Chapter 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, tính khả giải thích (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.

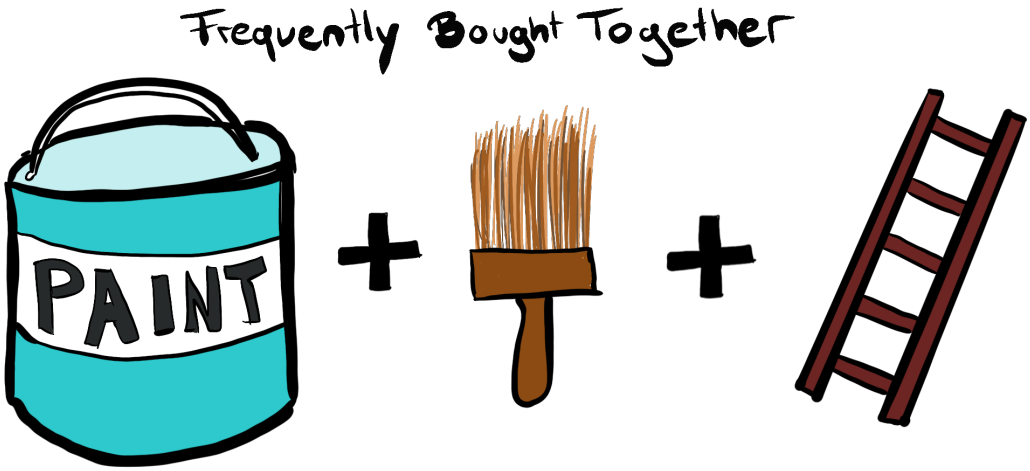
Hãy cùng chúng tôi tìm hiểu sâu hơn về lý do tại sao tính khả giải thích lại quan trọng như vậy. Khi nói đến mô hình dự đoán, bạn phải đánh đổi: Bạn chỉ muốn biết **những gì** được dự đoán? Ví dụ, xác suất mà khách hàng sẽ bỏ qua hoặc mức độ hiệu quả của một số loại thuốc đối với bệnh nhân. Hoặc bạn có muốn biết **tại sao** dự đoán được đưa ra và mức bù trừ có thể cho tính khả diễn giải bằng việc giảm hiệu suất dự đoán? Trong một số trường hợp, bạn không quan tâm tại sao lại đưa ra quyết định, chỉ cần biết rằng hiệu suất dự đoán trên tập dữ liệu kiểm tra (test dataset) là tốt. Nhưng trong các trường hợp khác, biết "tại sao" giúp bạn tìm hiểu thêm về vấn đề, dữ liệu và lý do tại sao một mô hình không hoạt động như mong muốn. Với một số mô hình, tính khả diễn giải có thể không cần thiết do chúng được sử dụng trong các môi trường rủi ro thấp (low-risk environment), nghĩa là các vấn đề của mô hình trong quá trình hoạt động sẽ không dẫn tới những hậu quả nghiêm trọng (ví dụ như một hệ thống gợi ý sản phẩm) hay các bài toán đã được nghiên cứu kĩ lưỡng trước đó (ví dụ như nhận dạng ký tự quang học - optical character recognition). Nhu cầu về tính khả diễn giải đến từ sự thiếu hoàn thiện trong quá trình thiết lập vấn đề (Doshi-Velez và Kim 2017), có nghĩa là đối với một số vấn đề hoặc nhiệm vụ nhất định thì dự đoán không thôi là chưa đủ (cái gì). Mô hình cũng phải giải thích cách nó tạo ra dự đoán (tại sao), bởi vì một dự đoán đúng chỉ giải quyết một phần vấn đề được đặt ra. Những lý do sau đây thúc đẩy nhu cầu cho tính khả giải thích (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, có thể do ăn thức ăn lạ, hoặc anh ta mới dính một cơn mùa đầu mùa. Người này sẽ cập nhật "mô hình" để tránh ăn thức ăn lạ đó hoặc sẽ mang ô khi trời mưa.

Khi các mô hình máy học 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 lời giải thích. Để tạo điều kiện 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 lời giải thích cho mọi thứ xảy ra. Hầu hết là việc họ không hiểu cách hoạt động của máy tính là điều hoàn toàn bình thường. 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. Chúng 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 . Lời giải thích của bác sĩ thú y đã hòa giải mâu thuẫn của chủ nhân con chó: "Con chó đã bị căng thẳng và cắn." Quyết định của máy càng ảnh hưởng đến cuộc sống của con người, thì máy càng quan trọng hơn trong việc giải thích hành vi của nó. 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 đầy đủ tình hình, nhưng nên có một 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, bạn nên đề xuất găng tay nếu tôi đã có mũ mùa đông trong giỏ hàng của mình. Thuật toán đề xuất bộ phim này, vì những người dùng đã thích các bộ phim khác mà tôi thích cũng rất thích bộ phim được đề xuất. Càng ngày, các công ty Internet càng thêm giải thích cho các khuyến nghị 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:



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

Trong nhiều ngành khoa học, có sự thay đổi từ phương pháp định tính sang định lượng (ví dụ: xã hội học, tâm lý học) và cả hướng tới học máy (sinh học, di truyền học). **Mục tiêu của khoa học** là thu được kiến thức, nhưng nhiều vấn đề được giải quyết bằng bộ dữ liệu lớn và mô hình máy học hộp đen. Bản thân mô hình trở thành nguồn kiến thức thay vì dữ liệu. Khả năng diễn giải giúp bạn có thể trích xuất kiến thức bổ sung này mà mô hình thu được.

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

Theo 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 bạn 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 bạn đã đào tạo 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 sử dụng trong quá khứ. Mục tiêu

chính của bạn là chỉ cấp khoản vay cho những người cuối cùng sẽ trả nợ. Sự thiếu toàn vẹn của việc xây dựng vấn đề trong trường hợp này nằm ở chỗ bạn không chỉ muốn giảm thiểu các khoản nợ, mà không phân biệt đối xử dựa trên một số đặc thù nhân khẩu (demographics). Đây là một ràng buộc bổ sung nằm trong quy trình xác lập vấn đề của bạn (cấp các khoản vay với rủi ro thấp và hợp lệ) 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. Trong một thí nghiệm nổi tiếng, Heider và Simmel (1944) đã cho những người tham gia xem video về các hình dạng trong đó một hình tròn mở ra "cánh cửa" để vào một "căn phòng" (đơn giản là hình chữ nhật). Những người tham gia mô tả hành động theo các hình dạng thay vì dùng ngôn ngữ thông thường, gán cả cảm xúc và đặc điểm tính cách cho các hình dạng. Robot là một ví dụ điển hình, như máy hút bụi của tôi, mà tôi đặt tên là "Doge". Nếu Doge gặp khó khăn, tôi nghĩ: "Doge muốn tiếp tục dọn dẹp, nhưng 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 máy hoặc thuật toán giải thích các dự đoán của nó 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 lời giải thích. Máy móc có thể cần phải định hình cảm xúc và niềm tin của chúng ta để tương tác với chúng ta. Máy móc phải "thuyết phục" chúng 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 một 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 cần bình luận. 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 lời 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 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 giải thích rằng có điều gì đó đang cản trở, và điều đó đủ để tôi tin tưởng vào hành vi của nó và hiểu được ý nghĩa chung của sự vụ đó.

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

Các mô hình học máy chỉ có thể được gỡ lỗi **(debugged) và kiểm soát (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, bạn 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, bạn có thể kiểm tra những đặc điểm 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 độ chệch 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): Dễ dàng cho con người để tin tưởng một hệ thống mà giải thích quyết định của nó so với một 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 **không có tác động đáng kể**. Tưởng tượng rằng có 1 người tên Mike đang 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 trên Facebook. Nếu mô hình dự đoán sai hay Mike không thể giải thích đầu ra của mô hình của mình, ta sẽ không gặp phải vấn đề gì nghiêm trọng. Ta hoàn toàn ổn nếu không có tính khả giải thích 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 hiểu quả do độ chệch. 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 đầy đủ để có đủ kinh nghiệm thực tế với mô hình và các vấn đề với mô hình đã được giải quyết theo thời gian. Một số ứng dụng 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 thực sự cần thiết.

Khả diễn giải có thể cho phép người dùng hoặc máy tính **thực sự làm chủ các hệ thống** khi người dùng đánh lừa hệ thống bằng cách khai khác

các lỗ hổng đến từ sự không khớp 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, và những người nộp đơn có mục tiêu nhận được khoản vay ngay cả khi ngân hàng không muốn cho họ một khoản vay. Sự không phù hợp giữa các mục tiêu này tạo ra các động cơ khuyến khích người nộp đơn tham gia vào hệ thống để tăng cơ hội nhận được khoản vay. 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. 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 là 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.

0.2 Phân loại các phương pháp khả giải thích

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

Nội tại hay sau huấn luyện? Tiêu chí này phân loại liệu ta đạt được tính khả giải thích 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 sau huấn luyện. Khả năng diễn giải instrinsic đề cập đến các mô hình học máy có tính khả giải thích nhờ vào tính đơn giản của cấu trúc, ví dụ như cây quyết định ngắn hoặc mô hình tuyến tính thưa. Tính khả giải thích sau huấn luyện đề cập việc áp dụng các phương pháp khả giải thích sau khi huấn luyện mô hình. Ví dụ, sự 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 giải thích nội tại. Ví dụ, pfi có thể dùng để tính cho cây quyết định.

Các phương pháp khả giải thích có thể phân loại dựa trên các kết quả.

- **Thống kê tổng quan đặc trưng (Feature summary statistic)**: Nhiều phương pháp khả giải thích 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, như là mức 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 hoạ và biểu diễn dưới dạng bảng là không phù hợp. Một trường hợp đó là tính phụ thuộc riêng. Phác hoạ phụ thuộc riêng là các đường cong và trung bình dự đoán đầu ra. -

- **Các nội bộ mô hình (model internals)** (các trọng số học được): Tính khả diễn giải của các mô hình khả giải thích mang tính nội tại nằm trong phần này. Các ví dụ là các trọng số của mô hình tuyến tính hoặc các cấu trúc học cây (learned tree structures) (các đặc trưng và mức ngưỡng được dùng để chia phần) của các cây quyết định. Các đường bị mờ trong nội bộ mô hình và thống kê tổng quan đặc trưng, 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 khả giải thích mà cho ra nội bộ mô hình sẽ theo định nghĩa cho mô hình cụ thể (xem tiêu chí tiếp theo). -

- **Điểm dữ liệu (data points)**: Mục này gồm tất cả các phương thức để tạo ra các điểm dữ liệu (đã tồn tại hoặc mới được tạo) để làm cho mô hình khả giải thích. Một phương pháp được gọi là giải thích phản chứng (counterfactual explanations). Để giải thích dự đoán của 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: lật (flip) trong lớp dự đoán). Một ví dụ khác là việc xác định các nguyên mẫu của các lớp dự đoán. Các phương pháp khả giải thích mà cho ra điểm dữ liệu mới phải có các điểm dữ liệu đó tính khả giải thích. Điều này hoạt động tốt cho hình ảnh và văn bản, nhưng ít hữu ích hơn cho dữ liệu dạng bảng với hàng trăm đặc trưng. -

- **Mô hình có thể diễn giải nội tại (Intrinsically interpretable model)**: Ta giải thích các mô hình hộp đen qua xấp xỉ chúng (toàn cục hoặc cục bộ) bằng một mô hình khả giải thích. Mô hình khả giải thích được diễn giải bằng cách xem các tham số nội bộ mô hình hoặc thống kê tổng quan tính năng.

Mô hình cụ thể (specific) hay mô hình agnostic? Các công cụ khả giải thích mô hình cụ thể được giới hạn cho một số phân loại mô hình. Việc giải thích các trọng số hồi quy trong một mô hình tuyến tính là khả giải thích theo mô hình cụ thể, vì - theo định nghĩa - việc giải thích các mô hình khả giải thích được về bản chất là mô hình cụ thể. Các công cụ hiệu quả cho giải thích của, ví dụ, mạng nơ-ron là mô hình cụ thể. Các công cụ mô hình 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 (sau 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 thông tin về cấu trúc.

Địa phương hay toàn cầu? Phương pháp khả giải thích diễn giải một

dự đoán riêng lẻ hay toàn bộ hành vi của mô hình? Hoặc là phạm vi ở giữa đâu đó? Thông tin thêm về tiêu chí phạm vi ở phần tiếp theo.

0.3 Phạm vi của khả giải thích

Một thuật toán huấn luyện một mô hình tạo ra các dự đoán. Mỗi bước có thể được đánh giá về tính minh bạch hoặc tính khả giải thích.

0.3.1 Tính minh bạch của thuật toán

Làm thế nào để thuật toán tạo ra mô hình?

Tính minh bạch của thuật toán là về cách thuật toán học một mô hình từ dữ liệu và loại mối quan hệ nào nó có thể học. Nếu bạn sử dụng các mạng thần kinh tích chập để phân loại hình ảnh, bạn có thể giải thích rằng thuật toán học các bộ phát hiện cạnh và bộ lọc trên các lớp thấp nhất. Đây là một sự hiểu biết về cách thức hoạt động của thuật toán, nhưng không phải cho mô hình cụ thể được học cuối cùng, và không phải về cách các dự đoán riêng lẻ được thực hiện. Tính minh bạch của thuật toán chỉ yêu cầu kiến thức về thuật toán chứ không phải về dữ liệu hoặc mô hình đã học. Cuốn sách này tập trung vào khả năng diễn giải mô hình và không minh bạch thuật toán. Các thuật toán như phương pháp bình phương tối thiểu cho các mô hình tuyến tính được nghiên cứu và hiểu rõ. Chúng được đặc trưng bởi tính minh bạch cao. Phương pháp học sâu (đẩy một gradient qua mạng với hàng triệu trọng số) thì ít được hiểu rõ và các hoạt động bên trong là trọng tâm của nghiên cứu đang diễn ra. Chúng được coi là ít minh bạch.

0.3.2 Tính khả giải thích của mô hình toàn cục, toàn diện

Làm thể nào để mô hình đã huấn luyện đưa ra các dự đoán?

Bạn có thể mô tả một mô hình theo cách hiểu của bạn nếu bạn có thể hiểu toàn bộ mô hình (Lipton 2016). Để giải thích đầu ra một mô hình toàn cục, bạn cần huấn luyện mô hình, hiểu thuật toán và dữ liệu. Mức độ này của tính khả giải thích là về cách hiểu mô hình đưa ra các quyết định, dựa trên cái nhìn toàn diện về các đặc trưng quan trọng của nó và từng thành phần đã học như trọng số, các tham số khác ,các cấu trúc và loại tương tác gì đang diễn ra giữa chúng. Tính khả giải thích của mô hình toàn cục là rất khó để đạt được trong thực tế. Bất kì mô hình nào vượt quá một số tham số

hoặc trọng số dường như không phù hợp với bộ nhớ của người trung bình. Tôi tin rằng bạn 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. Thông thường, khi mọi người cố gắng hiểu một mô hình, họ chỉ xem xét các phần của nó, chẳng hạn như các trọng số trong các mô hình tuyến tính.

0.3.3 Tính khả giải thích của mô hình toàn cục ở cấp độ modun

Làm thế nào để các phần của mô hình ảnh hưởng đến các dự đoán?

Một mô hình Naive Bayes với hàng trăm đặc trưng sẽ là quá tải đối với bộ nhớ của tôi và bạn. Và ngay cả khi chúng ta cố gắng nhớ tất cả các trọng số, chúng ta cũng sẽ không thể nhanh chóng đưa ra các dự đoán đối với các điểm dữ liệu mới. Ngoài ra bạn cần có sự phân phối chung của tất cả các đặc trưng trong đầu để ước tính tầm quan trọng của mỗi đặc trưng và cách các đặc trưng ảnh hưởng trung bình đến các dự đoán. Đây quả là một nhiệm vụ bất khả thi. Nhưng bạn có thể dễ dàng hiểu một trọng số đơn (single weight). Trong khi tính khả giải thích của mô hình toàn cục thường nằm ngoài tầm, thì việc hiểu ít nhất một vài mô hình ở cấp độ mô đun là một cơ hội tốt. Không phải tất cả các mô hình đều có thể giải thích được ở mức độ tham số. Đối vớ i các mô hình tuyến tính, những phần có thể giải thích như trọng số, cây được phân tách (đặc trưng được chọn cộng với các điểm cắt) và các nút lá dự đoán. Các mô hình tuyến tính thoạt nhìn chúng ta có thể giải thích hoàn hảo ở cấp độ mô dun nhưng việc giải thích một trọng số đơn 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 phòng. Điều đó có thể xảy ra bởi vì đã có đặc trưng kích thước ngôi nhà có giá trị tương quan cao. Trong thị trường mà mọi người thích những căn phòng lớn hơn, một ngôi nhà có ít phòng hơn có thể có giá trị hơn một ngôi nhà có nhiều phòng hơn nếu cả hai ngôi nhà có cùng kích thước. Những trọng số chỉ có ý

nghĩa trong ngữ cảnh của các đặc trưng khác trong mô hình. Nhưng những trọng số trong một mô hình tuyến tính vẫn có thể được giải thích tốt hơn những trọng số của mạng nơron sâu.

0.3.4 Tính khả giải thích cục bộ cho một dự đoán đơn

Tại sao mô hình đưa ra một dự đoán nhất định cho một trường hợp ví dụ (instance)?

Bạn có thể xem xét kĩ hơn với một trường hợp ví dụ và kiểm tra mô hình đưa ra dự đoán gì với dữ liệu đầu vào có sẵn và giải thích tại sao. Nếu bạn nhìn vào một dự đoán riêng lẻ, hành vi của mô hình phức tạp khác có thể thực hiện dễ dàng. Một cách cục bộ, sự dự đoán có thể chỉ phụ thuộc tuyến tính hoặc đơn điệu trên một số đặc trưng, thay vì phụ thuộc phức tạp vào chúng. Ví dụ, giá trị của một ngôi nhà có thể phụ thuộc phi tuyến vào kích thước của nó. Nhưng nếu bạn chỉ nhìn vào một ngôi nhà 100 mét vuông cụ thể, có khả năng đối với tập dữ liệu đó, sự dự đoán mô hình của bạn phụ thuộc tuyến tính vào kích thước. Bạn có thể tìm ra điều này bằng cách mô phỏng cách giá tị dự đoán thay đổi khi bạn tăng hoặc giảm 10 mét vuông kích thước. Sự giải thích cục bộ có thể chính xác hơn sự giải thích toàn cục. Cuốn sách này trình bày các phương pháp có thể làm cho các dự đoán riêng lẻ trở nên khả giải thích hơn trong phần về các phương pháp không theo mô hình được trình bày trong chương 5.

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

Tại sao mô hình đưa ra các dự đoán cụ thể cho một nhóm các trường hợp (instances)

Các dự đoán mô hình cho nhiều trường hợp có thể được giải thích bằng các phương pháp bằng các phương pháp giải thích mô hình toàn cục (ở cấp độ mô đun) hoặc bằng các giải thích về các trường hợp riêng lẻ. Các phương thức 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 thức 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 được liệt kê hoặc tổng hợp cho toàn bộ

nhóm.

0.4 Khả năng đánh giá tính khả giải thích

Trong học máy, chưa có sự thống nhất thực sự định nghĩa khả giải thích là cái gì. Cũng không có 1 sự định lượng rõ ràng. Nhưng có 1 số nghiên cứu ban đầu và cố gắng hình thành vài hướng tiếp cận để đánh giá, 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 giải thích:

Đánh giá ở mức độ ứng dụng (nhiệm vụ chính): Dưa sự giải thích vào sản phẩm và để người dùng kiểm tra, đánh giá sự 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 gãy 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 mô hình máy học của phần mềm đó. Điều này đòi hỏi người có kinh nghiệm và hiểu cách đánh giá chất lượng. Một nền tảng kiến thức cơ bản về vấn đề này thì luôn luôn tốt cho con người có thể giải thích những quyết định tương tự với máy tính.

Đánh giá ở mức độ con người (Nhiệm vụ đơn giản) là 1 mức độ đánh giá ứng dụng đơn giản. Cái khác biệt là những thử nghiệm đánh giá không bao gồm kiếm thức chuyên môn. Điều này làm chi phí đánh giá rẻ hơn (đặc biệt là kiến thức chuyên môn về X-quang) và dễ dàng để tìm người kiểm thử hơn. Một ví dụ là cho 1 người xem những giải thích khác nhau và để người đó lựa chọn sự giải thích tốt nhất.

Đánh giá ở mức độ hàm chức năng (nhiệm vụ ủy quyền) không yêu cầu con người. Công việc này tốt nhất khi lớp mô hình sử dụng đã được đánh giá bởi 1 người ở mức độ đánh giá con người. Ví dụ, có thể cho rằng người sử dụng hiểu về cây quyết định. Cây ngắn hơn sẽ có được một số điểm giải thích tốt hơn. Sẽ có ý nghĩa khi thêm các ràng buộc rằng hiệu suất dự đoán của cây vẫn tốt và không giảm quá nhiều so với cây lớn hơn.

Chương tiếp theo tập trung vào đánh giá các giải thích cho các dự đoán riêng lẻ ở cấp độ chức năng. Các thuộc tính có liên quan của giải thích mà chúng tôi sẽ xem xét để đánh giá của họ là gì?

0.5 Tính chất của sự diễn giải

Chúng tôi muốn giải thích các dự đoán của một mô hình học máy. Để đạt được điều này, chúng tôi dựa vào một số phương pháp giải thích, đó là một thuật toán tạo ra các giải thích. **Một lời giải thích thường liên quan đến các giá trị tính năng của một thể hiện với dự đoán mô hình của nó theo cách dễ hiểu của con người.** Các loại giải thích khác bao gồm một tập hợp các trường hợp dữ liệu (ví dụ: trong trường hợp của mô hình láng giềng k gần nhất). Ví dụ: chúng ta có thể dự đoán nguy cơ ung thư bằng máy vectơ hỗ trợ và giải thích dự đoán bằng phương pháp thay thế cục bộ, tạo ra các cây quyết định như là lời giải thích. Hoặc chúng ta có thể sử dụng mô hình hồi quy tuyến tính thay vì máy vectơ hỗ trợ. Mô hình hồi quy tuyến tính đã được trang bị một phương pháp giải thích (giải thích các trọng số).

Chúng tôi xem xét kỹ hơn các thuộc tính của phương pháp giải thích và giải thích (Robnik-Sikonja và Bohanec, 20188). Các tính chất này có thể được sử dụng để đánh giá phương pháp giải thích hoặc giải thích tốt như thế nào. Không rõ ràng cho tất cả các tính chất này làm thế nào để đo lường chúng một cách chính xác, vì vậy một trong những thách thức là chính thức hóa cách chúng có thể được tính toán.

Thuộc tính của phương pháp giải thích

- **Sức mạnh biểu cảm** là "ngôn ngữ" hoặc cấu trúc của các giải thích mà phương thức 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 thứ gì khác.

- **Độ trong mờ** mô tả phương pháp giải thích dựa vào mô hình học máy như thế nào, giống như các tham số của nó. Ví dụ, các phương pháp giải thích dựa trên các mô hình có thể hiểu được nội tại như mô hình hồi quy tuyến tính (đặc trưng cho mô hình) rất mờ. Các phương pháp chỉ dựa vào thao tác đầu vào và quan sát các dự đoán có độ trong mờ bằng không. Tùy thuộc vào kịch bản, mức độ mờ khác nhau có thể được mong muốn. Ưu điểm của độ trong suốt cao là phương pháp có thể dựa vào nhiều thông tin hơn để tạo ra các giải thích. Ưu điểm của độ trong mờ thấp là phương pháp giải thích dễ mang theo hơn.

- **Tính di động** mô tả phạm vi của các mô hình học máy mà phương pháp giải thích có thể được sử dụng. Các phương thức có độ trong suốt thấp có tính di động cao hơn vì chúng coi mô hình học máy như một hộp đen. Các mô hình thay thế có thể là phương pháp giải thích với tính di động cao nhất. Các phương thức chỉ hoạt động cho ví dụ: mạng thần kinh tái phát có tính di động thấp.

- **Độ phức tạp thuật toán** mô tả độ phức tạp tính toán của phương pháp tạo ra lời giải thích. Tính chất này rất quan trọng để xem xét khi thời gian tính toán là một nút cổ chai trong việc tạo ra các giải thích.

Thuộc tính của giải thích cá nhân

- **Độ chính xác:** Làm thế nào để một lời giải thích dự đoán dữ liệu chưa nhìn thấy? Độ chính xác cao đặc biệt quan trọng nếu giải thích được sử dụng cho các dự đoán thay cho mô hình học máy. Độ chính xác thấp có thể ổn nếu độ chính xác của mô hình học máy cũng thấp và nếu mục tiêu là giải thích mô hình hộp đen làm gì. Trong trường hợp này, chỉ có lòng trung thành là quan trọng.

- **Độ trung thực:** Giải thích gần đúng với dự đoán của mô hình hộp đen như thế nào? Độ trung thực cao là một trong những tính chất quan trọng nhất của một lời giải thích, bởi vì một lời giải thích với độ trung thực thấp là vô ích để 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 lời giải thích có độ trung thực cao, thì lời 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 thay thế cục bộ) hoặc thậm chí chỉ cho một trường hợp dữ liệu riêng lẻ (ví dụ: Giá trị Shapley).

- **Tính nhất quán:** Một lời giải thích khác nhau bao nhiêu giữa các mô hình đã được đào tạo về cùng một nhiệm vụ và điều đó tạo ra các dự đoán tương tự nhau? Ví dụ, tôi 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. Tôi tính toán các giải thích bằng một phương pháp mà tôi chọn và phân tích sự khác biệt của các giải thích. Nếu các giải thích rất giống nhau, các giải thích rất nhất quán. Tôi thấy

tính chất này hơi khó, vì hai mô hình có thể sử dụng các tính năng khác nhau, nhưng có được các dự đoán tương tự (còn được gọi là "Hiệu ứng Rashomon"). Trong trường hợp này, tính nhất quán cao là không mong muốn vì các giải thích phải rất khác nhau. Tính nhất quán cao là mong muốn nếu các mô hình thực sự dựa trên các mối quan hệ tương tự.

- **Tính ổn định:** Giải thích tương tự như thế nào cho các trường hợp tương tự? Trong khi tính nhất quán so sánh các giải thích giữa các mô hình, độ ổn định so sánh các giải thích giữa các trường hợp tương tự cho một mô hình cố định. Độ ổn định cao có nghĩa là các thay đổi nhỏ trong các tính năng của một thể hiện không thay đổi đáng kể lời giải thích (trừ khi các biến thể nhỏ này cũng thay đổi mạnh mẽ dự đoán). Sự thiếu ổn định có thể là kết quả của phương sai giải thích cao. Nói cách khác, phương thức giải thích bị ảnh hưởng mạnh mẽ bởi những thay đổi nhỏ của các giá trị tính năng của thể hiện được giải thích. Sự thiếu ổn định cũng có thể được gây ra bởi các thành phần không xác định của phương pháp giải thích, chẳng hạn như bước lấy mẫu dữ liệu, giống như phương pháp thay thế cục bộ sử dụng. Độ ổn định cao luôn là mong muốn.

- **Tính dễ hiểu:** Con người hiểu những lời giải thích như thế nào? Điều này trông giống như một tài sản nữa trong số nhiều người, nhưng đó là con voi trong phòng. Khó xác định và đo lường, nhưng cực kỳ quan trọng để có được đúng. Nhiều người đồng ý rằng tính dễ hiểu phụ thuộc vào khán giả. Các ý tưởng để đo lường mức độ dễ hiểu bao gồm đo kích thước của lời giải thích (số tính năng có trọng số khác không trong mô hình tuyến tính, số quy tắc quyết định, ...) hoặc kiểm tra xem mọi người có thể dự đoán hành vi của mô hình học máy tốt như thế nào những lời giải thích. Tính dễ hiểu của các tính năng được sử dụng trong giải thích cũng cần được xem xét. Một chuyển đổi phức tạp của các tính năng có thể ít dễ hiểu hơn các tính năng 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 có tuyên bố về độ tin cậy của mô hình rằng dự đoán là chính xác. Nếu mô hình dự đoán xác suất ung thư là 4% cho một bệnh nhân, liệu có chắc chắn bằng xác suất 4% mà một bệnh nhân khác, với các giá trị tính năng khác nhau, nhận được không? Một lời giải thích bao gồm sự chắc chắn

của mô hình là rất hữu ích.

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

- **Tính mới:** Giải thích có phản ánh liệu một trường hợp dữ liệu được giải thích có xuất phát từ một khu vực "mới" khác xa với việc phân phối dữ liệu đào tạo không? Trong những trường hợp như vậy, mô hình có thể không chính xác và lời giải thích có thể là vô ích. 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:** Có bao nhiêu trường hợp giải thích bao gồm? 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 Khả năng diễn giải thân thiện với con người

Chúng ta cùng đào sâu hơn và khám phá cách mà con người chúng ta diễn giải tốt như thế nào và là những ý nghĩa cho khả năng giải thích được của học máy. Những nghiên cứu về con người học có thể giúp chúng ta tìm ra câu trả lời. Miller (2017) đã tiến hành một cuộc khảo sát có quy mô lớn và đăng tải công bố bài báo về sự diễn giải, và chương này xây dựng trên các tổng hợp của anh ấy.

Trong chương này. Tôi muốn thuyết phục bạn về điều sau: Với một sự diễn giải cho một sự kiện. Con người thích những diễn giải ngắn (chỉ một hoặc nguyên nhân) mà nó đối lập với tình huống hiện tại với một tình huống mà nó sẽ không bao giờ xảy đến. Đặc biệt những nguyên nhân bất thường thì thường là những lời giải thích tốt. Các diễn giải là những tương tác xã hội giữa người giải thích với người được giải thích (người nhận lời giải thích) và do đó bối cảnh xã hội có một ảnh hưởng to lớn đến nội dung thực tế của sự diễn giải.

Khi bạn cần những giải thích với tất cả các nhân tố cho một dự đoán hay một hành vi cụ thể. Bạn không cần một lời giải thích phù hợp, mà chỉ cẩn như một sự quy kết nhân quả tuyệt đối. Bạn có thể muốn một sự quy kết nhân quả nếu bạn được yêu cầu chỉ ra tất cả các đặc trưng có ảnh hưởng hoặc bạn phải gỡ lỗi trong mô hình học máy. Trong trường hợp này, bỏ qua những điểm sau. Trong tất cả các trường hợp khác, những người không có kiến thức chuyên môn hay những người có ít thời gian là những người nhận lời diễn giải, phần sau đây sẽ thu hút bạn đọc.

0.6.1 Diễn giải 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 chúng ta chưa được liên lạc với cuộc sống ngoài hành tinh?

Hai câu hỏi đầu tiên có thể được trả lời với một lời 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. Chúng 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 thê 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 (Explaination)" đề cập đến quá trình diễn giải xã hội và nhận thức, nhưng cũng là sản phẩm của những quá trình đó. Kẻ diễn giãi có thể là con người hay một máy móc.

0.6.2 Diễn giải tốt là gì?

Phần này tiếp tục cô đọng bản tóm tắt của Miller về những diễn giải "tốt" và thêm ý nghĩa cụ thể cho việc học máy có thể hiểu được.

Những 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. Chúng 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 thấp hơn mà họ trông đợi. Nếu ứng dụng vay nợ của tôi bị từ chối, tôi không quan tâm đến tất cả các yếu tố mà thường được nói đến hoặc chống lại sự từ chối đó. Tôi chỉ quan tâm đến các yếu tố mà trong đó mà ứng dụng của tôi cần thay đổi để nhận được khoản vay. Tôi muốn biết sự đối lập giữa ứng dụng của tôi và phiên bản ứng dụng mà được chấp thuận. Nhận thấy rằng các vấn đề giải thích tương phản (contrasting explanations) là một phát hiện quan trọng cho học máy khả giải thích. Đối với hầu hết các mô hình khả diễn giải, bạn có thể trích xuất các giải thích mà ngầm tương phản một dự đoán của một thể hiện với dự đoán đó của một thể hiện dữ liệu nhân tạo hoặc là một trung bình các thể hiện. Người bác sĩ có thể thắc mắc: “Tại sao thuốc này lại không hiệu quả trên bệnh nhân của tôi?”. Và họ có thể muốn một giải thích mà có yếu tố tương phản 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 sự phản hồi khác. Các giải thích tương phản là dễ dàng để hiểu hơn các giải thích đầy đủ. Một 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 gene là đã có biểu hiện quá mức, cơ thể bệnh nhân nhanh chóng phá hủy thuốc thành các thành phần hóa học không còn hữu hiệu,... Một giải thích tương phản có đơn giản hơn nhiều: Đối lập với các bệnh nhân có phản hồi với thuốc, nhưng bệnh nhân không phản hồi có sự kết hợp các gene nhất định làm cho thuốc kém hiệu quả đi. Giải thích tốt nhất là một 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 đến học máy khả diễn giải: Con người không muốn một sự giải thích đầy đủ cho một dự đoán, tạo nên các giải thích tương phản là phụ thuộc vào ứng dụng vì nó yêu cầu một điểm liên quan đến 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 ngồi nhà họ sở hữu hoặc có thể một ngồi nhà nào khác trên trang web hoặc có thể với một ngồi nhà trung bình 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.

Những giải thích mang tính được lựa chọn (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. Chúng ta đã từng lựa chọn ra một hoặc hai nguyên do từ một số những nguyên nhân có thể khác nhau rồi xem như là sự giải thích cho nó. Để chứng minh, điều đó 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 tôi 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 bản mà kể những câu chuyện luân chuyển, mâu thuẫn (giải thích) về cái chết của một samurai. Đối với các mô hình học máy, điều

đó thuận lợi nếu một dự đoán tốt có thể tạo nên từ những đặc trưng khác nhau. Tập hợp lại các mô hình mà kết hợp nhiều mô hình với các đặc trưng khác nhau (các giải thích khác nhau) thường biểu diễn tốt vì tính trung bình trên các "câu chuyện" đó làm cho các dự đoán mạnh mẽ và chính xác hơn. Nhưng nó cũng có nghĩa rằng có nhiều hơn một giải thích có chọn lọc mà có một dự đoán nhất định được đưa ra.
Điều này có ý nghĩa gì đến học máy khả diễn giải: Làm cho sự giải thích rất ngắn gọn, chỉ 1 đến 3 lý do, thậm chí nếu ngữ cảnh có phức tạm hơn. Phương pháp LIME (phần 5.7) thực hiện tốt điều này.

Những giải thích mang tính xã hội (Explanations are social). Đấy là một phần trong một cuộc hội thoại hay sự 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 kỹ thuật 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, thứ mà không thể bị quản lý bởi một cơ sở trung tâm nào, cộng hưởng với đó là con người bảo mật tài sản của họ, điều này giải thích tại sao nó có nhu cầu và giả cả đều 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 của bạn và đối tượng mục tiêu. Để cho phần xã hội của mô hình học máy của bạn phụ thuộc hoàn toàn đúng vào mục đích ứng dụng cụ thể của bạn. Tìm các chuyên gia con người học (ví dụ: nhà tâm lý học và nhà xã hội học) để giúp bạn.

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 đến nhưng nguyên nhân "bất thường" được xem như nhưng giải thích tốt. Một ví dụ từ Štrumbelj và Kononenko (2011) là: Giả sử chúng ta có một bộ dữ liệu về kiểm tra tình huống giữa những giáo viên và học sinh. Học sinh tham gia vào một khóa học và vượt qua khóa

học trực tiếp 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 rớt khóa học. Các học sinh có thể có các chuẩn bị cấp bậc 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 hỏi kiểm tra học sinh đó). Chúng ta muốn dự đoán học sinh nào sẽ đỗ môn học và giải thích các dự đoán đó. Cơ hội để vượt qua là 100% nếu giáo viên hỏi thêm câu hỏi nào, hoặc hoặc xác suất của việc đỗ 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 thêm vào (ví dụ cứ 95 trong 100 lần). Một học sinh không học (10% cơ hội để vượt qua phần câu hỏi) là một trong những người không may mắn, nhận thêm câu hỏi mà học sinh đó trả lời sai. Tại sao học sinh đó rớt? tôi 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 cứ 2 trong 100 lần). Đối với một sinh mà đã không học bài cho các câu hỏi, chúng ta sẽ dự đoán một xác suất cao cho việc đỗ môn học vì nhận các câu hỏi là khó có khả năng. 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 là không may mắn và giáo hỏi thêm câu hỏi, học sinh đó không trả lời đúng và rớt môn. Vậy có lý do nào cho việc thất bại đó? tôi sẽ cho rằng giải thích tốt hơn hết đó là "vì giáo viên đó đã kiểm tra học sinh". Vì đã có 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 mà bất thường trong bất cứ hoàn cảnh nào (như là một loại đặc trưng ít có trong tập các tính năng phân loại) và các đặc trưng đó ảnh hướng đến dự đoán, nó nên có bao chứa một giải thích, thậm chí nếu các đặc trưng "bình thường" khác có cùng ảnh hưởng như các 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 một số phương pháp tìm kiếm tiêu biểu cho thấy là sự đóng góp hai ban công làm cho giá nhà cao chênh lệch cũng như kích thước nhà trên trung bình, vùng lân cận xung quanh tốt hay cải tạo gần đây, đặc trưng bất thường "hai ban công" có thể là sự giải thích tại sao giá nhà lại cao vậy.

Các giải thích mang tính trung thực (Explanations are truthful).

Một giải thích tốt chứng minh là đúng trong thực tế (kể cả trong các tình huống khác). Nhưng lại đáng lo ngại, đây không phải là yếu tố quan trọng nhất cho một giải thích tốt. Ví dụ, tính chọn lọc dường như quan trọng hơn tính trung thực. Một 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. Điều đó không đúng hoàn toàn mà chỉ một hoặc hai yếu tố, ví dụ, với nguyên nhân gây ra sự phá sản của một thị trường chứng khoán, nhưng sự thực là có hàng triệu nguyên nhân mà ảnh hưởng đến hàng triệu người đều dẫn đến hành động theo cùng một cách mà kết thúc lại là sự phá sản đấy.
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 thành thật (fidelity). Vậy nên nếu chúng ta nói rằng hai ban công làm tăng giá nhà, thì cũng nên áp dụng đối với các căn nhà khác (hoặc ít nhất các nhà tương đồng). Đối với con người, tính thành thật (fidelity) của một giải thích là không quan trọng như tính lựa chọn, nó đối lập và theo khía cạnh của xã hội.

Giải thích tốt mang tính phù hợp với niềm tin trước đây 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 mà nó không phù hợp với niềm tin có sẵn của họ. Ảnh hưởng này được gọi là sự sai lệch xác nhận (confirmation bias) (Nickerson 1998). Các giải thích không bị phân cách bởi các sai lệch này. Con người sẽ có xu hướng đánh giá thấp hoặc bỏ qua những giải thích không đồng tình với niềm tin của họ. Nhóm các niềm tin khác nhau giữa người này với người khác, nhưng cũng có những niềm tin dựa trên nhóm như là các thế giới quan chính trị.
Điều này có ý nghĩa gì đến học máy khả diễn giải: Các giải thích tốt là phù hợp với các niềm tin trước. Điều này khó có thể áp dụng lên học máy và có lẽ sẽ ảnh hưởng lớn hiệu suất dự đoán. Niềm tin trước của chúng ta về sự ảnh hưởng của kích thước căn nhà lên giá nhà là căn nhà càng lớn, giá càng cao. Giả sử là mô hình cũng tác động tiêu cực của kích thước nhà đối với giá dự đoán cho một ít ngôi nhà. Mô hình đã học được điều này vì nó cải thiện hiệu suất dự đoán (do một số tương tác phức tạp), nhưng hành vi này ảnh mâu thuẫn mạnh mẽ với niềm tin trước của chúng ta. Bạn có thể thực hiện áp dụng lên các ràng buộc đơn điệu (một đặc trưng chỉ có thể ảnh

hướng đến dự đoán theo một chiều) hoặc sử dụng một số mô hình tuyến tính có thuộc tính đó.

Giải thích tốt mang tính tổng quan và hiển nhiên. (Good explanations are general and probable). Một nguyên nhân có thể được giải thích cho nhiều sự kiện là rất khái quát tổng thể và có thể được coi là một lời giải thích tốt. Lưu ý là 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 thắng thế với nguyên nhân khái quát. Nguyên nhân bất thường là xác định hiếm khi trong các kịch bản được cho. Trong trường hợp không có sự tồn tại của sự kiện bất thường, một giải thích khái quát là được xem xét như là một 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 thủ thư, có khả năng anh ấy là một người nhút nhát hay nhút nhat mà thích đọc sách?) 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 đặc trưng hỗ trợ, là số các mẫu thể hiện mà sự giải thích được áp dụng chia cho tổng số thể hiện.