Học máy khả diễn giải

Giang Nguyen

KAIST, June 2020

Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

@ChristophMolnar

Chương 2: Tính khả diễn giải

Tính khả diễn giải không được định nghĩa bởi một đại lượng hay công thức toán học. Một định nghĩa chung nhất đó là "khả diễn giải là khi con người có thể hiểu được nguyên nhân của một kết quả" . Một định nghĩa khác đó là "Khả diễn giải là khi con người có thể dự đoán kết quả của một mô hình dựa trên những hiểu biết về nó". Khi một mô hình có tính khả diễn giải càng cao, ta càng dễ dàng diễn giải một quyết định hay một dự đoán của nó. Trong tài liệu này, khái niệm khả diễn giải (explainability) và khả diễn giải (interpretability) được sử dụng tương đương.

0.1 Tại sao ta cần tính khả diễn giải?

Nếu một mô hình thực hiện công việc một cách hiệu quả, tại sao ta phải quan tâm tới những gì xảy ra bên trong nó? Một vấn đề rõ ràng đó là các bài toán hiện nay được thực hiện trên máy tính và kiểm chứng bằng các thông số nhất định. Ví dụ như trong bài toán phân loại thì độ chính xác sẽ là thước đo. Tuy nhiên, các thước đo này thường không thỏa mãn được các bài toán trong thực tế khi mà môi trường và dữ liệu có thể thay đổi và khác với dữ liệu trong quá trình huấn luyện.

Ta hãy cùng tìm hiểu sâu hơn về lý do tại sao tính khả diễn giải lại quan trọng như vậy. Khi nói đến việc dự đoán của mô hình học máy, có một sự đánh đổi (trade-off): Ta chỉ muốn biết **những gì** được dự đoán? Ví dụ, xác suất mà khách hàng sẽ bỏ đi hoặc mức độ hiệu quả của một số loại thuốc đối với bệnh nhân. Hay ta muốn biết **tại sao** dự đoán được đưa ra. Mức độ khả diễn giải mà ta có tỉ lệ nghịch với hiệu năng của mô hình? Trong một số trường hợp, ta không quan tâm tới nguyên nhận của một dự đoán, mà chỉ cần hiệu suất dự đoán trên tập dữ liệu kiểm tra (test dataset) là tốt. Nhưng trong các trường hợp khác, việc biết "tại sao" giúp ta đào sâu vào bài toán, dữ liệu, và lý do tại sao một mô hình không hoạt động như mong muốn. Một số mô hình có thể không yêu cầu tính khả diễn giải vì chúng được sử dụng trong môi trường rủi ro thấp, có nghĩa là một sai lầm sẽ không gây ra hậu quả nghiêm trọng, (ví dụ: hệ thống nhận xét/ đánh giá phim) hoặc những phương pháp đã được nghiên cứu và đánh giá rộng rãi (ví dụ: nhận dạng ký tự quang học). Nhu cầu về tính khả diễn giải nảy sinh do sự thiếu hoàn thiện trong quá trình chuẩn hóa vấn đề (Doshi-Velez và Kim 2017), nghĩa là đối với một số vấn đề hoặc công việc nhất định thì việc dự đoán (cái gì) là vẫn chưa đủ. Mô hình cũng phải giải thích cách đưa ra dự đoán (lý do tại sao), bởi vì một dự đoán đúng chỉ giải quyết một phần vấn đề gốc rễ. Những lý do sau đây dẫn dắt ta đến sự cần thiết của tính khả diễn giải (Doshi-Velez và Kim 2017 và Miller 2017).

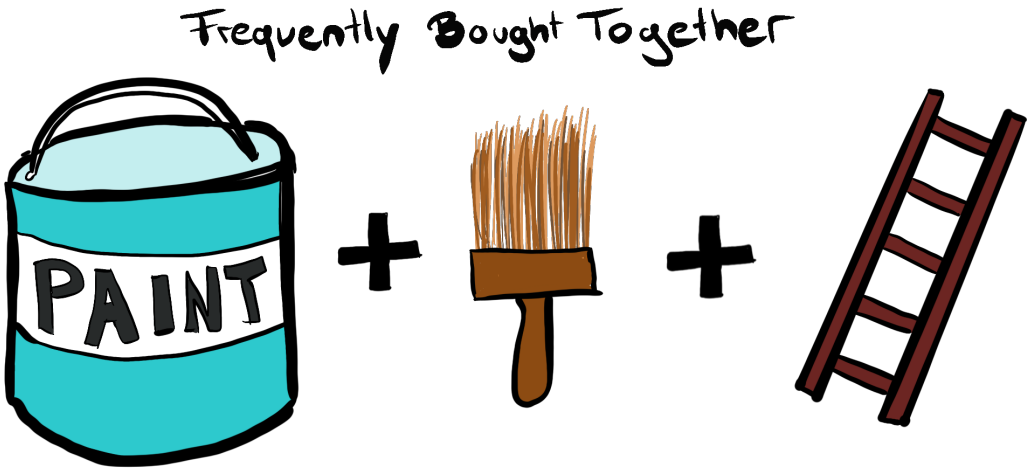
Sự tò mò và học hỏi Con người có bản năng tự cập nhật kiến thức khi có một sự kiện mới xảy ra xung quanh. Việc cập nhật kiến thức được thực hiện khi ta hiểu được câu trả lời tại sao cho các vấn đề đó. Giả dụ, khi một người cảm thấy mệt mỏi, anh ta sẽ tự hỏi "Tại sao mình bị ốm nhỉ?". Anh ta

sẽ nhớ lại việc mình làm mới đây. Anh ta nhận thấy rằng anh ta bị ốm mỗi khi ăn những quả dâu đỏ. Anh ta cập nhật nhận thức của mình và quyết định rằng dâu đỏ sẽ gây bệnh và tránh xa chúng.

Khi các mô hình học máy được sử dụng trong nghiên cứu, các phát hiện khoa học vẫn hoàn toàn bị che giấu nếu mô hình chỉ đưa ra dự đoán mà không có giải thích. Để tạo điều kiện cho việc học tập và thỏa mãn sự tò mò về lý do tại sao các dự đoán hoặc hành vi nhất định được tạo ra bởi máy móc, tính khả diễn giải và giải thích là rất quan trọng. Tất nhiên, con người không cần giải thích cho mọi thứ xảy ra. Việc hầu hết mọi người không hiểu cách hoạt động của máy tính là điều hoàn toàn bình thường. Tuy nhiên, các sự kiện bất ngờ khiến ta tò mò. Ví dụ: Tại sao máy tính của tôi tắt đột ngột?

Liên quan chặt chẽ đến học tập là mong muốn đi tìm ý nghĩa của sự vật sự việc. ta muốn hài hòa những mâu thuẫn hoặc sự thiếu nhất quán giữa các yếu tố trong vốn kiến thức của ta. "Tại sao con chó của tôi lại cắn tôi mặc dù nó chưa bao giờ làm như vậy trước đây"? Có một sự mâu thuẫn giữa hành vi trước đây của con chó và hành vi cắn người mới đây . giải thích của bác sĩ thú y gỡ bỏ mâu thuẫn của chủ nhân con chó: "Con chó đã bị căng thẳng nên mới cắn". Quyết định của máy móc càng ảnh hưởng đến cuộc sống của con người, thì chúng càng cần giải thích hành vi. Nếu một mô hình học máy từ chối đơn xin vay, điều này có thể hoàn toàn bất ngờ đối với những người nộp đơn. Họ chỉ có thể dung hòa sự mâu thuẫn này giữa kỳ vọng và thực tế bằng một số cách giải thích. Giải thích không thực sự phải trình bày đầy đủ về tình hình, nhưng nên chỉ ra nguyên nhân chính. Một ví dụ khác là bài toán đề xuất sản phẩm. Cá nhân tôi luôn nghĩ về lý do tại sao một số sản phẩm hoặc bộ phim nhất định được giới thiệu theo thuật toán cho tôi. Thông thường sẽ khá rõ ràng: Quảng cáo theo dõi tôi trên Internet bởi vì tôi mới mua một chiếc máy giặt, và tôi biết rằng trong những ngày tới tôi sẽ bị theo dõi bởi những quảng cáo về máy giặt. Đúng vậy, ta nên đề xuất găng tay nếu tôi đã có mũ mùa đông trong giỏ hàng của mình. Với thuật toán đề xuất phim, vì những người dùng đã thích các bộ phim khác mà tôi cũng thích nên bộ phim mà họ thích mới đây được đề xuất cho tôi. Càng ngày, các công ty Internet càng hoàn thiện việc giải thích cho các đề xuất của họ. Một ví dụ điển hình là các đề xuất về sản phẩm, dựa trên các kết hợp sản phẩm được mua thường xuyên:

Trong nhiều ngành khoa học, có sự thay đổi từ phương pháp định tính



Hình 1: Các sản phẩm được đề xuất mà thường xuyên được mua cùng nhau.

sang định lượng (ví dụ: xã hội học, tâm lý học) và cả xu hướng áp dụng học máy (sinh học, di truyền học). **Mục tiêu của khoa học** là thu thập kiến thức, nhưng nhiều vấn đề được giải quyết bằng dữ liệu lớn và mô hình học máy hộp đen. Bản thân mô hình trở thành nguồn kiến thức thay vì dữ liệu. Khả năng diễn giải giúp ta có thể trích xuất kiến thức mà mô hình đã học.

Các mô hình học máy đảm nhận các nhiệm vụ trong thế giới thực yêu cầu **tính an toàn** và đã được kiểm nghiệm. Hãy tưởng tượng một chiếc xe tự lái tự động phát hiện người đi xe đạp dựa trên hệ thống học sâu. Ta muốn chắc chắn 100% rằng các kiến thức trừu tượng mà hệ thống đã học không có vấn đề, bởi vì việc cán lên người đi xe đạp là không thể chấp nhận. Khi tai nạn xảy ra, giải thích có thể cho rằng đặc điểm đã học quan trọng nhất cho việc nhận dạng xe đạp là phát hiện được hai bánh của xe đạp và giải thích này giúp ta suy nghĩ về các trường hợp hy hữu như xe đạp có túi bên nên đã che một phần bánh xe.

Mặc định, các mô hình học máy thừa hưởng sai lệch (biases) từ dữ liệu huấn luyện. Điều này có thể biến các mô hình học máy của ta trở nên phân biệt chủng tộc, phân biệt đối xử với các nhóm thiểu số. Khả năng diễn giải là một công cụ gỡ lỗi hữu ích để **phát hiện thiên vị** trong các mô hình học máy. Có thể xảy ra trường hợp mô hình học máy mà ta đã huấn luyện để phê duyệt tự động hoặc từ chối các đơn đăng ký tín dụng phân biệt đối xử với một nhóm thiểu số đã bị tước quyền vay trong quá khứ. Mục tiêu chính của ta là chỉ cấp khoản vay cho những người có khả năng trả nợ. Sự thiếu

toàn vẹn của việc thiết lập vấn đề trong trường hợp này nằm ở chỗ ta không chỉ muốn giảm thiểu các khoản nợ xấu, mà còn muốn không phân biệt đối xử. Đây là một ràng buộc bổ sung nằm trong quy trình xác lập vấn đề của ta (cấp các khoản vay với rủi ro thấp và hợp lệ), nhưng nó không nằm trong hàm mất mát (loss function) mà mô hình học máy đã được tối ưu hóa.

Việc đưa máy móc và thuật toán vào trong cuộc sống đòi hỏi **tính khả diễn giải để được xã hội chấp nhận**. Con người gán niềm tin, ước muốn, ý định, v.v. cho các đối tượng. Robot là một ví dụ điển hình, như máy hút bụi của tôi, mà tôi đặt tên là "Doge". Nếu Doge bị mắc kẹt, tôi nghĩ: "Doge muốn tiếp tục dọn dẹp, nhưng muốn nhờ tôi giúp đỡ vì nó bị kẹt". Sau đó, khi Doge hoàn thành việc dọn dẹp và tìm kiếm tầng hầm để sạc pin, tôi nghĩ: "Doge muốn nạp năng lượng và có ý định tìm tầng hầm". Tôi cũng gán cho đặc điểm tính cách: "Doge hơi ngốc, nhưng theo một cách dễ thương". Đây là những suy nghĩ của tôi, đặc biệt là khi tôi phát hiện ra rằng Doge đã xô ngã một cái cây khi đang hút bụi nhà. Tôi sẽ dễ thông cảm cho Doge hơn. Một cỗ máy máy hoặc thuật toán mà giải thích các dự đoán của chúng sẽ được nhiều người chấp nhận hơn.

Giải thích được sử dụng để quản lý các tương tác xã hội. Bằng cách tạo ra một ý nghĩa chung, người giải thích ảnh hưởng đến hành động, cảm xúc và niềm tin của người nhận giải thích. Máy móc có thể cần phải định hình cảm xúc và niềm tin của ta để tương tác với ta. Máy móc phải "thuyết phục" ta, để chúng đạt được mục tiêu đã định. Tôi sẽ không hoàn toàn chấp nhận máy hút bụi của mình nếu nó không giải thích được hành vi của nó ở một mức độ nào đó. Máy hút bụi tạo ra giải thích nhằm tìm kiếm sự cảm thông, chẳng hạn như một "tai nạn" (như bị kẹt trên thảm nhà tắm ... một lần nữa) bằng cách giải thích rằng nó bị kẹt thay vì chỉ dừng lại làm việc mà không có giải thích. Điều thú vị là có thể có sự sai lệch giữa mục tiêu của cỗ máy (tạo niềm tin) và mục tiêu của người nhận (hiểu được dự đoán hoặc hành vi). Có lẽ giải thích đầy đủ cho lý do tại sao Doge bị mắc kẹt có thể là do pin rất yếu, một trong các bánh xe không hoạt động bình thường và có một lỗi lập trình khiến robot lặp đi lặp lại cùng một chỗ mặc dù gặp chướng ngại vật. Những lý do này (và một số lý do khác nữa) khiến robot gặp khó khăn, nhưng nó chỉ cần giải thích rằng có điều gì đó đang cản trở, điều đó đủ để tôi tin tưởng vào hành vi của nó và hiểu được ý nghĩa chung của sự vụ.

Các mô hình học máy chỉ có thể được gỡ lỗi **(debugged) và kiểm soát**

Hình 2: Doge, máy hút bụi của tôi, bị kẹt. Để giải thích cho vụ tai nạn, Doge nói với tôi rằng nó cần phải ở trên bề mặt phẳng.

(audited) khi chúng khả diễn giải. Thậm chí trong các môi trường rủi ro thấp, như mô hình đề xuất sản phẩm, tính khả diễn giải trở nên rất có giá trị cả trong và sau quá trình triển khai mô hình. Khi mô hình được sử dụng sau đó trong thực tế, các vấn đề có thể phát sinh. Tính khả diễn giải giúp ta hiểu được nguyên nhân của vấn đề, và từ đó, ta có giải pháp để bảo trì và sửa lỗi hệ thống. Một ví dụ khác như bộ phân loại chó husky và chó sói, nếu bộ phân loại này phân biệt nhầm một vài bức ảnh chó husky thành chó sói, bằng cách sử dụng các phương pháp học máy khả diễn giải, ta có thể nhận ra nguyên nhân của việc phân loại sai đến từ việc các bức ảnh đó có chứa tuyết (snow). Bộ phân loại, trong quá trình học tập, sử dụng tuyết như là một đặc trưng (feature) để phân loại một bức ảnh là chó sói. Việc này có thể có ý nghĩa về mặt phân biệt husky và chó sói trong tập dữ liệu huấn luyện, nhưng trong thực tế, đặc trưng này sẽ không có hiệu quả.

Khi một mô hình học máy có thể giải thích quyết định của nó, ta có thể kiểm tra những tiêu chí sau một cách dễ dàng (Doshi-Velez and Kim 2017):

1. Tính công bằng (Fairness): Đảm bảo rằng dự đoán không bị ảnh hưởng bởi sai lệch (bias) và có tình trạng "phân biệt đối xử" với các nhóm dữ liệu đặc biệt.
2. Tính riêng tư (Privacy): Đảm bảo rằng những thông tin nhạy cảm được bảo vệ.
3. Tính đáp ứng nhanh (Robustness): Đảm bảo rằng những thay đổi nhỏ ở

đầu vào không ảnh hưởng lớn tới kết quả đầu ra.

4. Tính nhân quả (Causality): Đảm bảo rằng chỉ những mối quan hệ có tính nhân quả được sử dụng.
5. Tính tin cậy (Trust): Đảm bảo mô hình đáng tin cậy hơn mô hình hộp đen.

Khi nào ta không cần tính khả diễn giải? Trong một số trường hợp, ta sẽ không cần hoặc thậm chí không muốn tính khả diễn giải của các mô hình học máy.

Khả diễn giải thực sự không cần thiết nếu mô hình khi hoạt động **không có tác động đáng kể lên đời sống**. Tưởng tượng rằng có 1 người tên Mike xây dựng một mô hình học máy để dự đoán bạn của anh ta sẽ đi đâu trong kì nghỉ tới dựa trên dữ liệu có sẵn của người bạn đó trên Facebook. Nếu mô hình dự đoán sai hoặc Mike không thể giải thích đầu ra của mô hình của mình, vấn đề sẽ chẳng có gì nghiêm trọng. Ta hoàn toàn ổn nếu không có tính khả diễn giải trong trường hợp này. Tuy nhiên, vấn đề sẽ thực sự nảy sinh nếu Mike muốn kinh doanh dựa trên mô hình này. Nếu mô hình hoạt động sai, công ty có thể bị thua lỗ hoặc mô hình có thể làm việc kém hiệu quả do sai lệch. Do đó, nếu mô hình có những tác động đáng kể tới tài chính hoặc xã hội, tính khả diễn giải trở nên cực kỳ quan trọng.

Khả diễn giải không cần thiết khi **vấn đề đã được tìm hiểu và phân tích một cách kỹ lưỡng**. Một số ứng dụng đã được nghiên cứu rộng rãi và tỉ mỉ để đưa vào thực tế ví dụ như nhận diện ký tự quang - OCR, xử lý hình ảnh từ phong bì và trích xuất địa chỉ. Các hệ thống này đã được nghiên cứu một cách kỹ lưỡng và các kết quả thực tế chứng minh tính hiệu quả của chúng. Ở đây, việc hiểu sâu hơn về vấn đề này đôi khi là không cần thiết.

Khả diễn giải có thể cho phép người dùng hoặc các chương trình **làm chủ các hệ thống** khi người dùng đánh lừa hệ thống bằng cách khai khác các lỗ hổng đến từ sự thiếu nhất quán giữa mục tiêu của người tạo ra và người sử dụng mô hình. Tính điểm tín dụng là một hệ thống như vậy bởi vì các ngân hàng muốn đảm bảo rằng các khoản vay chỉ được cung cấp cho những người nộp đơn có khả năng trả lại chúng, nhưng những người nộp đơn có mục tiêu nhận được khoản vay ngay cả khi ngân hàng không muốn. Sự thiếu nhất quán này tạo ra lỗ hổng để các ứng viên khai thác nhằm tăng cơ hội để đơn vay của họ được chấp nhận. Khi một ứng viên bị từ chối vay,

nếu biết việc có nhiều hơn 2 thẻ tín dụng ảnh hưởng tiêu cực đến điểm tín dụng, anh ta có thể hoàn trả lại thẻ thứ ba để cải thiện điểm và làm lại 1 thẻ mới sau khi gói vay được chấp nhận. Khi điểm tín dụng của ứng viên này tăng lên, xác suất thực tế để anh ta hoàn trả gói vay là không đổi. Hệ thống có thể bị lừa nếu dữ liệu vào ảnh hưởng tới các đặc trưng nhân quả - (causual features) mà không có tác động đến kết quả cuối cùng (khả năng hoàn trả khoản vay). Một ví dụ khác đó là Google phát triển một hệ thống gọi là Google Flu Trends để dự báo các đợt bùng phát cúm. Hệ thống này làm việc dựa các tìm kiếm trên Google liên quan tới bệnh cúm, tuy nhiên, hệ thống này làm việc không tốt. Phân bố của các truy vấn (queries) thay đổi khiến cho Google Flu Trends bỏ qua rất nhiều các đợt cúm bởi vì các "tìm kiếm trên Google không gây ra bệnh cúm". Khi người dùng tìm kiếm về các triệu chứng như sốt, rất ít khi người đó thực sự bị cúm. Một cách lý tưởng, các mô hình chỉ nên sử dụng các đặc trưng nhân quả (các đặc trưng thực sự ảnh hưởng tới đầu ra mong muốn).

0.2 Phân loại các phương pháp diễn giải

Các phương pháp cho học máy khả diễn giải có thể phân loại theo nhiều tiêu chí khác nhau.

Nội tại hay sau huấn luyện (Intrinsic or post-hoc)? Tiêu chí này phân loại liệu ta đạt được tính khả diễn giải qua việc giảm đi độ phức tạp của mô hình học máy nội tại hay qua áp dụng các phương pháp phân tích mô hình sau khi huấn luyện. Khả năng diễn giải intrinsic hay nội tại đề cập đến các mô hình học máy có tính khả diễn giải nhờ vào tính đơn giản của cấu trúc, ví dụ như cây quyết định ngắn (short decision trees) hoặc mô hình tuyến tính thưa (sparse linear models). Tính khả diễn giải sau huấn luyện đề cập việc áp dụng các phương pháp diễn giải sau khi huấn luyện mô hình. Ví dụ, sử dụng độ quan trọng của đặc trưng hoán vị (Permutation feature importance - PFI) là một phương pháp giải thích sau huấn luyện. Các phương pháp sau huấn luyện cũng có thể dùng trong các mô hình khả diễn giải nội tại. Ví dụ, PFI có thể dùng cho cây quyết định.

Kết quả của phương pháp diễn giải Các phương pháp diễn giải có thể phân loại dựa trên kết quả của chúng.

- **Thống kê tổng quan đặc trưng (Feature summary statistic)**: Nhiều phương pháp diễn giải cho ta các thống kê tổng quan cho mỗi đặc trưng. Nhiều phương pháp trả về một giá trị duy nhất với mỗi đặc trưng, ví dụ như độ quan trọng đặc trưng (feature importance), hay một kết quả phức tạp hơn, ví dụ như độ mạnh tương quan đặc trưng (feature interaction strength) bao gồm một số cho mỗi cặp đặc trưng.

- **Trực quan hoá tổng quan đặc trưng (Feature summary visualization)**: Đa số các thống kê tổng quan đặc trưng đều có thể được trực quan hoá. Một số các đặc trưng tổng quan chỉ có ý nghĩa nếu chúng được phác họa trực quan thay vì biểu diễn dưới dạng bảng. Một ví dụ đó là tính phụ thuộc riêng (partial dependence). Phác hoạ phụ thuộc riêng là các đường biểu diễn đặc trưng và trung bình dự đoán đầu ra. Cách tốt nhất để biểu diễn phụ thuộc riêng đó là vẽ chúng thành các đường thay vì đặt lên các trục tọa độ.

- **Các nội bộ mô hình (Model Internals)** (ví dụ như các trọng số đã được học): Tính khả diễn giải của các mô hình khả diễn giải mang tính nội tại nằm trong phần này. Các ví dụ là các trọng số của mô hình tuyến tính hoặc các cấu trúc cây đã được học (learned tree structures) (các đặc trưng và mức ngưỡng được dùng để phân nhánh cây) của các cây quyết định. Nội bộ mô hình và thống kê tổng quan đặc trưng khá giống nhau trong một vài trường hợp; ví dụ: mô hình tuyến tính, vì trọng số vừa là nội bộ của mô hình và vừa là thống kê tổng quan cho các đặc trưng. Một phương pháp khác để tạo ra nội bộ mô hình là phác hoạ các bộ phát hiện đặc trưng học được từ mạng nơ-ron tích chập (convolutional neural network). Các phương pháp diễn giải mà tạo ra nội bộ mô hình làm việc trên một số kiến trúc nhất định.

- **Điểm dữ liệu (Data Points)**: Mục này gồm tất cả các phương pháp trả về các điểm dữ liệu (đã tồn tại hoặc mới được sinh ra) để làm cho mô hình trở nên khả diễn giải. Một phương pháp trong số chúng được gọi là giải thích phản chứng (counterfactual explanations). Để giải thích dự đoán trên một mẫu dữ liệu (data instance), phương pháp này tìm một điểm dữ liệu tương tự bằng cách thay đổi một số đặc trưng mà làm cho kết quả dự đoán thay đổi (ví dụ: chuyển dự đoán từ 0 thành 1). Một ví dụ khác là việc xác định các nguyên mẫu (prototypes) của các lớp (class). Các phương pháp diễn giải mà trả về điểm dữ liệu mới phải đảm bảo điểm dữ liệu này là khả diễn giải. Các phương pháp này được dùng cho dữ liệu hình ảnh và văn bản, và ít hữu ích hơn cho dữ liệu dạng bảng vì chúng có rất nhiều đặc trưng.

- **Mô hình khả diễn giải nội tại (Intrinsically Interpretable Model)**: Ta giải thích các mô hình hộp đen bằng cách xấp xỉ chúng (toàn cục hoặc cục bộ) với một mô hình khả diễn giải nội tại (cây quyết định hoặc mô hình tuyến tính). Chúng sau đó được diễn giải bằng cách xem xét các tham số nội bộ mô hình hoặc thống kê tổng quan đặc trưng.

Mô hình cụ thể (specific) hay mô hình kiểu mẫu (agnostic)? . Phương pháp có thể áp dụng cho mọi loại mô hình học máy hay chỉ một số loại nhất định? Với phương pháp theo kiểu mô hình cụ thể, ví dụ như việc giải thích các trọng số hồi quy trong một mô hình tuyến tính là theo kiểu mô hình cụ thể, vì theo định nghĩa các mô hình khả diễn giải nội tại bản

thân chúng đang tự giải thích chúng rồi. Một ví dụ khác là các công cụ giải thích mạng nơ-ron. Các công cụ mẫu agnostic có thể dùng trên bất kỳ mô hình học máy nào và được áp dụng sau khi mô hình đó đã được huấn luyện. Các phương pháp agnostic hoạt động qua phân tích các cặp giá trị đầu vào và đầu ra của đặc trưng. Theo định nghĩa, các phương pháp này không được quyền truy cập vào nội bộ mô hình như trọng số hoặc cấu trúc mạng.

Cục bộ hay toàn cục? Phương pháp diễn giải một dự đoán riêng lẻ hay toàn bộ hành vi của mô hình hay đâu đó ở giữa đầu vào và đầu ra? Hãy tìm hiểu phạm vi của tính khả diễn giải trong phần tiếp theo.

0.3 Phạm vi của khả diễn giải

Một thuật toán tạo ra một mô hình đã được huấn luyện cho việc đưa ra các dự đoán. Mỗi bước trong thuật toán này có thể được sử dụng để đánh giá tính minh bạch hay khả diễn giải của mô hình.

0.3.1 Tính minh bạch của thuật toán

Cách thuật toán tạo ra mô hình?

Tính minh bạch của thuật toán là cách thuật toán học một mô hình từ dữ liệu và và các quan hệ nhân quả. Nếu ta sử dụng các mạng thần kinh tích chập để phân loại hình ảnh, ta có thể thấy rằng thuật toán học các bộ phát hiện cạnh (edge detectors) và bộ lọc (filters) tại các tầng (layer) thấp nhất. Đây chỉ là một trong các bước hoạt động của thuật toán. Tính minh bạch của thuật toán chỉ phân tích thuật toán mà ko quan tâm tới dữ liệu hoặc mô hình đã học. Tài này tập trung vào tính khả giễn giải thay vì sự minh bạch của các thuật toán. Các thuật toán như phương pháp bình phương tối thiểu (lease squares) cho các mô hình tuyến tính từ lâu đã được nghiên cứu cặn kẽ. Chúng được biết tới bởi tính minh bạch cao. Phương pháp học sâu (đưa gradient qua mạng với hàng triệu trọng số) vẫn chưa được nghiên cứu rõ ràng và đang thu hút sự quan tâm lớn lao của giới khoa học và chúng được coi là thiếu minh bạch, ít nhất cho tới thời điểm này.

0.3.2 Tính khả diễn giải toàn cục, toàn diện

Cách đưa ra dự đoán của một mô hình?

Ta có thể mô tả một mô hình theo cách hiểu của ta nếu ta có thể hiểu toàn bộ mô hình (Lipton 2016). Để giải thích đầu ra một cách toàn cục, ta cần huấn luyện mô hình, hiểu thuật toán và dữ liệu. Mức độ này của tính khả diễn giải là về cách hiểu mô hình đưa ra các quyết định, dựa trên cái nhìn toàn diện về các đặc trưng quan trọng của mô hình và từng thành phần đã học như trọng số, các tham số (parameters) khác, các cấu trúc và loại tương tác giữa các đặc trưng. Tính khả diễn giải toàn cục là rất khó để đạt được trong thực tế. Bất kì mô hình nào có số lượng trọng số lớn đều không phù hợp với bộ nhớ ngắn hạn của con người bình thường. Tôi tin rằng ta thực sự không thể tưởng tượng ra một mô hình tuyến tính với 5 đặc trưng bởi

vì điều đó có nghĩa là tưởng tượng trong không gian 5 chiều. Bất kì không gian đặc trưng nào có nhiều hơn 3 chiều thì đã quá giới hạn tưởng tượng của con người rồi. Thông thường, khi ta cố gắng hiểu một mô hình, ta chỉ xem xét các phần riêng lẻ của nó, chẳng hạn như các trọng số trong các mô hình tuyến tính.

0.3.3 Tính khả diễn giải toàn cục ở cấp độ mô đun

Cách các thành phần của mô hình ảnh hưởng tới dự đoán?

Một mô hình Naive Bayes với hàng trăm đặc trưng sẽ là quá tải đối với bộ nhớ của ta. Và ngay cả khi ta cố gắng nhớ tất cả các trọng số, ta cũng sẽ không thể nhanh chóng đưa ra các dự đoán đối với các điểm dữ liệu mới. Ngoài ra ta cần có phân phối chung của tất cả các đặc trưng trong đầu để ước tính tầm quan trọng của mỗi đặc trưng và mức độ các đặc trưng ảnh hưởng đến các dự đoán. Đây quả là một nhiệm vụ bất khả thi. Nhưng ta có thể dễ dàng hiểu một trọng số đơn (single weight). Trong khi tính khả diễn giải toàn cục thường nằm ngoài tầm với thì việc phân tích chúng ở cấp độ mô đun không phải lựa chọn tồi . Không phải tất cả các mô hình đều có thể giải thích được ở mức độ tham số hay mô đun. Đối với các mô hình tuyến tính, những phần có thể giải thích là trọng số. Với cây, ta có thể xem xét các nhánh (splits) và dự đoán trên nốt lá (leaf-node) . Các mô hình tuyến tính thoạt nhìn ta có thể giải thích hoàn hảo ở cấp độ mô đun tuy nhiên việc giải thích một trọng số đơn cần được đồng bộ với tất cả các trọng số khác. Việc giải thích của một trọng số đơn luôn đi kèm với chú thích rằng các đặc trưng đầu vào khác vẫn giữ nguyên giá trị, điều này không đúng với nhiều ứng dụng thực tế. Một mô hình tuyến tính dự đoán giá trị của một ngôi nhà, có tính đến cả kích thước của ngôi nhà và số lượng các phòng, có thể có một trọng số âm cho đặc trưng số phòng!!! Điều này hoàn toàn có thể xảy ra bởi vì đặc trưng diện tích có tương quan rất lớn với đặc trưng số phòng.. Trong một thị trường mà mọi người thích những căn phòng lớn, một ngôi nhà có ít phòng có thể có giá trị hơn một ngôi nhà có nhiều phòng nếu cả hai ngôi nhà có cùng kích thước. Những trọng số chỉ có ý nghĩa trong ngữ cảnh tồn tại các đặc trưng khác của mô hình. Dù vậy, những trọng số trong một mô hình tuyến tính vẫn có tính khả diễn giải cao hơn so với trọng số trong mạng nơ-ron sâu (deep neural networks).

0.3.4 Tính khả diễn giải cục bộ cho một dự đoán đơn lẻ

Cách mô hình tạo ra một dự đoán nhất định cho một mẫu dữ liệu (instance)?

Ta có thể xem xét với một mẫu dữ liệu và kiểm tra mô hình đưa ra dự đoán gì với dữ liệu đầu vào này và giải thích tại sao cho dự đoán đó. Nếu ta nhìn vào một dự đoán riêng lẻ, hành vi của mô hình phức tạp có thể được đơn giản hóa. Một cách cục bộ, dự đoán có thể chỉ phụ thuộc tuyến tính hoặc đơn điệu vào một số đặc trưng, thay vì phụ thuộc vào tất cả. Ví dụ, giá trị của một ngôi nhà có thể phụ thuộc phi tuyến vào kích thước của nó. Nhưng nếu ta chỉ nhìn vào một ngôi nhà 100 mét vuông cụ thể, có khả năng đối với tập con dữ liệu có kích thước là 100 mét vuông, dự đoán mô hình của ta phụ thuộc tuyến tính vào đặc trưng này. Ta có thể tìm ra điều này bằng cách mô phỏng dự đoán khi ta tăng hoặc giảm 10 mét vuông kích thước. Việc giải thích cục bộ có thể chính xác hơn việc giải thích toàn cục. Tài liệu này trình bày các phương pháp làm cho các dự đoán riêng lẻ trở nên khả diễn giải trong chương 5.

0.3.5 Tính khả diễn giải cục bộ cho một nhóm các dự đoán

Tại sao các mẫu dữ liệu có các dự đoán khác nhau?

Các dự đoán trên một nhóm mẫu có thể được giải thích bằng các phương pháp giải thích mô hình toàn cục (ở cấp độ mô đun) hoặc bằng các giải thích trên từng mẫu riêng lẻ. Các phương pháp toàn cục có thể được áp dụng bằng cách gom nhóm các trường hợp, xử lí chúng như thể nhóm là bộ dữ liệu hoàn chỉnh và sử dụng các phương pháp toàn cục với tập con này. Các phương pháp giải thích riêng lẻ có thể được sử dụng cho từng trường hợp và sau đó liệt kê hoặc gộp cho toàn bộ nhóm.

0.4 Đánh giá tính khả diễn giải

Trong cộng đồng học máy hiện nay, chưa có sự thống nhất thực sự về định nghĩa khả diễn giải là gì. Cũng không có 1 sự định lượng rõ ràng. Nhưng có 1 số nghiên cứu sơ khai và cố gắng hình thành vài hướng tiếp cận để đánh giá độ khả diễn giải, những hướng tiếp cận này sẽ được mô tả chi tiết hơn ở phần dưới.

Doshi-Velez and Kim (2017) đề xuất 3 tầng chính để đánh giá tính khả diễn giải:

Đánh giá ở mức độ ứng dụng (nhiệm vụ chính): đưa giải thích vào môi trường thực tế và để người dùng kiểm tra, đánh giá giải thích đó. Hãy tưởng tượng 1 ví dụ: Phần mềm phát hiện gãy xương có bao gồm 1 module học máy để định vị và đánh dấu các mảnh xương bị vỡ trên phim X-ray. Ở tầng ứng dụng, bác sĩ X quang sẽ sử dụng phần mềm phát hiện gãy xương và trực tiếp đánh giá chất lượng của giải thích đến từ mô hình học máy. Việc đánh giá này đòi hỏi việc cài đặt thí nghiệm phải tốt và bác sĩ có hiểu biết sâu về bài toán phát hiện gãy xương. Để so sánh, ta có thể dùng giải thích của con người, sau đó bác sĩ sẽ kiểm tra xem giải thích nào tốt hơn.

Đánh giá ở mức độ con người (Nhiệm vụ đơn giản) là một mức độ đánh giá ứng dụng đơn giản. Sự khác biệt là thử nghiệm đánh giá không yêu cầu kiến thức chuyên môn. Điều này giảm chi phí (đặc biệt là kiến thức chuyên môn về X-quang) và dễ dàng để tìm người kiểm thử. Một ví dụ là cho một người xem những giải thích khác nhau và để người đó lựa chọn giải thích tốt nhất.

Đánh giá ở mức độ chức năng (nhiệm vụ ủy quyền) không yêu cầu con người. Phương pháp này hoạt động tốt nhất khi vấn đề đã được kiểm nghiệm trước đó bởi con người. Ví dụ với cây quyết định, cây ngắn hơn sẽ có tính khả diễn giải tốt hơn (khoa học đã chứng minh trước đó cây ngắn hơn là khả diễn giải hơn).

Phần tiếp theo tập trung vào việc đánh giá giải thích cho các dự đoán riêng lẻ ở cấp độ chức năng. Các thuộc tính của giải thích mà ta sẽ xem xét để đánh giá chúng là gì?

0.5 Thuộc tính của giải thích

Ta muốn giải thích các dự đoán của một mô hình học máy. Để đạt được điều này, ta dựa vào một số phương pháp, trong đó các thuật toán tạo ra các giải thích. **Một cách giải thích liên hệ giá trị đặc trưng của một mẫu dữ liệu tới dự đoán theo cách con người có thể hiểu được**. Các cách giải thích khác bao gồm đưa ra một tập các mẫu dữ liệu (ví dụ phương pháp knn). Ví dụ: ta có thể dự đoán nguy cơ ung thư bằng máy vectơ hỗ trợ (SVM) và giải thích dự đoán bằng phương pháp đại diện cục bộ (local surrogate), tạo ra các cây quyết định thay cho giải thích. Hoặc ta có thể sử dụng mô hình hồi quy tuyến tính thay vì máy vectơ hỗ trợ (SVM) và mô hình hồi quy tuyến tính đã có phương pháp giải thích (giải thích các trọng số).

Ta hãy xem xét kỹ hơn các tính chất của các phương pháp giải thích và giải thích (Robnik-Sikonja và Bohanec, 2018). Các tính chất này có thể được sử dụng để đánh giá chất lượng của phương pháp giải thích hoặc của giải thích. Tất nhiên ta không thể đo lường một cách hoàn hảo, do vậy khó nhất là chuẩn hóa cách để đo lường chất lượng này.

Thuộc tính của các phương pháp giải thích:

- **Sức biểu đạt** là "ngôn ngữ" hay cấu trúc của các giải thích mà phương pháp có thể tạo ra. Một phương pháp giải thích có thể tạo ra các quy tắc IF-THEN, cây quyết định, tổng trọng số, ngôn ngữ tự nhiên hoặc bất kỳ thứ gì.

- **Độ minh bạch** mô tả phương pháp giải thích dựa vào mô hình học máy như thế nào, ví dụ như việc sử dụng các tham số của mô hình. Ví dụ, các phương pháp giải thích dựa trên các mô hình khả diễn giải nội tại như mô hình hồi quy tuyến tính có độ minh bạch rất cao. Các phương pháp chỉ dựa vào đầu vào và đầu ra có độ minh bạch bằng 0. Tùy thuộc vào từng vấn đề, ta có thể yêu cầu độ minh bạch khác nhau. Khi độ minh bạch cao, phương pháp có thể dùng nhiều thông tin hơn để tạo ra các giải thích. Khi độ minh bạch thấp, giải thích sẽ mô phạm và chung chung hơn.

- **Tính linh động** mô tả phạm vi của các loại mô hình học máy mà phương

pháp giải thích có thể được áp dụng. Các phương pháp có độ minh bạch thấp có tính linh động cao hơn vì chúng coi mô hình học máy như một hộp đen. Các mô hình đại diện (surrogate models) có thể là phương pháp giải thích với tính linh động cao nhất. Các phương pháp chỉ hoạt động cho một loại mạng cụ thể (ví dụ: mạng nơ-ron hồi tiếp) có tính di động thấp.

- **Độ phức tạp thuật toán** mô tả độ phức tạp tính toán của phương pháp tạo ra giải thích.Tính chất này rất quan trọng bởi vì tính toán đang là một vấn đề lớn trong việc tạo ra các giải thích với các phương pháp hiện nay.

Thuộc tính của các giải thích đơn lẻ:

- **Độ chính xác:** Nếu sử dụng giải thích, chất lượng của dự đoán đầu ra cho một mẫu dữ liệu mới là như thế nào? Đặc biệt nếu giải thích được sử dụng thay vì mô hình học máy, độ chính xác của giải thích là cực kỳ quan trọng. Tuy vậy, nếu mô hình học máy có độ chính xác thấp, ta cũng không thể đòi hỏi giải thích mang lại độ chính xác cao, nếu mục đích chỉ là để giải thích mô hình hộp đen. Trong trường hợp này, sự minh bạch là quan trọng hơn cả.

- **Độ trung thực:** Giải thích có thể xấp xỉ dự đoán của mô hình hộp đen như thế nào? Độ trung thực cao là một trong những tính chất quan trọng nhất của giải thích, bởi vì giải thích với độ trung thực thấp là vô ích trong việc giải thích mô hình học máy. Độ chính xác và độ trung thực có liên quan chặt chẽ. Nếu mô hình hộp đen có độ chính xác cao và giải thích có độ trung thực cao, thì giải thích cũng có độ chính xác cao. Một số giải thích chỉ cung cấp độ trung thực cục bộ, nghĩa là giải thích chỉ gần đúng với dự đoán mô hình cho một tập hợp con của dữ liệu (ví dụ: mô hình đại diện cục bộ - local surrogate models) hoặc thậm chí chỉ cho một mẫu dữ liệu riêng lẻ (ví dụ: Giá trị Shapley).

- **Tính nhất quán:** Các giải thích khác nhau như thế nào giữa các mô hình đã được huấn luyện cho cùng một nhiệm vụ và tạo ra các dự đoán tương tự nhau? Ví dụ, ta huấn luyện một máy vectơ hỗ trợ và mô hình hồi quy tuyến tính trên cùng một nhiệm vụ và cả hai đều tạo ra các dự

đoán rất giống nhau. Ta tạo ra các giải thích bằng một phương pháp đã chọn và phân tích sự khác biệt giữa các giải thích. Nếu các giải thích giống nhau, chúng có sự nhất quán. Điều này khá khó xảy ra, vì hai mô hình có thể sử dụng các đặc trưng khác nhau, nhưng có được các dự đoán tương tự (còn được gọi là "Hiệu ứng Rashomon"). Trong trường hợp này, ta không mong đợi tính nhất quán cao vì các giải thích là khác nhau. Tính nhất quán cao chỉ xuất hiện khi các mô hình sử dụng các đặc trưng giống nhau.

- **Tính ổn định:** Các giải thích tương đồng ra sao nếu các mẫu dữ liệu là tương đồng? Trong khi tính nhất quán so sánh các giải thích giữa các mô hình, độ ổn định so sánh các giải thích giữa các mẫu tương tự cho một mô hình cố định. Độ ổn định cao có nghĩa là các thay đổi nhỏ trong các đặc trưng của một thể hiện không làm thay đổi đáng kể giải thích (trừ khi các thay đổi nhỏ này cũng thay đổi mạnh mẽ dự đoán). Một phương pháp giải thích thiếu ổn định có thể do nó có phương sai lớn (high variance). Nói cách khác, phương pháp giải thích bị ảnh hưởng mạnh mẽ bởi những thay đổi nhỏ trong các giá trị đặc trưng của dữ liệu được giải thích. Sự thiếu ổn định cũng có thể được gây ra bởi các thành phần bất định của phương pháp giải thích, chẳng hạn như bước lấy mẫu dữ liệu, giống như trong phương pháp đại diện cục bộ. Ta luôn mong muốn độ ổn định cao.

- **Tính dễ hiểu:** Con người hiểu những giải thích như thế nào? Tính chất này trông có vẻ tầm thường nhưng kỳ thực lại là một vấn đề cực kỳ to lớn. Ta rất khó xác định và đo lường tính chất này. Nhiều người đồng ý rằng tính dễ hiểu phụ thuộc vào người dùng. Các ý tưởng để đo lường mức độ dễ hiểu bao gồm đo kích thước của giải thích (số đặc trưng có trọng số khác không trong mô hình tuyến tính, số luật quyết định, ...) hoặc kiểm tra xem người dùng có thể dự đoán hành vi của mô hình học máy bằng cách sử dụng giải thích hay không. Tính dễ hiểu của các đặc trưng được sử dụng trong giải thích cũng cần được cân nhắc kĩ lưỡng. Các biến đổi phức tạp của các đặc trưng cũng có thể làm giảm tính dễ hiểu của chúng so với ban đầu.

- **Sự chắc chắn:** Giải thích có phản ánh sự chắc chắn của mô hình học máy không? Nhiều mô hình học máy chỉ đưa ra dự đoán mà không đưa ra thông tin về độ tin cậy. Nếu mô hình dự đoán xác suất ung thư là 4%

cho một bệnh nhân, và cũng 4% cho một bệnh nhân với các giá trị đặc trưng khác, dự đoán nào là chắc chắn hơn? Giải thích bao hàm sự chắc chắn của mô hình sẽ rất hữu ích.

- **Mức độ quan trọng:** Giải thích phản ánh tầm quan trọng của các đặc trưng hoặc các phần của giải thích như thế nào? Ví dụ: nếu một luật quyết định được tạo ra như giải thích cho một dự đoán riêng lẻ, thì điều kiện nào của luật này là quan trọng nhất?

- **Tính mới:** Giải thích có phản ánh liệu một mẫu dữ liệu đến từ phân phối khác với phân phối của dữ liệu huấn luyện hay không? Trong trường hợp này, mô hình có thể hoạt động không chính xác và giải thích có thể là vô nghĩa. Khái niệm về tính mới có liên quan đến khái niệm về sự chắc chắn. Tính mới càng cao, càng có nhiều khả năng mô hình sẽ có độ chắc chắn thấp do thiếu dữ liệu.

- **Tính đại diện:** Giải thích có thể phủ (cover) bao nhiêu mẫu dữ liệu? Giải thích có thể bao trùm toàn bộ mô hình (ví dụ: giải thích các trọng số trong mô hình hồi quy tuyến tính) hoặc chỉ đại diện cho một dự đoán riêng lẻ (ví dụ: Giá trị Shapley).

0.6 Giải thích thân thiện với con người

Ta cùng đào sâu hơn và khám phá cách mà chúng coi một diễn giải là tốt và ý nghĩa thực sự của học máy khả diễn giải. Các nghiên cứu bên ngành nhân văn có thể giúp ta tìm ra câu trả lời. Miller (2017) đã tiến hành một cuộc khảo sát có quy mô lớn trên các công bố khoa học về sự diễn giải, và phần này được dựa trên tổng hợp của Miller.

Trong mục này, ta lưu ý một số điểm như sau: Với một diễn giải cho một sự kiện, con người ưu tiên những diễn giải ngắn (chỉ một hoặc hai nguyên nhân) tương phản với tình huống hiện tại. Đặc biệt những nguyên nhân bất thường sẽ mang lại những giải thích tốt. Diễn giải là tương tác xã hội giữa người giải thích với người được giải thích (người nhận giải thích) và do đó bối cảnh xã hội có ảnh hưởng to lớn đến nội dung thực tế của diễn giải.

Khi ta cần giải thích sử dụng tất cả đặc trưng cho một dự đoán hay một hành vi cụ thể, ta không cần giải thích thân thiện, mà chỉ cần nó mang tính nhân quả tuyệt đối. Để có thể gỡ lỗi của mô hình, việc nắm bắt các thành phần nhân quả là thực sự cần thiết.

0.6.1 Giải thích là gì?

Một lời diễn giải là **câu trả lời cho câu hỏi tại sao (Miller 2017)**

- Tại sao việc điều trị không hiểu quả trên bệnh nhân?
- Tại sao khoản vay của tôi lại bị từ chối?
- Tại sao ta chưa được liên lạc với người ngoài hành tinh?

Hai câu hỏi đầu tiên có thể được trả lời với giải thích "thông thường", trong khi đó câu hỏi thứ ba thuộc danh mục "Câu hỏi hiện tượng khoa học tổng quát và triết học" hơn. Ta tập trung vào những loại diễn giải "thông thường", bởi vì chúng liên quan đến khả năng giải thích của học máy. Các câu hỏi bắt đầu bằng "như thế nào" có thể thường được chuyển thể sang câu hỏi "tại sao": "Làm thế nào mà khoản vay của tôi lại bị từ chối?" có thể trở thành "Tại sao khoản vay của tôi bị từ chối?".

Theo đó, thuật ngữ "Diễn giải (Explanation)" đề cập đến quy trình để tạo ra nhận thức, nhưng cũng là sản phẩm của quá trình đó. Bộ diễn giải có thể là con người hay máy móc.

0.6.2 Thế nào là giải thích tốt?

Phần này tổng hợp cô đọng các kết quả của Miller về diễn giải "tốt" và bổ sung ý nghĩa cho học máy khả diễn giải.

Giải thích mang tính tương phản (Explanations are contrastive) (Lipton 1990). Con người thường không hỏi tại sao một dự đoán cụ thể nào đó được tạo ra, mà thường thắc mắc tại sao lại có dự đoán này mà không phải một dự đoán nào khác thay thế. Ta thường có xu hướng suy nghĩ về những trường hợp đối lập. Ví dụ, "dự đoán được đưa ra sẽ như thế nào nếu đưa vào một đầu vào X khác với trước đó". Đối với dự đoán giá nhà, người sở hữu căn nhà có thể quan tâm tại sao giá dự đoán cao hơn so với giá cả mà họ trông đợi. Nếu đơn vay nợ của tôi bị từ chối, tôi không quan tâm đến tất cả các yếu tố dẫn đến việc đơn vay bị từ chối . Tôi chỉ quan tâm đến các yếu tố mà trong đó mà đơn vay của tôi cần thay đổi để nhận được khoản vay. Tôi muốn biết sự đối lập giữa đơn vay của tôi và đơn đã được chấp thuận. Ta thấy rằng các giải thích tương phản (contrasting explanations) là một phát hiện quan trọng cho học máy khả diễn giải. Đối với hầu hết các mô hình khả diễn giải, ta có thể trích xuất các giải thích tương phản một dự đoán của một mẫu bằng các mẫu dữ liệu đối kháng. Bác sĩ có thể thắc mắc: "Tại sao thuốc này lại không hiệu quả trên bệnh nhân của tôi?". Và họ có thể muốn giải thích mà có yếu tố tương phản giữa bệnh nhân của họ với một bệnh nhân mà thuốc đó có hiệu quả và có sự tương đồng đối với những bệnh nhân không có tác dụng khác. Các giải thích tương phản là dễ dàng để hiểu hơn các giải thích đầy đủ. Giải thích đầy đủ cho câu hỏi của bác sĩ trước đó tại sao thuốc lại không hiệu bảo có thể bao gồm: Bệnh nhân đó đã có bệnh duy trì suốt 10 năm, 11 genes là đã có biểu hiện quá mức, cơ thể bệnh nhân nhanh chóng phân giải thuốc thành các thành phần hóa học nên không còn hữu hiệu,... Giải thích tương phản đơn giản hơn nhiều: Đối lập với các bệnh nhân có phản ứng với thuốc, những bệnh nhân không phản ứng có sự kết hợp các gene nhất định làm cho thuốc kém hiệu quả đi. Giải thích tốt nhất là giải thích mà làm nổi bật sự khác nhau giữa đối tượng quan tâm với đối tượng tham chiếu đến. **Điều này có ý nghĩa gì tới học máy khả diễn giải:** Con người không muốn một giải thích đầy đủ cho một dự đoán, nên tạo ra các giải thích tương phản sẽ phụ thuộc vào ứng dụng vì nó yêu cầu sự so sánh. Và điều này có thể phụ thuộc vào các điểm dữ liệu được diễn giải,

nhưng cũng phụ thuộc vào người nhận sự giải thích. Một người dùng trên một trang web dự đoán giá nhà có thể muốn có một sự giải thích cho những giá nhà đối lập với ngôi nhà họ sở hữu hoặc có thể ngôi ngồi nhà nào khác trên trang web hoặc có thể với một ngôi nhà phổ thông trong khu vực lân cận. Giải pháp cho việc tạo tự động các giải thích tương phản có thể cũng liên quan đến việc tìm các nguyên mẫu hay mẫu đặc trưng trong dữ liệu.

Giải thích mang tính chọn lọc (Explanations are selected). Con người không mong đợi những giải thích mà bao quát thực tế và nguyên nhân đầy đủ của một sự kiện. Ta đã từng lựa chọn ra một hoặc hai nguyên nhân từ một số những nguyên nhân khác nhau rồi tạo ra giải thích. Thực tế, điều này diễn ra hàng ngày trên các bản tin TV:

- "Sự tụt giảm giá cổ phiếu được cho là có nguyên nhân từ những phản ứng dữ dội xuất phát từ bản cập nhật phần mềm mới nhất của công ty."

- "Tsubasa và đội của anh ấy đã thua trận vì yếu kém trong khâu phòng thủ: họ đã để đối thủ tự do có nhiều chỗ trống để thực hiện chiến thuật của mình."

- "Sự không tin tưởng ngày càng gia tăng vào các tổ chức thành lập và chính phủ của chúng ta là những yếu tố chính mà làm giảm đi tỷ lệ cử tri bầu cử."

Việc một sự kiện có thể được giải thích bằng nhiều nguyên nhân khác nhau được gọi là Hiệu ứng Rashomon. Rashomon là một bộ phim Nhật kể những câu chuyện tương phản, mâu thuẫn (giải thích) về cái chết của một Samurai. Đối với các mô hình học máy, điều này thuận lợi nếu một dự đoán tốt có thể tạo nên từ những đặc trưng khác nhau. Việc kết hợp nhiều mô hình (ensemble) với các đặc trưng khác nhau (các giải thích khác nhau) thường hoạt động tốt vì tính trung bình trên các "câu chuyện" đó làm cho các dự đoán linh hoạt và chính xác hơn. Nhưng nó cũng có nghĩa rằng ta có nhiều giải thích có chọn lọc hơn với một dự đoán nhất định được đưa ra. **Điều này có ý nghĩa gì đến học máy khả diễn giải:** Nó làm cho việc diễn giải ngắn gọn, chỉ một đến ba lý do, thậm chí nếu ngữ cảnh có phức tạp hơn. Phương pháp LIME (phần 5.7) vẫn hoạt động tốt.

Những giải thích mang tính xã hội (Explanations are social). Chúng là một phần trong hội thoại hay tương tác giữa người giải thích và

người nhận sự giải thích đó. Bối cảnh xã hội quyết định nội dung và bản chất của các giải thích. Nếu tôi muốn giải thích cho một người am hiểu công nghệ tại sao tiền điện tử lại có giá trị rất lớn, tôi sẽ nói những điều như là: "Các tính chất phi tập trung (decentralized), phân phối (distributed), sổ cái dựa trên blockchain, không thể bị quản lý bởi một trung tâm nào, cộng hưởng với đó là con người có thể bảo mật tài sản của họ, điều này giải thích tại sao nó có nhu cầu và giá cả cao". Nhưng để giải thích với bà của tôi thì tôi sẽ nói rằng: "Bà nghe này: tiền điện tử cũng có chút giống như là đồng tiền vàng máy tính. Con người thích và trả nhiều thứ bằng vàng, còn giới trẻ thích và trả nhiều thứ bằng tiền vàng máy tính." **Điều này có ý nghĩa gì đến học máy khả diễn giải:** Hãy chú ý đến môi trường xã hội của ứng dụng học máy và đối tượng ta hướng đến.

Các giải thích tập trung vào sự bất thường (Explanations focus on the abnormal). Con người tập trung nhiều hơn vào các nguyên nhân bất thường để giải thích các sự kiện (Kahnemann and Tversky, 1981). Đây là những nguyên nhân có xác suất nhỏ nhưng vẫn xảy ra. Việc loại bỏ những nguyên nhân bất thường này sẽ làm thay đổi rất nhiều kết quả (giải thích tương phản). Con người xem xét những nguyên nhân "bất thường" và xem chúng như những giải thích tốt. Một ví dụ từ Štrumbelj và Kononenko (2011) là: Giả sử ta có một bộ dữ liệu về việc kiểm tra của giáo viên đối với học sinh. Học sinh tham gia vào một khóa học và vượt qua khóa học ngay sau khi trình bày thành công một bài thuyết trình. Giáo viên có tùy chọn để hỏi thêm học sinh các câu hỏi để kiểm tra kiến thức. Những học sinh nào không trả lời được những câu hỏi đó sẽ trượt khóa học. Các học sinh có thể có các mức độ chuẩn bị khác nhau, nó chuyển thành các xác suất khác nhau cho việc trả lời chính xác các câu hỏi của giáo viên (nếu họ quyết định kiểm tra học sinh đó). Ta muốn dự đoán học sinh nào sẽ qua môn học và giải thích các dự đoán đó. Cơ hội để vượt qua là 100% nếu giáo viên không có thêm câu hỏi nào, hoặc hoặc xác suất của việc qua môn học phụ thuộc vào sự chuẩn bị của học sinh và kết quả xác suất của việc trả lời câu hỏi chính xác. Kịch bản 1: Giáo viên thường hỏi các học sinh các câu hỏi bổ sung (ví dụ 95 trong 100 lần). Một học sinh không học (10% cơ hội để vượt qua phần câu hỏi) không may bị hỏi và học sinh đó trả lời sai. Tại sao học sinh đó rớt? Ta sẽ cho rằng vì lỗi của học sinh đó là không học bài. Kịch bản 2: Giáo viên hiếm khi hỏi thêm các câu hỏi (ví dụ 2 trong 100 lần). Đối với một sinh không học bài, ta

sẽ dự đoán một xác suất cao cho việc đỗ môn học vì khó có khả năng học sinh đó bị hỏi thêm. Tất nhiên, một trong các học sinh không chuẩn bị cho các câu hỏi, học sinh đó cũng có 10% cơ hội vượt qua các câu hỏi. Khi học sinh không may bị hỏi thêm câu hỏi, học sinh đó không trả lời đúng và rớt môn. Vậy có lý do nào cho việc trượt môn đó? Ta sẽ cho rằng giải thích tốt nhất đó là "vì giáo viên đó đã kiểm tra học sinh". Vì nhiều khả năng là giáo viên không kiểm tra, vậy nên hành động của giáo viên lúc đó là bất thường. **Điều này có ý nghĩa gì đến học máy khả diễn giải:** Nếu một trong các đặc trưng đầu vào cho một dự đoán là bất thường trong bất cứ hoàn cảnh nào (ví dụ một loại đặc trưng hiếm trong tập các đặc trưng) và các đặc trưng đó ảnh hưởng đến dự đoán, nó nên xuất hiện trong giải thích, thậm chí nếu các đặc trưng "bình thường" khác có cùng ảnh hưởng như đặc trưng bất thường. Một đặc trưng bất thường trong ví dụ dự đoán giá nhà có thể là một ngôi nhà khá đắt có hai ban công. Thậm chí nếu ta biết hai ban công có mức độ ảnh hưởng đến giá nhà tương đương với các đặc trưng: diện tích trung bình, hàng xóm tốt, mới được cải tạo; giá trị đặc trưng bất thường này vẫn sẽ là giải thích tốt nhất cho việc tại sao giá nhà lại cao như vậy.

Các giải thích mang tính trung thực (Explanations are truthful). Giải thích tốt cần được chứng minh là đúng trong thực tế (và các tình huống khác). Nhưng trớ trêu, đây không phải là yếu tố quan trọng nhất cho giải thích tốt. Ví dụ, tính chọn lọc dường như quan trọng hơn tính trung thực. Giải thích mà chọn chỉ một hoặc hai nguyên nhân khả thi hiếm khi bao quát hết danh sách các nguyên nhân có liên quan. Tính chọn lọc bỏ quan một phần sự thực. Chỉ một hay hai nguyên nhân là không đúng; ví dụ nếu thị trường chứng khoán sụp đổ, có cả triệu nguyên nhân dẫn đến vấn đề này. **Điều này có ý nghĩa gì đến học máy khả diễn giải:** Sự giải thích nên dự đoán sự kiện một cách trung thực nhất có thể, cũng như trong học máy đôi khi được gọi là tính trung thực hoặc minh bạch (fidelity). Vậy nên nếu ta nói rằng hai ban công làm tăng giá nhà, thì điều này cũng nên đúng đối với các căn nhà khác (hoặc ít nhất các nhà tương đồng). Đối với con người, tính trung thực (fidelity) của giải thích là không quan trọng bằng tính chọn lọc, tương phản, và xã hội.

Giải thích tốt mang tính nhất quán với niềm tin trước đó của người được giải thích (Good explanations are consistent with prior beliefs

of the explainee). Con người có xu hướng bỏ qua các thông tin không phù hợp với niềm tin có sẵn của họ. Vấn đề này được gọi là sai lệch xác nhận (confirmation bias) (Nickerson 1998). Con người sẽ có xu hướng đánh giá thấp hoặc bỏ qua những giải thích không nhất quán với niềm tin của họ. Các niềm tin này khác nhau với mỗi người, nhưng cũng có những niềm tin thống nhất như thế giới quan về chính trị. **Điều này có ý nghĩa gì đến học máy khả diễn giải:** Các giải thích tốt cần phù hợp với các niềm tin trước đó của người nhận giải thích. Điều này khó có thể áp dụng lên học máy và có lẽ sẽ ảnh hưởng lớn tới hiệu suất dự đoán. Niềm tin trước của ta về sự ảnh hưởng của kích thước căn nhà lên giá nhà là căn nhà càng lớn thì giá càng cao. Giả sử với một số ngôi nhà, diện tích lớn có tác động tiêu cực tới giá trị của chúng. Mô hình đã học được điều này trong huấn luyện (do một số tương tác phức tạp), nhưng điều này mâu thuẫn mạnh mẽ với niềm tin trước đó của ta.

Giải thích tốt mang tính tổng quan và hiển nhiên (Good explanations are general and probable). Một nguyên nhân có thể được sử dụng để giải thích cho nhiều sự kiện là rất khái quát tổng thể và có thể được coi là giải thích tốt. Lưu ý rằng điều này là mâu thuẫn với nhận định rằng các nguyên nhân bất thường tạo nên các giải thích tốt. Theo tôi nhận thấy, nguyên nhân bất thường cần được ưu tiên hơn nguyên nhân khái quát. Nguyên nhân bất thường hiếm khi xuất hiện trong các kịch bản. Trong trường hợp không có sự tồn tại của sự kiện bất thường, giải thích khái quát là được xem xét như là giải thích tốt. Cũng nên nhớ rằng con người có xu hướng đánh giá nhầm xác suất của các sự kiện tham gia vào. (Joe là một người trông coi thư viện, có khả năng anh ấy là một người nhút nhát hay nhút nhát dẫn đến thích đọc sách nên muốn làm người trông coi thư viện?) Một ví dụ tốt là "Căn nhà là đắt đỏ vì nó to", điều này khá phổ biến khái quát, giải thích tốt cho câu hỏi tại sao căn nhà lại đắt hoặc rẻ. **Điều này có ý nghĩa gì đến học máy khả diễn giải:** Tính khái quát tổng quan có thể dễ dàng được đo lường bởi các hỗ trợ đặc trưng, tính bằng số các mẫu mà giải thích được áp dụng chia cho tổng số mẫu.