ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

GIÁ TRỊ SHAP TRONG HỌC MÁY GIẢI THÍCH

SEMINAR BỘ MÔN KHOA HỌC MÁY TÍNH

Lê Nhựt Nam

Ngày 22 tháng 4 năm 2026

Nội dung trình bày

SHAP là gì?

SHAP (SHapley Additive exPlanations) là một phương pháp tiếp cận dựa trên lý thuyết trò chơi để giải thích kết quả đầu ra cho bất kỳ mô hình học máy nào.

- Được công bố bởi Lundberg và Lee vào năm 2017.
- Dựa trên giá trị Shapley từ lý thuyết trò chơi hợp tác (cooperative game theory).
- Cung cấp khả năng diễn giải **cục bộ** cho các dự đoán của mô hình.
- Trả lời câu hỏi **"Tại sao?"** trong khi mô hình trả lời **"Bao nhiêu?"**

Điểm then chốt của SHAP

SHAP định lượng sự đóng góp của mỗi đặc trưng vào một dự đoán cụ thể.

SHAP trong ngữ cảnh Học Máy

Ánh xạ từ Lý thuyết Trò chơi:

- **Trò chơi (Game)**: Tái tạo kết quả đầu ra của mô hình
- **Người chơi (Player)**: Các đặc trưng trong mô hình
- **Phần thưởng (Reward)**: Giá trị dự đoán
- **Liên minh (Coalition)**: Tập con các đặc trưng

Tính chất chính:

- **Cục bộ (Local)**: Một quan sát = Một trò chơi
- **Cộng tính (Additive)**: Giá trị SHAP tổng bằng chênh lệch dự đoán
- **Công bằng (Fair)**: Thỏa mãn các tiên đề Shapley
- **Tổng quát (Universal)**: Hoạt động với mọi mô hình

Ví dụ

Mô hình dự đoán thu nhập dựa trên: Tuổi (Age), Giới tính (Sex), Nghề nghiệp (Occupation).

Ký hiệu toán học

Ký hiệu

- $\mathcal{F} = \{1, 2, ..., F\}$: Tập hợp tất cả các đặc trưng
- $S \subseteq \mathcal{F}$: Một liên minh (tập con) các đặc trưng
- $f_S(x_S)$: Mô hình được huấn luyện chỉ với các đặc trưng trong S
- x_i: Giá trị của đặc trưng i cho quan sát x
- ϕ_i: Giá trị SHAP cho đặc trưng i

Vấn đề trung tâm

Vấn đề trung tâm

Đặc trưng i đóng góp bao nhiêu vào dự đoán $f(x)$ so với dự đoán cơ sở $f_\emptyset$?

$$\sum_{i=1}^{F} \phi_i = f(x) - f_\emptyset \tag{1}$$

Tập lũy thừa của các đặc trưng

Định nghĩa

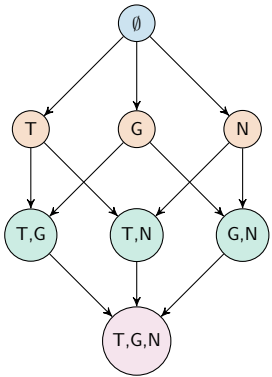
Tập lũy thừa $\mathcal{P}(\mathcal{F})$ (powerset) chứa tất cả các tập con có thể của các đặc trưng

Tập lũy thừa của các đặc trưng

Ví dụ với 3 đặc trưng:
$\mathcal{F} = \{\text{Tuổi, Giới tính, Nghề nghiệp}\}$

Tập lũy thừa có $2^F = 2^3 = 8$ phần tử:

1. $\emptyset$ (không đặc trưng)
2. {Tuổi}
3. {Giới tính}
4. {Nghề nghiệp}
5. {Tuổi, Giới tính}
6. {Tuổi, Nghề nghiệp}
7. {Giới tính, Nghề nghiệp}
8. {Tuổi, Giới tính, Nghề nghiệp}



Định nghĩa của Đóng góp Biên

Định nghĩa

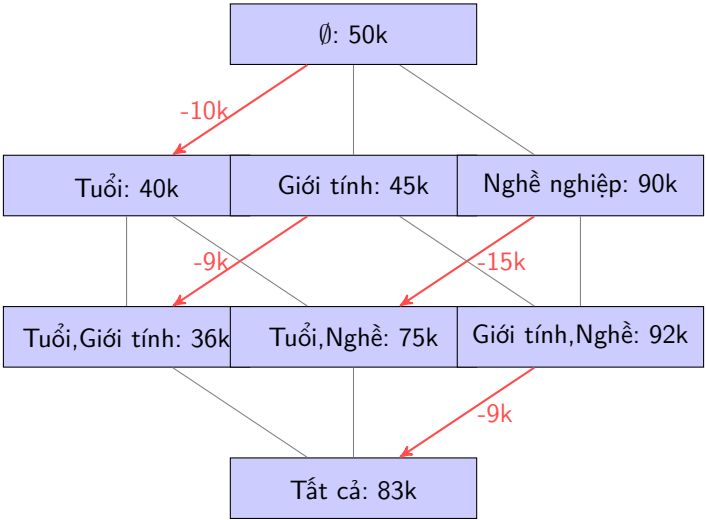
Đóng góp biên của đặc trưng i vào liên minh S (trong đó $i \notin S$):

$$\Delta_i(S) = f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \tag{2}$$

Ví dụ: Thêm Tuổi vào các Liên minh khác nhau

- $\Delta_{\text{Tuổi}}(\emptyset) = f_{\{\text{Tuổi}\}}(x) - f_\emptyset = 40k - 50k = -10k$
- $\Delta_{\text{Tuổi}}(\{\text{Giới tính}\}) = f_{\{\text{Tuổi,Giới tính}\}}(x) - f_{\{\text{Giới tính}\}} = 36k - 45k = -9k$
- $\Delta_{\text{Tuổi}}(\{\text{Nghề nghiệp}\}) = f_{\{\text{Tuổi,Nghề nghiệp}\}}(x) - f_{\{\text{Nghề nghiệp}\}} = 75k - 90k = -15k$
- $\Delta_{\text{Tuổi}}(\{\text{Giới tính,Nghề nghiệp}\}) = f_{\text{Tất cả}}(x) - f_{\{\text{Giới tính,Nghề nghiệp}\}} = 83k - 92k = -9k$

Trực quan hóa Đóng góp biên



Công thức Trọng số Shapley

Trọng số cho Đóng góp biên

Trọng số cho đóng góp biên của đặc trưng i vào liên minh S có kích thước $|S|$ được cho bởi:

$$w(S) = \frac{|S|!(F - |S| - 1)!}{F!}. \tag{3}$$

Ý nghĩa

- Trọng số đảm bảo **tầm quan trọng bằng nhau** cho mọi kích thước liên minh.
- Tổng trọng số cho tất cả liên minh kích thước s bằng $\frac{1}{F}$.
- Thỏa mãn các tiên đề **hiệu quả** và **đối xứng**.

Công thức Trọng số Shapley

Trọng số cho Đóng góp biên

Trọng số cho đóng góp biên của đặc trưng i vào liên minh S có kích thước $|S|$:

$$w(S) = \frac{|S|!(F - |S| - 1)!}{F!} \tag{4}$$

Ví dụ: Trọng số cho Đặc trưng Tuổi ($F = 3$)

- $w(\emptyset) = \frac{0! \cdot 2!}{3!} = \frac{2}{6} = \frac{1}{3}$
- $w(\{\text{Giới tính}\}) = w(\{\text{Nghề nghiệp}\}) = \frac{1! \cdot 1!}{3!} = \frac{1}{6}$
- $w(\{\text{Giới tính, Nghề nghiệp}\}) = \frac{2! \cdot 0!}{3!} = \frac{2}{6} = \frac{1}{3}$

Cách tính Trọng số Thay thế

Số Cạnh mỗi Cấp

Với mô hình có f đặc trưng, tổng số đóng góp biên là:

$$\text{Số cạnh tại cấp } f = f \cdot \binom{F}{f} \tag{5}$$

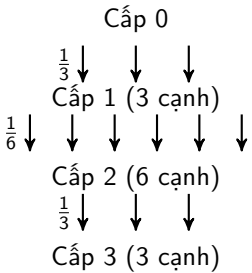
Cách tính Trọng số Thay thế

Ví dụ Tính toán ($F = 3$):

- Cấp 1: $1 \cdot \binom{3}{1} = 3$ cạnh
- Cấp 2: $2 \cdot \binom{3}{2} = 6$ cạnh
- Cấp 3: $3 \cdot \binom{3}{3} = 3$ cạnh

Trọng số:

- Cấp 1: $\frac{1}{3}$ mỗi cạnh
- Cấp 2: $\frac{1}{6}$ mỗi cạnh
- Cấp 3: $\frac{1}{3}$ mỗi cạnh



Công thức Giá trị SHAP

Định nghĩa

Giá trị SHAP cho đặc trưng i là:

$$\phi_i(x) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!(F - |S| - 1)!}{F!} \cdot \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{6}$$

Các thành phần của giá trị SHAP:

- $S \subseteq \mathcal{F} \setminus \{i\}$: Tất cả liên minh không chứa đặc trưng i.
- $\frac{|S|!(F - |S| - 1)!}{F!}$: Trọng số Shapley cho liên minh S.
- $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$: Đóng góp biên.

Ví dụ Tính toán Hoàn chỉnh

Tính giá trị SHAP cho đặc trưng Tuổi:

$$\begin{aligned}
\phi_{\text{Tuổi}} &= \frac{1}{3} \cdot (40k - 50k) + \frac{1}{6} \cdot (36k - 45k) \\
&\quad + \frac{1}{6} \cdot (75k - 90k) + \frac{1}{3} \cdot (83k - 92k) \\
&= \frac{1}{3} \cdot (-10k) + \frac{1}{6} \cdot (-9k) \\
&\quad + \frac{1}{6} \cdot (-15k) + \frac{1}{3} \cdot (-9k) \\
&= -3.33k - 1.5k - 2.5k - 3k \\
&= \boxed{-10.33k}
\end{aligned}$$

Nhận xét: Đặc trưng Tuổi làm giảm dự đoán thu nhập \$10,330 cho quan sát này.

Ví dụ 1: Phân loại Nhị phân

Dự đoán Sống sót trên tàu Titanic

Các đặc trưng:

- x_1: Hạng vé (1/2/3)
- x_2: Giới tính (Nam/Nữ)
- x_3: Tuổi
- x_4: Giá vé

Quan sát:

- Hạng vé: Hạng 3
- Giới tính: Nam
- Tuổi: 25
- Giá vé: \$7.25

Giá trị SHAP:

- $\phi_{\text{Hạng vé}} = -0.15$ (giảm khả năng sống sót)
- $\phi_{\text{Giới tính}} = -0.38$ (giảm mạnh)
- $\phi_{\text{Tuổi}} = -0.05$ (giảm nhẹ)
- $\phi_{\text{Giá vé}} = -0.02$ (giảm nhẹ)

Xác suất cơ sở: 0.38
Dự đoán cuối: $0.38 - 0.60 = -0.22$
(Sau logistic: xác suất sống sót 0.08)

Ví dụ 2: Mô hình dựa trên Cây

Dự đoán Giá Nhà với Random Forest

Đóng góp Đặc trưng cho Nhà $450,000

Đặc trưng	Giá trị	Giá trị SHAP
Giá cơ sở	-	$300,000
Vị trí (Trung tâm)	Có	+$80,000
Diện tích	2,500 sq ft	+$45,000
Phòng ngủ	3	+$5,000
Tuổi nhà	5 năm	+$15,000
Gara	2 xe	+$5,000
Tổng		**$450,000**

Nhận xét: Vị trí đóng góp nhiều nhất vào giá cao hơn (+$80,000), tiếp theo là diện tích (+$45,000).

Ví dụ 3: Phân loại Đa lớp

Phân loại Loài Hoa Iris
Ba lớp: Setosa, Versicolor, Virginica

Quan sát:

- Độ dài đài: 6.3 cm
- Độ rộng đài: 2.8 cm
- Độ dài cánh: 5.1 cm
- Độ rộng cánh: 1.8 cm

Giá trị SHAP theo Lớp:

Đặc trưng	Setosa	Versic.	Virgin.
Dài đài	-0.15	+0.05	+0.10
Rộng đài	-0.20	+0.08	+0.12
Dài cánh	-0.45	-0.10	+0.55
Rộng cánh	-0.20	-0.03	+0.23
Tổng	-1.00	0.00	+1.00

Nhận xét: Dự đoán mạnh mẽ là Virginica (Độ dài cánh là yếu tố phân biệt chính).

Các Tính chất Cơ bản

1. Độ chính xác cục bộ (Hiệu quả)

$$f(x) = f_{\emptyset} + \sum_{i=1}^{F} \phi_i(x) \tag{7}$$

Giá trị SHAP tổng chính xác bằng chênh lệch giữa đầu ra mô hình và giá trị kỳ vọng

2. Tính vắng mặt (Missingness)

Nếu đặc trưng i không có trong mô hình, thì $\phi_i = 0$

Các Tính chất Cơ bản

3. Tính nhất quán/ đơn điệu (Consistency/Monotonicity)

Nếu mô hình f' phụ thuộc vào đặc trưng i nhiều hơn mô hình f, thì:

$$\phi_i(f', x) \geq \phi_i(f, x) \tag{8}$$

4. Tính đối xứng (Symmetry)

Nếu các đặc trưng i và j đóng góp như nhau vào tất cả liên minh:

$$f_{S\cup\{i\}}(x) = f_{S\cup\{j\}}(x) \quad \forall S : i, j \notin S \tag{9}$$

Thì: $\phi_i(x) = \phi_j(x)$

Các Tính chất Cơ bản

5. Tính tuyến tính (Linearity)

Với tổ hợp tuyến tính các mô hình:

$$\phi_i(af + bg, x) = a\phi_i(f, x) + b\phi_i(g, x) \tag{10}$$

Định lý về tính duy nhất của giá trị SHAP

Giá trị SHAP là nghiệm **duy nhất** thỏa mãn đồng thời tất cả các tính chất này.

Độ phức tạp Tính toán

Không gian cao chiều

Tính toán SHAP chính xác yêu cầu đánh giá 2^F mô hình!

- 10 đặc trưng: 1,024 mô hình
- 20 đặc trưng: 1,048,576 mô hình
- 50 đặc trưng: 1.1×10^{15} mô hình

Các phương pháp xấp xỉ

1. **Dựa trên lấy mẫu**: Lấy mẫu Monte Carlo các liên minh.
2. **Linear SHAP** [6]: Chính xác cho mô hình tuyến tính.
3. **Tree SHAP** [3, 8]: Thời gian đa thức cho mô hình dựa trên cây.
4. **Deep SHAP** [5, 7]: Xấp xỉ DeepLift cho mạng nơ-ron.
5. **Kernel SHAP** [9]: Hồi quy tuyến tính có trọng số dựa trên LIME.

Thuật toán Tree SHAP

Cải tiến chính

Với mô hình dựa trên cây, tính giá trị SHAP chính xác trong thời gian $O(TLD^2)$:

- T: Số lượng cây
- L: Số lượng lá tối đa
- D: Độ sâu tối đa

Thuật toán Tree SHAP

Algorithm 1 Tree SHAP (Đơn giản hóa)

Khởi tạo xác suất đường đi $p = 1$
Khởi tạo đóng góp đặc trưng $\phi = 0$
for mỗi đường đi từ gốc đến lá **do**
    Cập nhật p dựa trên điều kiện phân chia
    for mỗi đặc trưng i trong đường đi **do**
        $\phi_i \leftarrow \phi_i + p \times$ đóng góp
    end for
end for
return ϕ

Mở rộng của SHAP

Khả năng Diễn giải Toàn cục

- **Tầm quan trọng Đặc trưng SHAP**: $I_i = \mathbb{E}[|\phi_i|]$
- **Biểu đồ Phụ thuộc SHAP**: Cách ϕ_i thay đổi theo x_i
- **Giá trị Tương tác SHAP**: Tương tác bậc hai

Các phương pháp Nâng cao

- **SHAP-IQ** [2]: Phát hiện tương tác.
- **TimeSHAP** [10]: Giải thích chuỗi thời gian.
- **GraphSHAP** [11]: Giải thích mạng nơ-ron đồ thị.
- **Asymmetric Shapley** [1]: Diễn giải nhân quả.

Tài liệu tham khảo I

[1] José Bento et al. "TimeSHAP: Explaining Recurrent Models with Time-Series Data". In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & (2021).

[2] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. "Feature Relevance Quantification in Explainable AI: A Causal Problem". In: International Conference on Artificial Intelligence and Statistics (2020).

[3] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". In: arXiv preprint arXiv:1802.03888 (2018).

[4] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: Advances in Neural Information Processing Systems 30 (2017).

Tài liệu tham khảo II

[5] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions with Deep Learning". In: arXiv preprint arXiv:1705.07874 (2018).

[6] Scott M. Lundberg et al. "From Local Explanations to Global Understanding with Explainable AI for Trees". In: Nature Machine Intelligence 2.1 (2020), pp. 56–67.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledg (2016), pp. 1135–1144.

[8] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". In: International Conference on Machine Learning (2017).

Tài liệu tham khảo III

[9] Mukund Sundararajan and Amir Najmi. "The Many Shapley Values for Model Explanation". In: International Conference on Machine Learning (2020).

[10] Hao Yuan et al. "Explainability in Graph Neural Networks: A Taxonomic Survey". In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[11] Hao Yuan et al. "On Explainability of Graph Neural Networks via Subgraph Exploration". In: International Conference on Machine Learning (2021).