 6	~ 22
Dữ liệu thực tế (gồm 81 đặc trưng)	~ 18	~ 67.5

Bảng 5.3: So sánh thời gian chạy của các mô hình

Random Forest (RF): Trong tất cả biểu đồ, RF là mô hình có hiệu suất tốt nhất. MSE của mô hình này thấp hơn đáng kể so với các mô hình còn lại. Điều này có thể lý giải là do **RF** kết hợp nhiều **CART**, kĩ thuật trên giúp giảm phương sai, cải thiện hiệu suất của dự đoán. Mặc dù **RF** vượt trội hơn **CART**, các kết quả cho thấy tốc độ hội tụ thực tế của **RF** vẫn chậm hơn đáng kể so với tốc độ **minimax** cho mô hình cộng tính thưa.

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

Ở Hình 5.2 và Hình 5.3, ta thấy độ dốc (slope) trên biểu đồ log-log là hoàn toàn phù hợp với **minimax rate** đã nêu, cụ thể là độ dốc -0.17 và -0.08 lần lượt với $s = 10, s = 20$. Đường cong của **Honest CART** và **CART** ở Hình 5.1 có dạng cong phù hợp với hình dạng mà công thức ở Định lý 4.4 nêu ra. Nên về mặt thực nghiệm, hai mô hình này gần đạt giới hạn lý thuyết.

Khi sử dụng dữ liệu thực (Hình 5.4 và Hình 5.5), ta nhận thấy **RF** vẫn là thuật toán có hiệu suất ổn định nhất. Song, kết quả vẫn có một vài điểm khác biệt so với khi sử dụng dữ liệu mô phỏng:

- **Honest CART** có hiệu suất thấp hơn so với **Dishonest CART**.
- **Honest CART** có độ dốc âm nhất, thay vì **RF**.

$\Rightarrow$ 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.

CHƯƠNG 6

KẾT LUẬN

6.1. Những phát hiện và đóng góp chính

Bài báo này đã xác định được các **giới hạn lý thuyết** (theoretical lower bounds) về rủi ro dự kiến (expected risk) cho cây ALA (honest ALA trees) khi được huấn luyện trên các mô hình cộng tính (additive models). Trong khi đó, các mô phỏng của chúng tôi cho thấy những kết quả này cũng có thể đúng với các mô hình cây không-trung thực (non-honest counterparts). Một đóng góp quan trọng là lập luận rằng các bộ ước lượng như **CART** có **thiên kiến quy nạp** (inductive bias) chống lại cấu trúc toàn cục (global structure). Thiên kiến này không phải do thuật toán phân tách tham lam (greedy splitting) mà do đặc tính chỉ sử dụng giá trị trung bình ở lá (leaf-only averaging property) của lớp các bộ ước lượng này.

Bằng chứng thực nghiệm cho thấy **Random Forest (RF)** không bị áp dụng các giới hạn này, từ đó ủng hộ luận điểm của Breiman (2001) rằng sự đa dạng của các cây trong một rừng giúp giảm phương sai và cải thiện hiệu suất dự đoán. Tuy nhiên, tốc độ của RF vẫn chậm hơn đáng kể so với **minimax rates** (tốc độ tối ưu) cho các mô hình cộng tính thưa (sparse additive models), cho thấy vẫn còn những giới hạn cơ bản chưa được hiểu rõ.

6.2. Tổng kết thực nghiệm trên dữ liệu thực tế

Kết quả thực nghiệm trên bộ dữ liệu thực tế cho thấy Random Forest (RF) tiếp tục duy trì hiệu suất dự đoán ổn định và nhìn chung vượt trội hơn so với các biến thể của CART, phù hợp với quan sát từ dữ liệu mô phỏng. Tuy nhiên, khác với dữ liệu giả lập, Honest CART lại thể hiện hiệu suất thấp hơn Dishonest CART. Ngoài ra, Honest CART có độ dốc âm lớn nhất thay vì RF, cho thấy sự khác biệt đáng kể về hành vi mô hình khi áp dụng trong môi trường thực tế.

Nguyên nhân có thể xuất phát từ việc dữ liệu thực tế thường chứa nhiều nhiễu hơn và ít

khi tuân thủ chặt chẽ giả định cộng tính. Dù vậy, tốc độ hội tụ của các mô hình vẫn hợp lý và tương thích với các giới hạn lý thuyết đã được chứng minh trước đó. Kết quả này nhấn mạnh rằng mặc dù phân tích lý thuyết và dữ liệu mô phỏng cung cấp nền tảng quan trọng, việc kiểm chứng trên dữ liệu thực tế là không thể thiếu để đánh giá đúng mức độ ứng dụng và khả năng tổng quát hóa của các mô hình học máy.

6.3. Điểm mạnh và hạn chế của phương pháp tiếp cận

Phương pháp tiếp cận của bài báo là **phân tích một thuật toán như một đối tượng nghiên cứu sơ cấp**, và tìm cách phân tích hiệu suất của nó dưới các mô hình sinh dữ liệu khác nhau để làm rõ **thiên kiến quy nạp của nó**. Cách tiếp cận này khác với phương pháp thống kê cổ điển, vốn bắt đầu từ một bài toán ước lượng và sau đó tìm kiếm các quy trình ước lượng tối ưu. **Điểm mạnh** của phương pháp này là nó phù hợp hơn với phân tích dữ liệu hiện đại, khi chúng ta thường không có đủ kiến thức về dạng hàm của quá trình sinh dữ liệu.

Tuy nhiên, bài báo chỉ mới chạm tới bề nổi trong việc điều tra thiên kiến quy nạp của các thuật toán cây, RF và gradient boosting. Nó chưa trực tiếp giải quyết vấn đề mất khả năng phát hiện cấu trúc toàn cục do việc phân vùng không gian hiệp biến. Hơn nữa, việc tập trung vào các "rừng cây" (forests) cũng cản trở khả năng giữ được tính giải thích (interpretability) của mô hình.

6.4. Ý nghĩa đối với lĩnh vực

Nghiên cứu này có một số ý nghĩa quan trọng:

- **Cải thiện các thuật toán cây:** Các kết quả cho thấy cần phải sửa đổi các thuật toán cây để chúng có thể học được cấu trúc toàn cục dễ dàng hơn. Gợi ý tự nhiên là áp dụng một số hình thức thu hẹp phân cấp (hierarchical shrinkage) hoặc gộp toàn cục (global pooling), hoặc kết hợp các phương pháp dựa trên cây với các phương pháp tuyến tính hoặc cộng tính (như RuleFit của Friedman và Popescu).
- **Xác định thuật toán phù hợp:** Phân tích thiên kiến quy nạp của các thuật toán trên các mô hình hồi quy sinh dữ liệu khác nhau sẽ giúp các nhà khoa học dữ liệu xác định thuật toán nào nên sử dụng trong một ứng dụng cụ thể, đặc biệt trong các tình huống mà không có tập

dữ liệu kiểm tra độc lập (held-out test set).

- **Cung cấp cảm hứng cho các nghiên cứu mới:** Cuộc điều tra này có thể tạo cảm hứng để cải tiến các thuật toán hiện có và khuyến khích một cách tiếp cận mới trong phân tích thuật toán, phù hợp hơn với thực tiễn dữ liệu hiện đại.

6.5. Định hướng nghiên cứu tương lai

- **Tiếp tục phân tích thiên kiến quy nạp:** Mở rộng phân tích sang các mô hình hồi quy sinh dữ liệu khác để làm rõ hơn thiên kiến quy nạp của CART và các thuật toán cây khác.
- **Ứng dụng phân tích cho các thuật toán khác:** Áp dụng cùng loại phân tích này cho các thuật toán học máy khác.
- **Nghiên cứu các phương pháp kết hợp:** Tiếp tục khám phá các ý tưởng kết hợp các phương pháp dựa trên cây với các phương pháp tuyến tính hoặc cộng tính để tận dụng lợi thế thống kê của cả hai.
- **Khám phá các phương pháp cải tiến:** Nghiên cứu các ý tưởng cải tiến như thu hẹp phân cấp (hierarchical shrinkage) hoặc gộp toàn cục (global pooling) để giúp các mô hình cây học cấu trúc toàn cục hiệu quả hơn.
- **Phát triển các thuật toán mới:** Dựa trên các phát hiện về thiên kiến quy nạp, tạo ra các thuật toán mới có thể giải quyết các hạn chế đã được xác định.

TÀI LIỆU THAM KHẢO

[1] G´erard Biau. "Analysis of a random forests model". In: *The Journal of Machine Learning Research* (2012).

[2] Geoffrey I. Webb Claude Sammut. *Encyclopedia of Machine Learning and Data Mining*. 2nd. Truy cập ngày 19/08/2025. Springer, 2017.

[3] G´erard Biau Erwan Scornet and Jean-Philippe Vert. "Consistency of random forests". In: *The Annals of Statistics* (2015).

[4] Miran Firdausi. *Life Expectancy Based on Individual Lifestyle (dataset)*. Kaggle repository. Accessed: 2025-08-19. 2025. URL: https://www.kaggle.com/datasets/miranfirdausi/life-expectancy-based-on-individual-lifestyle.

[5] Martin J Wainwright Garvesh Raskutti and Bin Yu. "Minimax-optimal rates for sparse additive models over kernel classes via convex programming". In: *Journal of Machine Learning Research* (2012).

[6] Kam Hamidieh. *Superconductivty Data*. UCI Machine Learning Repository. 2018. URL: https://archive.ics.uci.edu/dataset/464/superconductivty+data.

[7] Trevor Hastie and Robert Tibshirani. "Generalized additive models". In: *Statistical Science* (1986).

[8] TS. Phạm Việt Hùng. *Giáo trình Lý thuyết Xác suất*. Truy cập ngày 18/08/2025. 2020.

[9] Jason Klusowski. "Sparse learning with cart". In: *Advances in Neural Information Processing Systems* (2020).

[10] Jason Klusowski. "Universal consistency of decision trees in high dimensions". In: *arXiv preprint arXiv:2104.13881* (2021).

[11] Veeranjaneyulu Sadhanala and Ryan J Tibshirani. "Additive models with trend filtering". In: *The Annals of Statistics* (2019).

[12] Erwan Scornet. “Trees, forests, and impurity-based variable importance”. In: *arXiv preprint arXiv:2001.04295* (2020).

[13] Charles J Stone. “Optimal global rates of convergence for nonparametric regression”. In: *The Annals of Statistics* (1982).

[14] Vasilis Syrgkanis and Manolis Zampetakis. “Estimation and inference with trees and forests in high dimensions”. In: *Conference on Learning Theory* (2020).

[15] Terence Tao. *An Introduction to Measure Theory*. Truy cập ngày 18/08/2025. American Mathematical Society, 2011.

[16] Ryan Tibshirani and Larry Wasserman. *Nonparametric Regression*. Truy cập ngày 20/08/2025. 2019.

[17] Stefan Wager and Susan Athey. “Estimation and inference of heterogeneous treatment effects using random forests”. In: *Journal of the American Statistical Association* (2018).