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

23127438 - 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 25 tháng 8 năm 2025

GIỚI THIỆU CHUNG

Cây quyết định là một thuật toán quan trọng, vừa là mô hình có khả năng diễn giải phù hợp cho các quyết định mang tính rủi ro cao, vừa là thành phần nền tảng cho các phương pháp ensemble như random forests và gradient boosting. Tuy nhiên, các thuộc tính thống kê của chúng vẫn chưa được hiểu rõ. Các công trình trước đây chủ yếu tập trung vào việc chứng minh sự nhất quán điểm (pointwise consistency) cho thuật toán CART trong bối cảnh hồi quy phi tham số cổ điển.

Trong bài báo này, chúng em tiếp cận theo một hướng khác, chủ trương nghiên cứu hiệu suất tổng quát hóa của cây quyết định trên các mô hình sinh dữ liệu hồi quy khác nhau. Điều này cho phép chúng em làm sáng tỏ "thiên kiến quy nạp" (inductive bias) của chúng, tức là những giả định mà thuật toán thực hiện (hoặc không thực hiện) để tổng quát hóa trên dữ liệu mới, từ đó hướng dẫn người thực hành về thời điểm và cách thức áp dụng các phương pháp này.

Chúng em tập trung vào các mô hình cộng tính thưa (sparse additive models), một lớp mô hình có độ phức tạp thống kê thấp nhưng vẫn đủ linh hoạt. Chúng em chứng minh một cận dưới sắc bén cho lỗi tổng quát hóa bình phương (squared error generalization lower bound) cho một lớp lớn các thuật toán cây quyết định khi được áp dụng trên các mô hình này. Đáng ngạc nhiên là cận dưới này tệ hơn nhiều so với tốc độ minimax tối ưu. Sự kém hiệu quả này không xuất phát từ tính "tham lam" (greediness) của thuật toán, mà từ việc mất khả năng phát hiện cấu trúc toàn cục khi chỉ lấy trung bình trên mỗi lá. Để chứng minh các cận dưới này, chúng em phát triển một bộ công cụ kỹ thuật mới, thiết lập một mối liên hệ độc đáo giữa việc ước lượng bằng cây quyết định và lý thuyết tốc độ-méo (rate-distortion theory).

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	**Mô tả**
CART	Cây quyết định phân loại và hồi quy
RF	Random Forest

DANH SÁCH KÝ HIỆU

Ký hiệu	Mô tả
$\mathbf{x}, \mathbf{y}, \mathbf{z}, \cdots$	Các vector được ký hiệu bởi chữ cái in đậm để phân biệt với các biến số thông thường
D_n	Tập huấn luyện gồm n mẫu độc lập
$x^{(i)}$	Mẫu huấn luyện thứ i
x_j	Thành phần thứ j của vector x
$\mathbb{R}$	Tập các số thực
$[a, b]$	Tập các số thực trong đoạn từ a đến b
$[n]$	Tập các số nguyên dương từ 1 đến n
$\inf, \sup$	Cận dưới đúng nhất (infimum), cận trên đúng nhất (supremum)
$X \sim \nu$	Biến ngẫu nhiên X tuân theo phân phối ν
$\mathbb{P}, \mathbb{E}, \mathrm{Var}$	Xác suất, Kỳ vọng và Phương sai
$C \subset X$	Cell (ô): hình chữ nhật trong $[0,1]^d$ hoặc subcube trong $\{0,1\}^d$
$N(C)$	Số lượng mẫu thuộc cell C
$p = \{C_1, \ldots, C_J\}$	Phân hoạch (partition) không gian đặc trưng thành các cell
$\hat{f}(x; p, D_n)$	Bộ ước lượng ALA tree với phân hoạch p và tập dữ liệu D_n
$R(\hat{f})$	Sai số trung bình bình phương
$R^*(f, \nu, n)$	Rủi ro nhỏ nhất đạt được với phân hoạch tối ưu
$C(S, z)$	Subcube trong $\{0,1\}^d$, cố định $x_j = z_j$ với $j \in S$
$\phi_j(x_j)$	Thành phần đơn biến trong mô hình cộng tính
$\sigma^2 = \mathbb{E}[\varepsilon^2 \mid x]$	Phương sai của nhiễu (giả định đồng nhất)

TÓM LƯỢC BÁO CÁO

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 thẩm định tín dụng. 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.

Tóm lại, **Chương 1** đã giới thiệu tổng quan về bối cảnh, mục tiêu và những đóng góp khoa học chính của bài báo. Chương này đã thiết lập nền tảng cho thấy sự cần thiết của việc phân tích sâu hơn về hiệu suất của cây quyết định, một vấn đề còn nhiều hạn chế trong các nghiên cứu trước đây. Để định vị rõ hơn giá trị và tính mới mẻ của nghiên cứu này, **Chương 2** tiếp theo sẽ đi sâu vào việc phân tích các công trình liên quan, từ đó làm nổi bật khoảng trống nghiên cứu mà bài báo đã giải quyết.

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 pointwise** cho CART khi giả định rằng hàm conditional mean là Lipschitz continuous.

Các nghiên cứu về tính nhất quán pointwise:

- Biau (2012) và Wager & Athey (2018) đã chứng minh tính nhất quán pointwise cho CART, nhưng buộc phải thay đổi splitting criterion để đảm bảo mesh của partition học được co về 0.
- Các nghiên cứu này gặp hạn chế về tính thực tiễn do cần *modify* algorithm gốc.

Tiến bộ với additive models:

- Scornet et al. (2015) 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 *additive regression model*.
- Giả định sinh này đơn giản hóa tính toán bằng cách tránh một số phụ thuộc phức tạp giữa các splits có thể tích lũy trong quá trình recursive splitting.

2.2. Nghiên cứu về Sparse Additive Models

Mở rộng cho sparse additive models:

- Klusowski (2020, 2021) đã mở rộng phân tích cho sparse additive models, 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).

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) 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 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) 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) 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.

CHƯƠNG 3

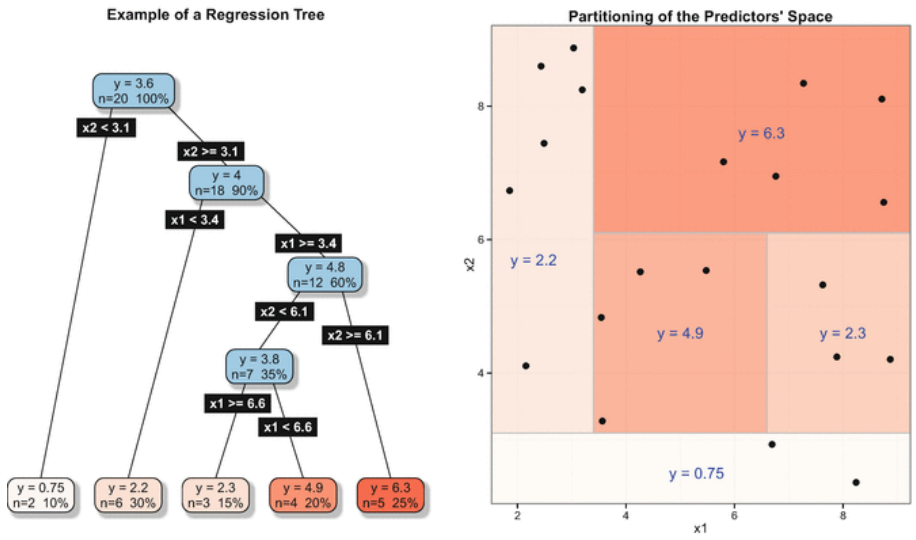
KIẾN THỨC NỀN TẢNG

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 [1]

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 [3]. 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 [2]:** 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(Thiên lệch) và Variance(Phương sai), nghĩa là khi giảm đi Bias thường sẽ làm tăng Variance 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 bias càng giảm nhưng ngược lại variance tăng.

Để làm sáng tỏ hơn nhận định này, bài báo đã chứng minh một cách chặt chẽ mối quan hệ giữa Bias và 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(Thiên lệch):

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

  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 hay ho ở đây 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 của nhóm 4.1

ác giả

CHƯƠNG 5

THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

CHƯƠNG 6

KẾT LUẬN

TÀI LIỆU THAM KHẢO

[1] Geoffrey I. Webb Claude Sammut. *Encyclopedia of Machine Learning and Data Mining*. 2nd. Truy cập ngày 19/08/2025. Springer, 2017. URL: https://link.springer.com/referencework/10.1007/978-1-4899-7687-1.

[2] TS. Phạm Việt Hùng. *Giáo trình Lý thuyết Xác suất*. Truy cập ngày 18/08/2025. 2020. URL: https://www.google.com.vn/books/edition/An_Introduction_to_Measure_Theory/HoGDAwAAQBAJ?hl=en&gbpv=0.

[3] Terence Tao. *An Introduction to Measure Theory*. Truy cập ngày 18/08/2025. American Mathematical Society, 2011. URL: https://www.google.com.vn/books/edition/An_Introduction_to_Measure_Theory/HoGDAwAAQBAJ?hl=en&gbpv=0.