Chương 8: Nhìn về tương lai

Vậy tương lai của học máy khả diễn giải là gì? Chương này gồm những phỏng đoán cá nhân của tác giả về tương lai của học máy khả diễn giải. Tôi đã mở đầu cuốn sách này với những mẩu chuyện nhỏ bi quan và sẽ kết thúc bằng một cái nhìn lạc quan.

Những phỏng đoán của tôi dựa vào ba tiên đề sau:

1. **Sự số hóa: Bất cứ thông tin hữu ích nào đều sẽ được số hóa.** Hãy nghĩ về việc thanh toán trực tuyến và tiền điện tử, về sách điện tử, âm nhạc và video. Nghĩ về tất cả các dữ liệu trong đời sống thường nhật, hành vi của con người, quá trình sản xuất công nghiệp, ... Những tiên lượng về việc số hóa dữ liệu ở khắp nơi: từ các máy tính/cảm biến/bộ nhớ giá rẻ, mở rộng thị trường (winner takes it all), các mô hình kinh doanh mới, các chuỗi giá trị, áp lực giá cả và hơn thế nữa.

2. **Tự động hóa: Khi một bài toán có thể được tự động và giá thành tự động hóa trở nên thấp hơn giá thành để vận hành, vấn đề sẽ được tự động hóa.** Thậm chí ngay trước sự có mặt của máy tính, ta đã có tự động hóa. Ví dụ, các máy dệt tự động hay các máy hơi nước. Nhưng máy tính và chu trình số hóa đưa tự động hóa lên một tầm cao mới. Chỉ đơn giản với việc bạn dùng các vòng lặp for trong lập trình, viết các hàm tính toán với Excel hay trả lời thư tự động cũng đã cho ta thấy tiềm năng vô hạn của tự động hóa. Các máy bán vé tự động phụ trách bán vé tàu (không cần nhân viên thu ngân), các mát giặt tự động là ủi, các máy đặt hàng tự động việc thanh toán ... Việc tự động hóa các công việc tiết kiệm cho ta rất nhiều thời gian và tiền bạc, do đó vấn đề này được đầu tư một lượng tài lực cũng như vật lực lớn lao. Ta đang chứng kiến sự tự động hóa trong biên dịch ngoại ngữ, lái xe, ở mức độ nhỏ, như nghiên cứu khoa học.

3. **Hiểu nhầm: Ta đang không thể chỉ ra mục tiêu rõ ràng với các điều kiện đưa ra.** Nghĩ về thần đèn trong một cái chai và luôn sẵn sàng đáp ứng các ước muốn của bạn. "Tôi muốn trở thành người giàu nhất thế gian" - Bạn thành người giàu nhất thế giới, nhưng có một vấn đề đó là tiền bạn giữ sẽ bị lạm phát. "Tôi muốn trở nên hạnh phúc cho cả phần đời còn lại" - Bạn sẽ hạnh phúc trong 5 phút sau và thần đèn giết bạn. "Tôi ước thế giới hòa bình" - Thần đèn hủy diệt cả nhân loại.

Do ta xác định mục tiêu sai, hoặc bởi vì ta không hề biết đến hoặc đong đếm tất cả các ràng buộc. Một tập đoàn có mục tiêu đơn giản là kiếm tiền cho các cổ đông của nó. Nhưng việc chỉ ra này không phản ánh đúng mục đích thật sự với tất cả các ràng buộc mà ta muốn đảm bảo: Ví dụ, ta không mong muốn công ty sẽ giết người để kiếm tiền, gây ô nhiễm các dòng sông, hoặc in tiền giả. Con người có pháp luật, quy định, chế tài, trình tự phải tuân theo, và công đoàn để hỗ trợ mục tiêu đó. Một ví dụ khác ta có thể tự thấy đó là Paperclips, một trò chơi trong đó nhiệm vụ là bạn phải tạo ra nhiều chiếc kẹp giấy nhất có thể. CẢNH BÁO: game gây nghiện. Tôi không muốn tiết lộ quá nhiều ở đây. Trong học máy, sự không hoàn hảo của việc xác định mục tiêu đến từ việc trừu tượng hóa dữ liệu không hoàn hảo (các mẫu dữ liệu lỗi, lỗi đo đạc, ...), các hàm mất mát không có ràng buộc, thiếu kiến thức về các ràng buộc, sự khác nhau về phần phối dữ liệu trong huấn luyện và thực tế, ...

Sự số hóa đang dẫn dắt tự động hóa. Sự xác định mục tiêu không hoàn hảo xung đột với việc tự động hóa. Tôi cho rằng xung đột này được gây ra một phần bởi các phương pháp diễn giải.

Ta đã xác định được mục tiêu, quả cầu pha lê đã sẵn sàng, hãy cùng nhìn xem ta sẽ đi tới đâu!

0.1 Tương lai của học máy

Nếu không có học máy sẽ không có học máy khả diễn giải. Do đó ta cần phỏng đoán học máy sẽ đi về đâu trước khi ta nói tới tính khả diễn giải.

Học máy luôn đi kèm với rất nhiều hứa hẹn và kỳ vọng. Nhưng hãy bắt đầu với một cái nhìn thiếu lạc quan một chút: Trong khi khoa học phát triển rất nhiều các các công cụ học máy mới ra đời, theo tôi, khá khó để tích hợp chúng vào đời sống và tạo ra các sản phẩm phục vụ con người. Không phải vì ta không thể, chỉ đơn giản vì nó sẽ tốn thời gian cho các công ty và viện nghiên cứu bắt kịp với nhau. Trong cơn lũ của trí tuệ nhân tạo hiện nay, các công ty mở ra các phòng nghiên cứu về trí tuệ nhân tạo, các nhóm nghiên cứu học máy và thuê các nhà khoa học dữ liệu, các chuyên gia học máy, các kỹ sư AI, ... nhưng trên thực tế, việc này không ổn. Các công ty thường thậm chí không có dữ liệu ở định dạng cần thiết và các nhà khoa học dữ liệu phải đợi đến vài tháng. Đôi khi các công ty có những kỳ vọng quá cao về trí tuệ nhân tạo và khoa học dữ liệu được thổi phồng bởi truyền thông và các nhà học học dữ liệu không thể đáp ứng. Và thường không ai biết cách để đưa các nhà khoa học dữ liệu vào các mô hình làm việc có sẵn bên cạnh rất nhiều vấn đề khác. Điều này dẫn tới dự đoán đầu tiên của tôi:

Học máy sẽ phát triển chậm nhưng đều.

Số hóa đang được đẩy mạnh và mong muốn về tự động hóa cũng vậy. Thậm chí nếu việc đưa học máy vào cuộc sống diễn ra chậm chạp đi chăng nữa, học máy vẫn sẽ phát triển và có nhiều ứng dụng vào các doanh nghiệp, sản phẩm, cũng như các mặt của đời sống.

Tôi tin rằng ta cần giải thích rõ ràng hơn cho những người ngoại đạo về tính ứng dụng mạnh mẽ của học máy. Tôi biết rất nhiều nhà khoa học dữ liệu được trả lương rất cao để thực hiện các công việc tính toán trên Excel hoặc các tác vụ doanh nghiệp, và làm việc dựa trên các truy vấn SQL thay vì ứng dụng học máy. Nhưng có một số ít công ty đã sử dụng học máy thành công, đơn cử như các công ty tập đoàn công nghệ lớn. Ta cần tìm các phương pháp ưu việt hơn để đưa học máy vào quy trình sản xuất và làm việc, huấn luyện con người và phát triển các công cụ học máy dễ sử dụng. Tôi tin rằng học máy sẽ trở nên dễ sử dụng hơn nhiều: Ta có thể đang thấy học máy đang ngày càng dễ tiếp cận, ví dụ thông qua các dịch vụ đám mây (cloud services) ("Machine Learning as a service"). Một khi học máy phát triển tới

một ngưỡng nhất định - hiện tại chúng ta qua những bước đầu tiên - dự đoán tiếp theo của tôi là.

Học máy sẽ vận hành rất nhiều công việc.

Dựa trên nguyên lý rằng "những thứ có thể tự động hóa sẽ được tự động hóa", tôi kết luận rằng ngay khi có thể, các công việc sẽ được định hình lại dưới bài toán có dạng dự đoán kết quả và được giải quyết bằng học máy. Học máy là một dạng của tự động hóa hoặc ít nhất là một phần của nó:

- Sắp xếp/Ra quyết định/Hoàn thành tài liệu (ví dụ: các công ty bảo hiểm, các doanh nghiệp pháp lý hoặc tư vấn).

- Đưa ra quyết định dựa trên dữ liệu ví dụ như các bài toán liên quan tới tín dụng.

- Phát triển thuốc.

- Quản lý sản phẩm trong các dây chuyền lắp ráp.

- Xe tự hành.

- Chẩn đoán bệnh tật.

- Dịch ngôn ngữ. Với cuốn sách này, tôi sử dụng một dịch vụ dịch tên là DeepL. Công cụ này sử dụng các mạng nơ ron để cải thiện chất lượng dịch bằng cách dịch từ tiếng anh sang tiếng Đức sau đó dịch ngược lại tiếng Anh.

- Và rất nhiều các ứng dụng khác.

Sự đột phá của học máy không chỉ là thành quả của các máy tính hiện đại, sự dồi dào của dữ liệu, các phần mềm tân tiến, mà còn:

Các công cụ khả diễn giải là chất xúc tác cho quá trình tích hợp học máy.

Dựa trên tiên đề rằng mục tiêu của học máy có thể không bao giờ được định rõ, điều này dẫn đến rằng học máy khả diễn giải sẽ cần thiết để thu hẹp khoảng cách giữa các mục tiêu đúng đắn và các mục tiêu được xác định sai. Trong rất nhiều lĩnh vực, tính khả diễn giải có thể là chất xúc tác cho việc đưa học máy vào ứng dụng. Một vài dẫn chứng như sau: Rất nhiều người tôi đã nói chuyện không sử dụng các mô hình học máy bởi vì chúng không thể giải thích. Tôi tin rằng tính khả diễn giải sẽ giải quyết vấn đề này và giúp học máy trở nên hấp dẫn đối với các doanh nghiệp và cá nhân đòi hỏi sự minh bạch. Bên cạnh các mục tiêu được xác định sai, rất nhiều ngành khác cũng đòi hỏi tính khả diễn giải, và coi nó thuộc về phạm trù pháp lý, do sự lo ngại rủi ro hoặc để nhìn rõ bản chất vấn đề. Học máy tự động quá hóa trình

mô hình và giúp con người đi xa hơn từ dữ liệu và các bài toán: Điều này làm tăng rủi ro với cách thiết lập thí nghiệm, phân phối huấn luyện, cách lấy mẫu, mã hóa dữ liệu, kỹ thuật đặc trưng, ... Các công cụ diễn giải giúp ta chỉ ra vấn đề dễ dàng hơn.

0.2 Tương lai của khả diễn giải

Ta hãy nhìn vào tương lai của học máy khả diễn giải.

Tập trung vào các công cụ giải thích kiểu mẫu.

Sẽ dễ dàng hơn nhiều nếu ta có thể tự động hóa tính khả diễn giải khi ta đặt nó bên ngoài các mô hình học máy. Ưu điểm của việc giải thích kiểu mẫu nằm ở tính mô đun hóa của nó. Ta có thể dễ dàng thay đổi các mô hình học máy. Ta cũng có thể dễ dàng thay đổi phương pháp diễn giải. Với các lý do này, các phương pháp kiểu mẫu sẽ có thể được mở rộng tốt hơn. Đó là tại sao tôi tin rằng các phương pháp kiểu mẫu sẽ trở nên phổ biến trong tương lai. Tuy nhiên các phương pháp khả diễn giải thông thường vẫn sẽ có chỗ đứng riêng của chúng.

Học máy sẽ được tự động hóa cùng với tính khả diễn giải.

Một xu thế ta có thể thấy hiện này đó là việc tự động hóa việc huấn luyện các mô hình. Điều này bao gồm việc các kỹ thuật và việc lựa chọn đặc trưng được tự động hóa, cũng như tự động hóa việc tối ưu các siêu tham số (hyperparameters), so sánh các mô hình khác nhau, gộp hoặc chồng các mô hình. Kết quả là mô hình dự đoán tốt nhất có thể. Khi ta dùng các phương pháp giải thích kiểu mẫu, ta có thể áp dụng chúng một cách tự động vào bất cứ mô hình nào được sinh ra từ quá trình học máy tự động. Một cách khác, ta cũng có thể tự động bước thứ hai này như sau: Tự động tính toán độ quan trọng của các đặc trưng, phác họa sự phụ thuộc riêng, huấn luyện một mô hình thay thế, và vân vân... Chẳng ai có thể ngăn cản bạn tự động hóa việc diễn giải. Việc diễn giải thực tế cần sự hỗ trợ của con người. Tưởng tượng rằng: Bạn tải lên một tập dữ liệu, chỉ định mục tiêu cho việc dự đoán và nhấn nút để huấn luyện ra mô hình tốt nhất, và cuối cùng chương trình gửi về tất cả các diễn giải về mô hình. Đã có một vài sản phẩm đầu tiên và tôi cho rằng với rất nhiều ứng dụng, việc tự động hóa các dịch vụ liên quan tới học máy là khả dĩ. Ngày nay bất cứ ai đều có thể xây dựng các trang web mà không cần biết HTML, CSS và Javascript, nhưng vẫn có rất nhiều nhà

phát triển web. Tương tự như vậy, tôi tin rằng tất cả mọi người sẽ đều có thể huấn luyện các mô hình học máy mà không cần kiến thức về lập trình, và tất nhiên vẫn sẽ có chỗ cho các chuyên gia trong lĩnh vực này.

Ta không phân tích dữ liệu, mà ta phân tích các mô hình. Dữ liệu thô bản thân chúng hầu như không có tác dụng. (Có thể tôi nói quá. Sự thực là bạn cần hiểu biết kỹ lưỡng về dữ liệu nếu muốn thực hiện phân tích chúng). Tôi không quan tâm về dữ liệu. Tôi quan tâm về những kiến thức hàm chứa trong đống dữ liệu đó. Học máy khả diễn giải là một cách tiếp cận tuyệt vời để có thể chắt lọc thông tin quý báu từ dữ liệu. Ta có thể kiểm tra mô hình một cách kỹ lưỡng, mô hình tự động nhận dạng liệu và bằng cách nào đặc trưng liên quan đến một dự đoán (rất nhiều mô hình có sẵn khả năng lựa chọn đặc trưng), mô hình tự động phát hiện cách các quan hệ được thể hiện, và – nếu được huấn luyện đúng – mô hình cuối cùng sẽ xấp xỉ rất tốt thực tế.

Rất nhiều công cụ phân tích đang được sử dụng dựa trên các mô hình dữ liệu (bởi vì chúng được dựa trên các giả thiết về phân phối).

- Các phép thử giả thiết đơn giản như Student's t-test.

- Các phép thử giả thiết với điều chỉnh nhân tố gây nhiễu (thường với GLMs).

- Các phân tích về phương sai (ANOVA).

- Các hệ số tương quan (hệ số hồi quy tuyến tính chuẩn hóa liên quan tới hệ số tương quan Pearson).

- vân vân ...

Những gì tôi đang nói ở đây không có gì mới. Vậy tại sao ta lại chuyển từ phân tích các mô hình dựa trên các giả thiết và minh bạch sang các mô hình hộp đen. Bởi vì việc tạo ra tất cả các giả thiết đó có vấn đề: Chúng thường không chính xác (dù cho bạn tin rằng hầu hết thế giới tuân theo phân phối Gaussian), khó kiểm tra, rất không mềm dẻo, và khó để tự động hóa. Trong rất nhiều lĩnh vực, các mô hình dựa trên các giả thiết thường có hiệu năng kém nhất trên tập dữ liệu mới so với các mô hình hộp đen. Điều này chỉ đúng với các tập dữ liệu lớn, do các mô hình khả diễn giải với những giả thiết tốt thường hoạt động tốt hơn với các tập dữ liệu nhỏ so với các mô hình hộp đen. Cách tiếp cận theo mô hình hộp đen yêu cầu rất nhiều dữ liệu để làm việc tốt. Với việc số hóa mọi thứ, ta sẽ có lượng dữ liệu lớn hơn và do đó cách tiếp cận này sẽ trở nên phù hợp hơn. Ta không cần tạo ra các giả thiết, ta xấp

xỉ thực tế gần nhất có thể (trong khi tránh quá khớp trên tập huấn luyện). Tôi cho rằng ta nên phát triển tất cả các công cụ mà ta có trong thống kê để trả lời các câu hỏi (kiểm tra giả thiết, đo đạc tương quan, đo đạc tương tác, công cụ trực quan hóa, khoảng tin cậy, giá trị p, khoảng dự đoán, phân bố xác suất) và viết lại chúng cho các mô hình hộp đen. Nói một cách nào đó, điều này đang xảy ra:

- Ta hãy lấy một mô hình tuyến tính: Hệ số hồi quy chuẩn hóa là độ quan trọng đặc trưng. Với việc tính toán độ quan trọng đặc trưng hoán vị (permutation feature importance measure), ta có một công cụ làm việc trên mọi mô hình.

- Trong một mô hình tuyến tính, các hệ số tính toán ảnh hưởng của một đặc trưng tới đầu ra được dự đoán. Tổng quát hóa của việc này là phác họa đồ thị riêng (partial dependence plot).

- Kiểm tra liệu A hay B tốt hơn: Với việc này ta cũng sử dụng các hàm phụ thuộc riêng. Thứ ta chưa có (theo quan điểm của tôi) đó là các bài kiểm tra thống kê cho các mô hình hộp đen bất kỳ.

Các nhà khoa học dữ liệu sẽ tự động hóa họ. Tôi tin rằng các nhà khoa học dữ liệu sẽ tự động hóa bản thân họ khỏi những công việc liên quan tới phân tích và dự đoán. Để điều này xảy ra, các bài toán phải được định nghĩa rõ ràng và phải có các quy trình chuẩn bên cạnh chúng. Ngày này, các thói quen và quy trình đang bị khuyết, nhưng các nhà khoa học dữ liệu và cộng sự đang làm việc với chúng. Khi học máy trở thành một phần của các lĩnh vực khác nhau trong cuộc sống, rất nhiều vấn đề sẽ được tự động.

Robot và các chương trình sẽ giải thích chúng. Ta cần nhiều hơn các giao diện với máy móc và các chương trình mà sử dụng học máy. Ví dụ: Một chiếc xe tự hành báo cáo tại sao nó lại dừng đột ngột ("70% là một cậu bé đang qua đường"). Một chương trình chấm điểm tín dụng phải giải thích cho nhân viên ngân hàng tại sao một đơn vay lại bị từ chối ("Ứng viên có quá nhiều thẻ tín dụng và công việc không ổn định"). Một cánh tay robot giải thích tại sao nó di chuyển một sản phẩm từ băng chuyền sang thùng rác ("Sản phẩm có lỗi ở đáy").

Tính khả diễn giải có thể thúc đẩy việc nghiên cứu trí thông minh của máy móc.

Tôi có thể tưởng tượng được khi ta có nhiều nghiên cứu hơn về tại sao các chương trình và máy móc có thể giải thích bản thân chúng, ta có thể cải

thiện hiểu biết của ta về trí thông minh và giúp tạo nên các máy móc thông minh tốt hơn.

Cuối cùng, tất cả các dự đoán đều dựa trên sự quan sát và ta phải xem tương lại sẽ mang đến điều gì. Thử lập luận và tiếp tục học tập nào!

Chương 9: Đóng góp cho cuốn sách này

Cám ơn vì đã đọc cuốn sách về học máy khả diễn giải của tôi. Cuốn sách đang được phát triển liên tục. Nó sẽ được cải thiện theo thời gian và nhiều chương sẽ được thêm vào. Rất giống với cách phần mềm được phát triển

Tất cả các văn bản và mã nguồn cho cuốn sách này là mã nguồn mở và có thể tìm thấy ở github. Trên trang Github bạn cũng có thể đề xuất các sửa đổi và tạo các vấn đề nếu bạn tìm thấy lỗi hoặc muốn thêm phần nào đó.

Chương 10: Trích dẫn

Nếu bạn thấy cuốn sách này hữu ích cho học tập, nghiên cứu, hoặc công việc, sẽ rất tuyệt nếu bạn có thể trích dẫn cuốn sách này. Bạn có thể trích dẫn như sau:

```
Molnar, Christoph. ‘‘Interpretable machine learning.
A Guide for Making Black Box Models Explainable’’, 2019.
https://christophm.github.io/interpretable-ml-book/.
```

Hoặc dùng bibtex:

```
@book{molnar2019,
  title      = {Interpretable Machine Learning},
  author     = {Christoph Molnar},
  note       = {\url{https://christophm.github.io/
                interpretable-ml-book/}},
  year       = {2019},
  subtitle   = {A Guide for Making Black Box Models Explainable}
}
```

Tôi luôn tò mò về việc liệu các phương pháp diễn giải được dùng ra sao trong nghiên cứu cũng như trong công nghiệp. Nếu bạn dùng cuốn sách này như một nguồn tham khảo, sẽ rất tuyệt nếu bạn có thể viết vài dòng cho tôi về việc bạn dùng nó như thế nào. Tất nhiên việc này sẽ tùy vào cá nhân bạn và chỉ để thỏa mãn tính hiếu kỳ của tôi cũng như để thúc đẩy việc trao đổi thông tin giữa chúng ta. Thư điện tử của tôi là: christoph.molnar.ai@gmail.com .

Chương 11: Các bản dịch

Bạn muốn dịch cuốn sách này? Cuốn sách này được cấp phép theo Giấy phép Quốc tế Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Có nghĩa là ban được phép dịch và đưa nó lên mạng Internet. Bạn có nghĩa vụ nhắc đến tôi như là tác giả gốc và không được phép bán cuốn sách này.

Nếu bạn muốn dịch cuốn sách này, bạn có thể nhắn cho tôi một tin và tôi có thể liên kết bản dịch của bạn ở đây. Địa chỉ email của tôi là: christoph.molnar.ai@gmail.com .

Danh sách các bản dịch:

Tiếng Việt: - Link | Dịch giả chính Giang Nguyen và Duy-Tung Nguyen.

Tiếng Trung: - Link | Dịch hoàn chỉnh bởi Mingchao Zhu

- Link | Được dịch bởi hầu hết các chương bởi CSDN, cộng đồng lập trình viên online tại trung quốc.

- Link | Đã dịch một số chương. Trang này cũng bao gồm nhiều câu hỏi và trả lời về các vấn đề khác nhau.

Tiếng Hàn:

- Link | Dịch hoàn chỉnh bởi TooTouch

- Link | Dịch một nửa bởi An Subin

Tây Ban Nha:

- Link | Các chương đầu được dịch bởi Federico Fliguer

Nếu bạn biết bất cứ bản dịch nào khác, tôi sẽ rất vui nếu bạn cho tôi biết để tôi có thể liệt kê ra đây. Bạn có thể gửi thử cho tôi qua: christoph.molnar.ai@gmail.com .