5 quan điểm cơ bản về khoa học dữ liệu

Tóm tắt :

Khoa học dữ liệu là một lĩnh vực liên ngành về các quá trình và các hệ thống rút trích tri thức hoặc hiểu biết từ dữ liệu ở các dạng khác nhau, kể ở dạng cấu trúc hay phi cấu trúc. Trong bài viết này tác giả muốn nhấn mạnh các quan điểm mà các nhà khoa học dữ liệu cần phải tuân thủ trong quá trình nghiên cứu để tránh những sai sót do thiên vị, không nắm rõ dữ liệu phân tích, đơn giản hóa hay quá phức tạp trong xây dựng mô hình phân tích và cuối cùng là tôn trọng những gì dữ liệu vốn có.

Khoa học dữ liệu giống như…

Đã bao nhiêu lần bạn nhìn thấy một bài báo bắt đầu theo cách này? Bây giờ, bạn đã thấy! Nhưng tôi muốn tránh hoàn thành câu đó, và do đó tôi sẽ bỏ qua các định nghĩa và suy luận ở đây, mặc dù chúng là rất quan trọng và các đồng nghiệp của tôi (và học sinh) biết rằng tôi rất chú tâm khi cố gắng để giải thích bất cứ điều gì. Thay vào đó, tập trung ở đây sẽ là những quan điểm cơ bản quan trọng đối với tất cả sinh viên và học viên của khoa học dữ liệu.

Hãy để tôi bắt đầu với một câu chuyện: Một người đàn ông giàu có muốn đầu tư vào việc nuôi và đào tạo một lớp học về đua ngựa với mục tiêu là sẽ giành chiến thắng càng nhiều cuộc đua càng tốt. Ông đã quyết định tài trợ cho các nghiên cứu của ba nhà khoa học đẳng cấp thế giới để thực hiện ước mơ của mình: Một nhà sinh lý học, một nhà hóa sinh, và một nhà vật lý. Sau thời gian thích hợp dành cho nghiên cứu và phát triển, nhà sinh lý học đã được gọi đến để giải thích giải pháp của mình. Cô trình bày một chế độ tập luyện toàn diện hàng ngày sẽ đảm bảo ngựa có sức mạnh, nhanh và sức chịu đựng cao, nếu những con ngựa được đào tạo từ nhỏ theo kế hoạch của mình. Người đàn ông giàu chúc mừng, cảm ơn và thanh toán các giải pháp cô đưa ra. Tiếp theo nhà hóa sinh học được gọi đến để giải thích các giải pháp của ông đưa ra. Nhà hóa sinh đã trình bày kế hoạch chế độ ăn uống đầy đủ những con ngựa nên theo từ thời thơ ấu đến tuổi trưởng thành, kể cả trước khi cuộc đua, trong ngày đua và các bữa ăn sau cuộc đua và Ông đảm bảo những nguyên tắc chế độ ăn uống của ngựa sẽ tạo ra sức mạnh, nhanh và có sức chịu đựng cao. Người đàn ông giàu có đã chúc mừng, cảm ơn và chi trả các giải pháp cho nhà hóa sinh. Cuối cùng, nhà vật lý đã được gọi đến để giải thích giải pháp của mình. Người đàn ông giàu có đang mong chờ giải pháp của nhà vật lý, vì các nhà vật lý sinh ra là để giải quyết vấn đề và chắc chắn sẽ có một giải pháp tuyệt vời. Nhà vật lý bắt đầu "giả định một con ngựa hình cầu....".

Hình ảnh minh họa

Điều gì đã sai? Nhà khoa học thứ ba cho rằng một mô hình quá đơn giản sẽ đúng cách đặt một con ngựa vào các thể loại cụ thể của "ngựa đua nhanh", trong khi hai nhà khoa học trước hiểu rằng đây là một vấn đề đa chiều (đa biến), cũng

như có thể được chứng minh bằng cách dành một khoảng thời gian chất lượng với một cuốn sách về phân tích nhân tố.

Mặc dù các giải pháp khác nhau, cả ba nhà khoa học đã bắt đầu một cách khôn ngoan bằng cách là làm theo các nguyên tắc. Do vậy, quan điểm cơ bản thứ nhất về khoa học dữ liệu, đó là: Bắt đầu với kết thúc trong tâm trí! Quan điểm này là nền tảng cho khoa học, kỹ thuật, thiết kế, kinh doanh, giáo dục, y tế, an ninh, kế hoạch tài chính, thể thao và có lẽ là tất cả các lĩnh vực và hoạt động của con người. Tương tự như vậy, bất cứ khi nào chúng ta tiến hành phân tích dữ liệu lớn (dữ liệu khoa học) nhiệm vụ, dự án, chúng ta nên hỏi: Mục tiêu là gì? Chúng ta đang cố gắng đạt được những gì? Làm thế nào để chúng ta đạt được mục tiêu? Nếu có thể, chúng ta nên xác định số lượng các mục tiêu cuối cùng với số liệu - kết quả đo lường được, với một số ước tính về "ngưỡng thành công". Hơn nữa, kiến thức về mục tiêu cuối cùng của chúng ta sẽ thường xuyên có các kích thích quan trọng cho việc lựa chọn các thành phần phù hợp cho dự án: Tuyển dụng một nhóm làm việc, lựa chọn các bộ dữ liệu chính xác, chọn các tính năng từ các dữ liệu cần được phân tích và xác định những thuật toán nào cần được sử dụng. Thông thường, khai thác dữ liệu được mô tả một cách tối tệ và đúng như vậy, khi các học viên sử dụng nó như là một "cuộc câu cá" để xem điều gì xẩy ra. Trong khi một số dữ liệu thăm dò không có giám sát là điều cần thiết (để đảm bảo rằng chúng ta không bỏ lỡ "những dữ liệu đang nói với chúng ta" và tìm ra tất cả các mô hình, xu hướng, tương quan, và các cụm trong bộ dữ liệu), tuy nhiên chúng ta nên đặt yếu tố rõ ràng lên đầu nếu đó là những gì chúng ta đang hướng tới để đạt được. Mặt khác, mục đích cuối cùng của chúng ta (đặc biệt là trong các lĩnh vực nêu trên) thường rõ ràng hơn nhiều: bán hàng hoá nhiều hơn, giữ khách hàng hài lòng, khám phá liệu pháp chữa bệnh cho một số bệnh, thiết kế một sản phẩm chức năng mạnh mẽ, khám phá các đặc tính một số hiện tượng khoa học mới, giành Cup Châu Mỹ. Hoặc tìm hiểu cách lai tạo và huấn luyện những con ngựa chiến thắng.

Tiềm ẩn trong các tuyên bố trên là quan điểm cơ bản thứ hai của khoa học dữ liệu, đó là: Hiểu biết dữ liệu của bạn! Để hiểu biết được các dữ liệu sẽ là tốt nhất cho một dự án, và các tính năng nào cần chọn, chúng ta phải biết rõ ràng dữ liệu của chúng ta. Nhưng tôi muốn nói đến một cái gì đó nhiều hơn thế, tốt hơn nên gọi đó là "Tập Dữ liệu". Trong quá trình xử lý dữ liệu, chúng tôi kiểm tra nhiều khía cạnh của dữ liệu: Giá trị min/max, tổng hợp các giá trị, như: Trung bình, trung vị, tổng,... danh sách các giá trị dữ liệu riêng biệt (nếu chúng ta làm việc với các thuộc tính dữ liệu rời rạc được xác định), các biểu đồ dữ liệu và các tham số phân bố (quartile, deciles, ...), các đơn vị vật lý, các yếu tố quy mô, sự phụ thuộc lẫn nhau (ví dụ, các tham số có nguồn gốc, chẳng hạn như C = B / A, trong đó A, B và C thuộc bộ dữ liệu), giá trị còn thiếu, giá trị NULL, chỉ số (được sử dụng để nhận diện đối tượng dữ liệu, nhưng không phải là thuộc tính của đối tượng) và nhiều hơn nữa. Nếu bạn đang làm việc với dữ liệu có nhãn (đối với phân loại, phân tích dự đoán, hoặc dự án học có giám sát), thì bắt buộc phải xác định thuộc tính dữ liệu nào là nhãn lớp hoặc biến dự đoán. Một khía cạnh khác của quan điểm "hiểu biết dữ liệu của bạn" là nhớ tập trung vào các dữ liệu có thể thực hiện được (ví dụ: Các mô hình dữ liệu cân bằng được ưa thích, đôi khi được gọi là Razor của Occam, hoặc quy tắc của Einstein: "Các mô hình phải được thực hiện càng đơn giản càng tốt, nhưng không quá đơn giản", tránh "con ngựa hình cầu!"). Bằng cách tập trung vào các phần tử dữ liệu và các

biến đầu ra thông tin, hướng dẫn và cung cấp thông tin chi tiết về mục tiêu cuối cùng, bạn sẽ giảm được sự phân tâm và tạp âm ảnh hưởng đến tín hiệu.

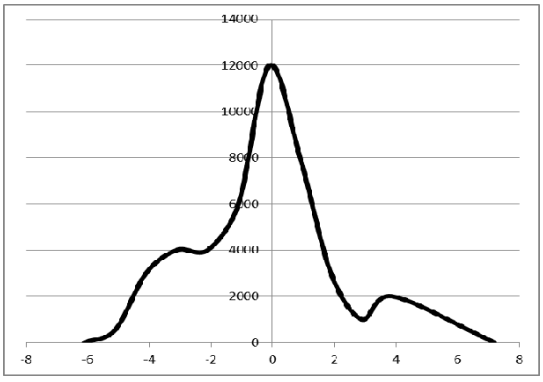
Trong thực tế, tôi cho rằng các tính năng này là bản chất của sự khám phá dữ liệu lớn: (a) việc thu thập các bộ dữ liệu lớn hiện nay cho phép chúng ta tìm ra những điều rất bất thường, ngạc nhiên, bất ngờ và thậm chí là thái quá trong lĩnh vực nghiên cứu (ví dụ, những điều chưa biết); và (b) phân bố dữ liệu tín hiệu tiếng ồn cao mà đôi khi dữ liệu lớn thu được lại cao hơn (ngoài các dữ liệu phản ánh: Trung bình, trung vị, mode và phương sai), điều này cho thấy các biến thể thú vị trong các đối tượng mà chúng ta đang điều tra.

Các hoạt động của ba nhà khoa học đã đáp ứng một cách chính xác các quan điểm cơ bản thứ ba của khoa học dữ liệu, đó là: Hãy nhớ rằng đây là khoa học! hay nói cách khác chúng ta đang thử nghiệm với các lựa chọn dữ liệu, kết hợp dữ liệu, thuật toán, kết hợp (cụm) của thuật toán, các đo lường chính xác và nhiều hơn nữa. Tất cả các mục này, tại một số điểm, được kiểm tra tính hợp lệ và tính khả thi của chúng đối với vấn đề mà bạn đang cố gắng giải quyết. Chúng ta có thể biết từ những kinh nghiệm trước đây về sự kết hợp dữ liệu, tính năng và các thuật toán nhất định sẽ đáp ứng được nhu cầu của chúng ta, nhưng ngay cả những kinh nghiệm đã học được (không phải là đoán) cũng chưa chắc đã áp dụng được. Hãy nhớ câu châm ngôn "Sự phán đoán tốt đến từ kinh nghiệm và kinh nghiệm xuất phát từ những phán đoán không tốt". Vì vậy, những lựa chọn tốt cho các thành phần thử nghiệm của dự án khoa học dữ liệu được rút ra từ kinh nghiệm, đặc biệt là từ các dự án thất bại. Hơn nữa, khoa học là một quá trình, liên quan đến quan sát, suy luận, tạo ra giả thuyết, thiết kế thực nghiệm, thu thập dữ liệu, thử nghiệm giả thuyết, ước lượng lỗi và sàng lọc giả thuyết. Bằng cách tuần tự làm theo các bước và theo chu kỳ (càng nhiều càng tốt để giảm sai sót và để tối ưu hóa độ chính xác), chúng ta có thể tránh được những sai sót dẫn đến những kết luận không đúng.

Quan điểm cơ bản thứ tư của khoa học dữ liệu là: Dữ liệu không bao giờ là hoàn hảo, nhưng tình yêu dữ liệu của bạn là vô điều kiện! Đây là khả năng thách thức lớn nhất và nguyên tắc bổ ích nhất để làm theo. Chúng ta thường cho rằng các dữ liệu tốt là những dữ liệu hoàn toàn sạch và phân phối bình thường. Thực tế là thế giới thực hiếm khi cung cấp cho chúng ta dữ liệu như vậy, chúng ta phải nhìn một cách toàn diện, đầy đủ bộ dữ liệu đó. Đối với tôi các bất thường trong dữ liệu như các dị thường, đuôi dài, bất đối xứng, và "mụn cóc" khác... thường nói với chúng ta điều gì đó rất quan trọng về lĩnh vực mà chúng ta đang nghiên cứu và/hoặc về các đối tượng trong miền đó. Ví dụ, giá trị ngoại lai thường bị sa thải và cắt bớt từ dữ liệu, đặc biệt là trong một số lĩnh vực khoa học mà tôi biết và yêu thích. Điều này là tốt nếu bạn có thể chắc chắn rằng, những giá trị đó chỉ đơn giản là tạp âm hoặc thành phần giả tạo trong các dữ liệu. Tuy nhiên, nếu những thứ đó đại diện cho một đối tượng hoàn toàn mới hoặc một kiểu hành vi mới thì cần phải xem xét.

Vì vậy, tôi thích gọi phát hiện ngoại lai bởi một tên tốt hơn (dữ liệu nhiều hơn): Khám phá ngạc nhiên. Các đặc điểm bất thường của dữ liệu (mà không tuân theo các tiêu chuẩn) là những điều mới lạ, thú vị và đáng ngạc nhiên. Hãy dành một chút thời gian với những tính năng: Đuôi dài, Q-Q plot, các giá trị ngoại lai và các tính năng khác (chẳng hạn như phân phối nhiều mode). Tìm hiểu những đặc điểm khác nhau trong dữ liệu của bạn (chẳng hạn như sự đa dạng thú vị của tính năng hiển thị trong phân phối dữ liệu trong hình 1).

Hình 1: Biểu đồ dữ liệu này từ một bộ sưu tập dữ liệu lớn giả định cho thấy nhiều đỉnh núi, thung lũng và đuôi trong phân bố. Mỗi đặc điểm của biểu đồ cung cấp những hiểu biết có giá trị có thể có trong tổng thể.



Tôi cho rằng các tính năng này là trong thực tế, bản chất của sự khám phá dữ liệu lớn: (a) các bộ sưu tập dữ liệu lớn hiện nay cho phép chúng ta tìm thấy rất không bình thường, đáng ngạc nhiên, bất ngờ, và thậm chí cả những thứ kỳ quặc trong phạm vi của chúng ta về nghiên cứu (ví dụ, các ẩn số chưa biết); và (b) phân bố dữ liệu tín hiệu tiếng ồn cao mà sản lượng dữ liệu lớn có nhiều khoảnh khắc cao hơn (ngoài các trung bình, trung vị, mode và phương sai) cho thấy các biến thể thú vị trong các đối tượng mà chúng ta đang điều tra. Áp dụng một số kiểm tra thống kê phi tham số về dữ liệu của bạn và bước vào một thế giới mới của phát hiện dựa vào dữ liệu. Nếu chưa có, bạn sẽ sớm "yêu dữ liệu của bạn" vì tính đa dạng của nó. Trên thực tế, các nhà khoa học đã khuyến cáo, giá trị cao được đặt trên sự đa dạng trong các khuyến nghị (nghĩa là chỉ đơn giản là giới thiệu sản phẩm "rõ ràng" cho người tiêu dùng không phải là sẽ giành chiến thắng và giữ chân được những khách hàng, trong khi cung cấp những điều thú vị, khác thường những sản phẩm có liên quan là một người chiến thắng chắc chắn, giống như là chiến thắng cuộc đua ngựa).

Cuối cùng, quan điểm cơ bản thứ năm về khoa học dữ liệu là: Quá phức tạp là một tội lỗi đối với khoa học dữ liệu! Trong khi chúng ta đã chỉ trích nhà vật lý trong trò đùa về việc trang bị cho mô hình ngựa quá đơn giản (với một mô tả hình học đơn giản), chúng ta cũng có thể có một "nhà khoa học" thứ tư trong trò đùa (chọn người yêu thích của bạn!), Người đã tạo ra mô hình "Rube Goldberg" Con ngựa chiến thắng - một mô hình được thiết kế quá kỹ lưỡng và quá phức tạp. Đây là một "tội lỗi" đối với khoa học dữ liệu theo nghĩa: Vì quan điểm số 3 (khoa học dữ liệu là khoa học), chúng ta nên kiểm tra, xác nhận và xác minh độ chính xác của các mô hình bằng cách sử dụng dữ liệu thử nghiệm và dữ liệu "trước đây không nhìn thấy". Quy trình khoa học này bảo vệ chúng ta khỏi bị quá phức tạp (phương sai cao) và quá đơn giản (độ chệch cao) trong các giải pháp mô hình của chúng ta. Nếu chúng ta bỏ qua các nguyên tắc của khoa học tốt, thì chúng ta có xu hướng lạm dụng và thiên vị. Ngoài ra, bằng cách áp dụng các quan điểm số (2) và số (4) một cách nghiêm túc, chúng ta phải biết được sự khác biệt trong giá trị dữ liệu, do đó cần phải cảnh báo sớm trong quá trình mô hình cuối cùng vẫn có thể chấp nhận được (nếu không được chấp nhận hơn) khi không chắc chắn về mô hình phản ánh sai sự thật trong dữ liệu của chúng ta.

Tóm lại, tất cả các sinh viên và học viên của khoa học dữ liệu nên tránh "ngựa hình cầu" và tuân thủ theo năm nguyên tắc cơ bản về khoa học dữ liệu, đó là: Bắt đầu với kết thúc trong tâm trí; hiểu, biết dữ liệu của bạn; hãy nhớ rằng đây là khoa học; dữ liệu là không bao giờ hoàn hảo, nhưng tình yêu dữ liệu của bạn là mãi mãi; quá phức tạp là một tội lỗi chống lại khoa học dữ liệu.

Anh Tuấn (dịch)

Nguồn: http://www.statisticsviews.com/details/feature/5459931/Five-Fundamental-Concepts-of-Data-Science.html