ĐẠI HỌC MỎ - ĐỊA CHẤT

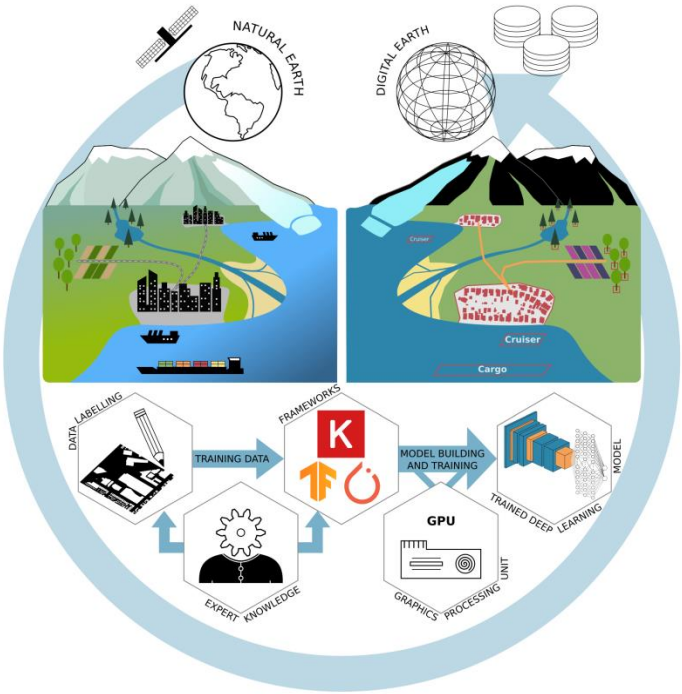
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO HỌC THUẬT

MỘT SỐ THUẬT TOÁN THỐNG KÊ TRONG KHOA HỌC DỮ LIỆU

Người thực hiện báo cáo

ThS. Trương Xuân Bình



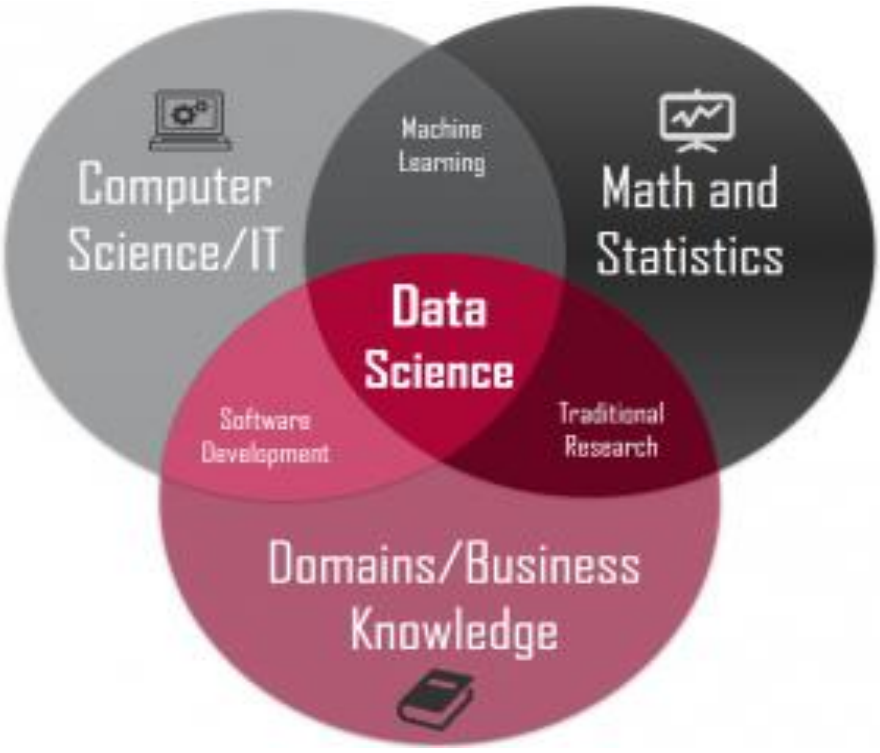
Hà Nội – 2023

MỤC LỤC

MỞ ĐẦU

Bất kể bạn có quan điểm thế nào về khoa học dữ liệu, chúng ta không thể bỏ qua tầm quan trọng của dữ liệu cũng như khả năng phân tích, sắp xếp và bối cảnh hóa dữ liệu. Dựa trên kho dữ liệu việc làm khổng lồ và phản hồi của nhân viên, Glassdoor xếp hạng Nhà khoa học dữ liệu đứng số 1 trong danh sách 25 việc làm tốt nhất ở Mỹ. Vì vậy, công việc này sẽ luôn tồn tại, nhưng những chi tiết cụ thể gì trong công việc của một nhà khoa học dữ liệu sẽ thay đổi và phát triển. Khi các công nghệ như Machine Learning trở nên phổ biến hơn bao giờ hết, và các lĩnh vực mới nổi như Deep Learning đạt được sức hút đáng kể trong giới nghiên cứu và kỹ sư – và các công ty thuê họ – các nhà khoa học dữ liệu tiếp tục làn sóng đổi mới và các tiến bộ công nghệ đáng kinh ngạc.

Mặc dù có khả năng code giỏi là rất quan trọng, nhưng khoa học dữ liệu không phải là công nghệ phần mềm (trên thực tế, sẽ tốt nếu có kiến thức về Python). Khả năng của các nhà khoa học dữ liệu nằm ở giao lộ của code, thống kê và tư duy phản biện. Như Josh Wills đã nói, “Nhà khoa học dữ liệu là người giỏi thống kê hơn bất kỳ lập trình viên nào và giỏi lập trình hơn bất kỳ nhà thống kê nào”. Cá nhân tôi biết quá nhiều kỹ sư phần mềm đang tìm cách chuyển đổi thành nhà khoa học dữ liệu và sử dụng một cách mù quáng các khung học máy như TensorFlow hoặc Apache Spark vào dữ liệu của họ mà không có sự hiểu biết thấu đáo về các lý thuyết thống kê đằng sau chúng. Vì vậy, cần học thống kê, một khung lý thuyết cho Machine Learning từ các lĩnh vực thống kê và phân tích chức năng.



Tại sao nên học statistical learning?

Điều quan trọng là phải hiểu các ý tưởng đằng sau các kỹ thuật khác nhau, để biết làm thế nào và khi nào sử dụng chúng. Trước tiên, người ta phải hiểu các phương pháp đơn giản hơn, để nắm bắt những phương pháp tinh vi hơn. Điều quan trọng là phải đánh giá chính xác hiệu suất của một phương pháp, để biết nó hoạt động tốt hay không. Ngoài ra, đây là một lĩnh vực nghiên cứu thú vị, có các ứng dụng quan trọng trong khoa học, công nghiệp và tài chính. Cuối cùng, thống kê là một thành phần cơ bản trong đào tạo của một nhà khoa học dữ liệu hiện đại. Ví dụ về các vấn đề thống kê bao gồm:

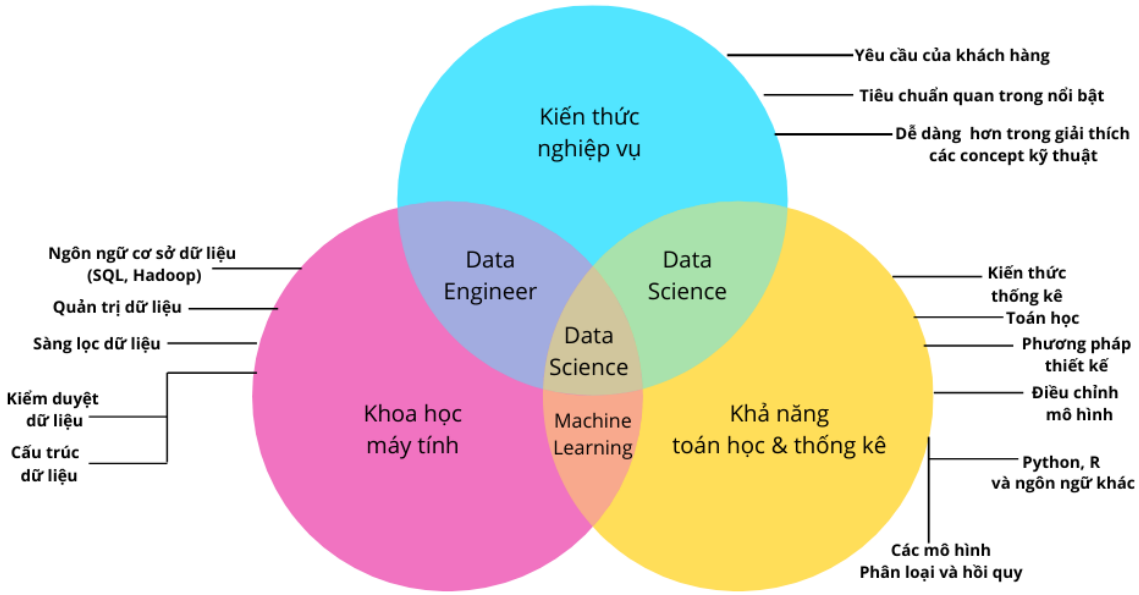
- Xác định các yếu tố nguy cơ ung thư tuyến tiền liệt.
- Phân loại âm vị được ghi dựa trên biểu đồ log-periodogram.
- Dự đoán liệu một người nào đó sẽ bị đau tim trên cơ sở nhân khẩu học, chế độ ăn uống và đo lường lâm sàng.
- Tùy chỉnh một hệ thống phát hiện email rác.
- Xác định các số trong một mã zip viết tay.
- Phân loại một mẫu mô thành một trong các nhóm ung thư.
- Thiết lập mối quan hệ giữa tiền lương và các biến nhân khẩu học trong dữ liệu khảo sát dân số.

Trong học kỳ cuối cùng của tôi ở trường đại học, tôi đã làm một nghiên cứu độc lập về khai thác dữ liệu. Lớp học bao gồm các tài liệu mở rộng đến từ 3 cuốn sách: Intro to Statistical Learning (Hastie, Tibshirani, Witten, James), Doing Bayesian Data Analysis (Kruschke), và Time Series Analysis and Applications (Shumway, Stoffer). Chúng tôi đã thực hiện rất nhiều bài tập về phân tích Bayes, Markov Chain Monte Carlo, mô hình phân cấp, học có giám sát và không giám sát. Trải nghiệm này làm tăng sự quan tâm của tôi trong lĩnh vực học thuật về khai thác dữ liệu (data mining) và thuyết phục tôi đi chuyên sâu hơn về lĩnh vực này. Gần đây, tôi đã hoàn thành khóa học trực tuyến về thống kê trên Stanford Lagunita, bao gồm cả tự đọc cuốn Intro to Statistical Learning book. Dựa trên kinh nghiệm bản thân, tôi muốn chia sẻ 10 kỹ thuật thống kê từ cuốn sách mà tôi tin rằng bất kỳ nhà khoa học dữ liệu nào cũng nên học để xử lý các bộ dữ liệu lớn hiệu quả hơn.

Trước khi tiếp tục với 10 kỹ thuật này, tôi muốn phân biệt giữa statistical learning và machine learning. Đã viết một trong những bài viết nổi tiếng nhất trên Medium về machine learning, tôi tự tin rằng mình có chuyên môn về những khác biệt này:

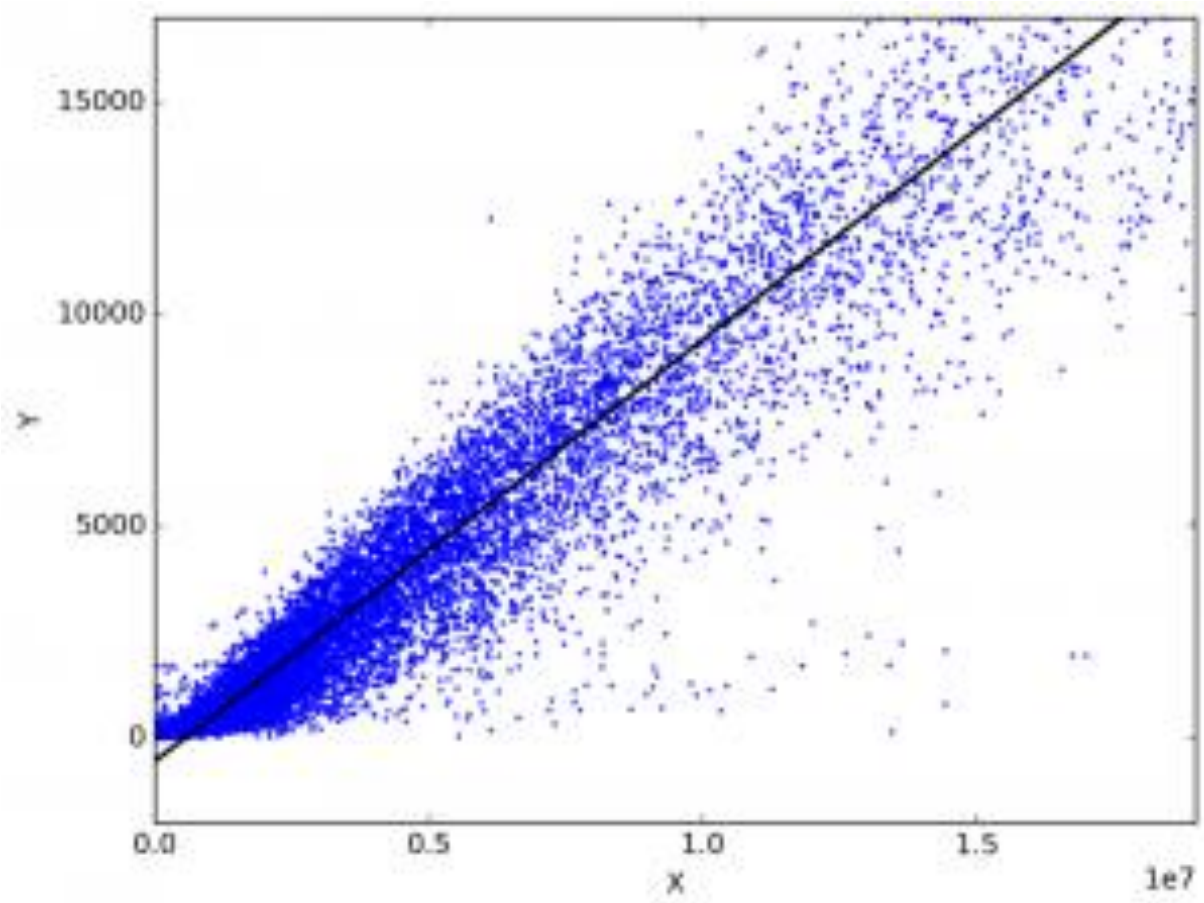
- Machine learning là một lĩnh vực nhỏ của Trí tuệ nhân tạo.
- Statistical learning là một lĩnh vực nhỏ của Thống kê.

- Machine learning nhấn mạnh vào các ứng dụng quy mô lớn và độ chính xác của dự đoán.
- Statistical learning nhấn mạnh vào mô hình và tính dễ hiểu của mô hình, và độ chính xác và không chắc chắn.
- Nhưng sự khác biệt đã trở nên ngày càng mờ nhạt, và có rất nhiều "giao thoa chéo".
- Machine learning có ưu thế trong Marketing!



1. Hồi quy tuyến tính

Trong thống kê, hồi quy tuyến tính là một phương pháp để dự đoán một biến mục tiêu bằng cách khớp mối quan hệ tuyến tính tốt nhất giữa biến phụ thuộc và biến độc lập. Sự phù hợp tốt nhất được thực hiện bằng cách đảm bảo rằng tổng của tất cả các khoảng cách giữa hình dạng và các quan sát thực tế tại mỗi điểm càng nhỏ càng tốt. Với mỗi hình dạng, sự phù hợp của hình dạng là cao nhất khi không có vị trí của hình tạo ra ít lỗi hơn. 2 loại chính của hồi quy tuyến tính là Hồi quy tuyến tính đơn giản và Hồi quy tuyến tính đa biến. Hồi quy tuyến tính đơn giản sử dụng một biến độc lập duy nhất để dự đoán một biến phụ thuộc. Nhiều hồi quy tuyến tính sử dụng nhiều hơn một biến độc lập để dự đoán một biến phụ thuộc.



Chọn bất kỳ 2 thứ mà bạn sử dụng trong cuộc sống hàng ngày và có liên quan. Giống như, tôi có dữ liệu về chi tiêu hàng tháng, thu nhập hàng tháng và số chuyến đi mỗi tháng trong 3 năm qua. Bây giờ tôi cần trả lời các câu hỏi sau:

Chi tiêu hàng tháng của tôi cho năm tới là gì?

Yếu tố nào (thu nhập hàng tháng hoặc số chuyến đi mỗi tháng) quan trọng hơn trong việc quyết định chi tiêu hàng tháng của tôi?

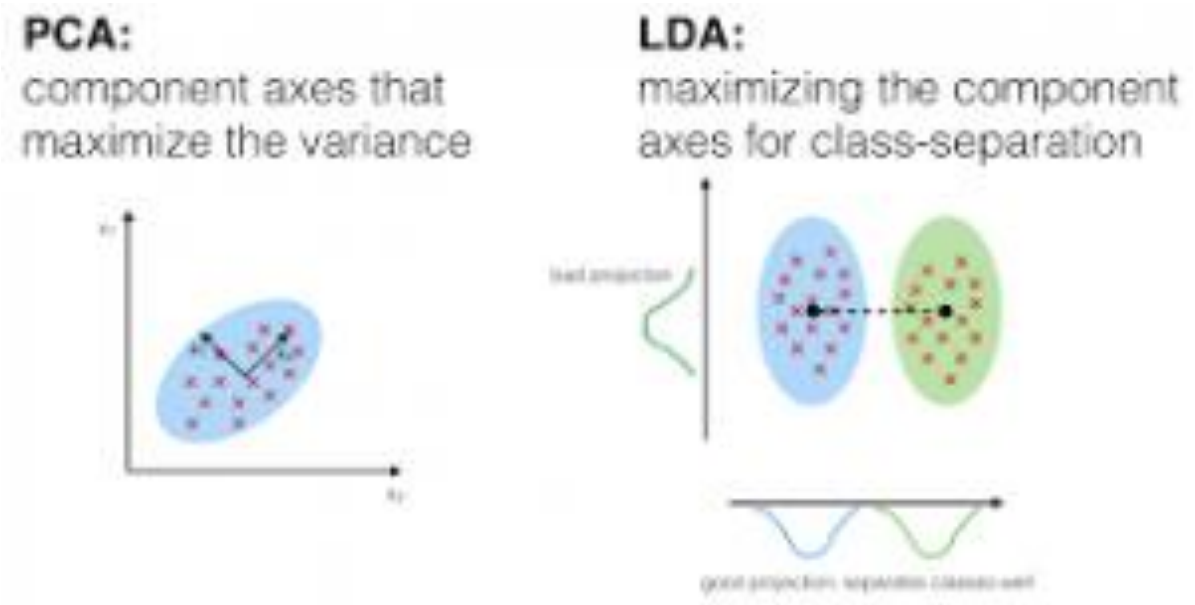
Thu nhập hàng tháng và các chuyến đi mỗi tháng tương quan với chi tiêu hàng tháng như thế nào?

2. Thuật toán Phân loại

Phân loại là một kỹ thuật khai thác dữ liệu gán các danh mục cho một tập hợp dữ liệu để hỗ trợ cho các dự đoán và phân tích chính xác hơn. Cũng đôi khi được gọi là Cây quyết định, phân loại là một trong một số phương pháp nhằm làm cho việc phân tích các bộ dữ liệu rất lớn có hiệu quả. 2 kỹ thuật phân loại chính nổi bật: Hồi quy logistic và phân tích phân biệt.

Hồi quy logistic là phân tích hồi quy thích hợp để tiến hành khi biến phụ thuộc là nhị phân. Giống như tất cả các phân tích hồi quy, hồi quy logistic là phân tích dự đoán. Hồi quy logistic được sử dụng để mô tả dữ liệu và để giải thích mối quan hệ giữa một biến nhị phân phụ thuộc và một hoặc nhiều biến độc lập. Các loại câu hỏi mà hồi quy logistic có thể kiểm tra:

- Xác suất bị ung thư phổi (Có so với Không) thay đổi thế nào theo mỗi pound thừa cân và cho mỗi gói thuốc lá hút mỗi ngày?
- Lượng calo cơ thể tiêu thụ, lượng chất béo và tuổi tham gia có ảnh hưởng thế nào đến cơn đau tim (Có so với Không)?



Trong Phân tích Phân biệt, 2 hoặc nhiều nhóm hoặc cụm hoặc quần thể được biết là một tiên nghiệm và 1 hoặc nhiều quan sát mới được phân loại thành 1 trong các quần thể đã biết dựa trên các đặc điểm đo được. Mô hình phân tích phân biệt phân chia các yếu tố dự đoán X riêng biệt thành từng lớp phản ứng và sau đó sử dụng định lý Bayes để ước tính xác suất của các loại phản ứng với mỗi giá trị của X. Các mô hình như vậy có thể là tuyến tính hoặc quadratic.

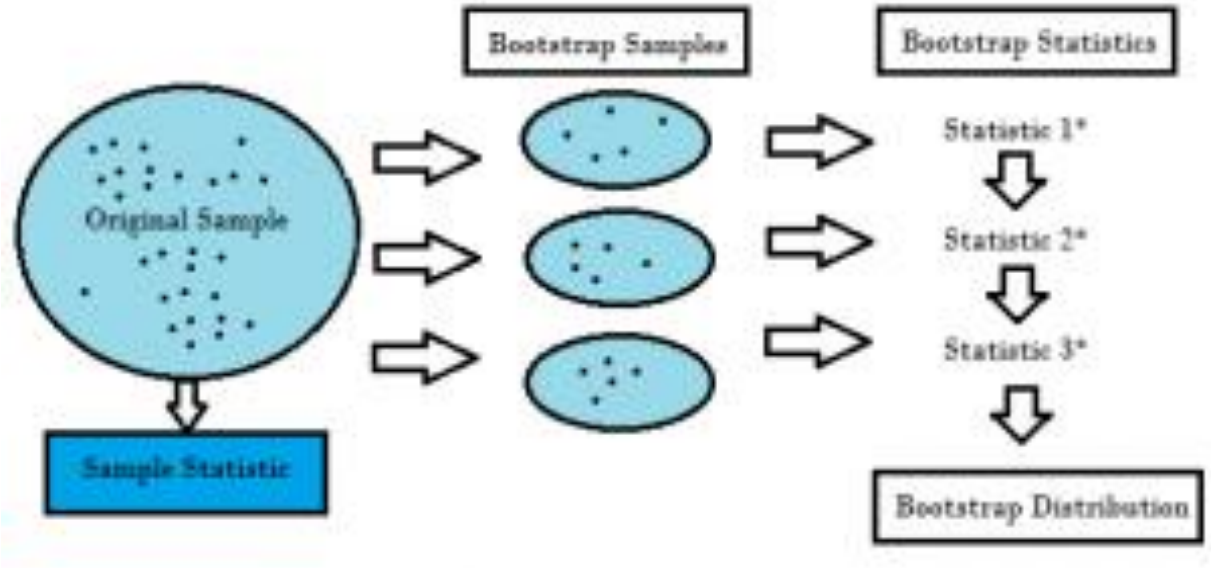
- Phân tích phân biệt tuyến tính tính "điểm số phân biệt" cho mỗi quan sát để phân loại lớp biến số phản ứng của nó. Những điểm số này có được bằng cách tìm kết hợp tuyến tính của các biến độc lập. Nó giả định rằng các quan sát trong mỗi lớp được rút ra từ phân phối Gaussian đa biến và hiệp phương sai của các biến dự đoán là phổ biến trên tất cả các cấp k của biến phản ứng Y.
- Phân tích phân biệt quadratic cung cấp một cách tiếp cận khác. Giống như LDA, QDA giả định rằng các quan sát từ mỗi lớp Y được rút ra từ phân phối Gaussian.

Tuy nhiên, không giống như LDA, QDA giả định rằng mỗi lớp có ma trận hiệp phương sai riêng. Nói cách khác, các biến dự đoán không được coi là có phương sai chung trên mỗi cấp độ k trong Y.

3. Phương pháp lấy mẫu lại

Lấy mẫu lại là phương pháp lấy lại nhiều mẫu nhỏ từ mẫu dữ liệu gốc. Nó là một phương pháp không tham số của suy luận thống kê. Nói cách khác, phương pháp lấy mẫu lại không liên quan đến việc sử dụng các bảng phân phối chung để tính các giá trị xác suất p gần đúng.

Lấy mẫu lại tạo ra một phân phối lấy mẫu duy nhất trên cơ sở dữ liệu thực tế. Nó sử dụng các phương pháp thử nghiệm, thay vì phương pháp phân tích, để tạo ra phân phối lấy mẫu duy nhất. Nó mang lại các ước tính không thiên vị vì nó dựa trên các mẫu không thiên vị của tất cả các kết quả có thể có của dữ liệu được nghiên cứu bởi nhà nghiên cứu. Để hiểu khái niệm lấy mẫu lại, bạn nên hiểu các thuật ngữ Bootstrapping và Xác thực chéo:



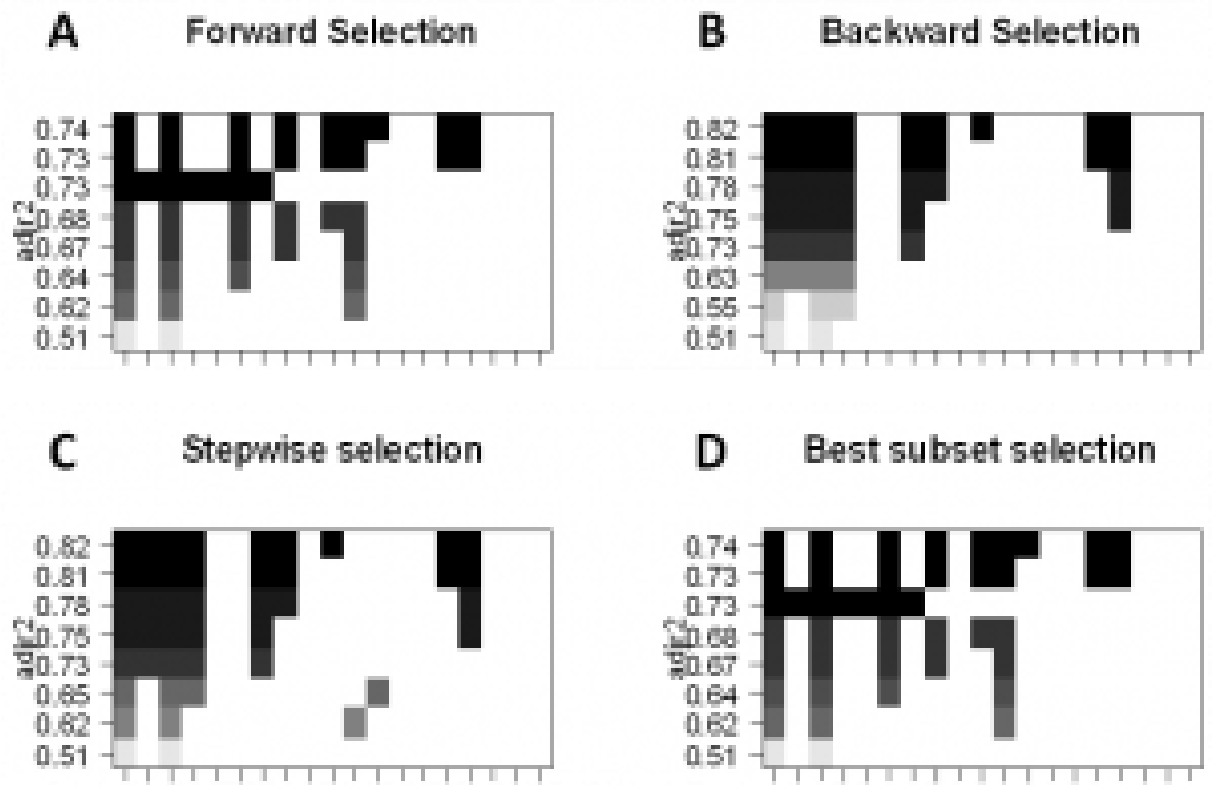
- Bootstrapping là một kỹ thuật giúp trong nhiều tình huống như xác nhận hiệu suất mô hình dự đoán, phương pháp tập hợp, ước tính độ lệch và phương sai của mô hình. Nó hoạt động bằng cách lấy mẫu với sự thay thế từ dữ liệu gốc và lấy các điểm dữ liệu không được chọn của làm trường hợp thử nghiệm. Chúng ta có thể thực hiện điều này nhiều lần và tính điểm trung bình cho ước tính hiệu suất mô hình
- Mặt khác, xác nhận chéo là một kỹ thuật để xác nhận hiệu suất của mô hình và nó được thực hiện bằng cách chia dữ liệu đào tạo thành k phần. Chúng ta lấy bộ

phận k – 1 làm bộ huấn luyện. Chúng tôi lặp lại rằng k lần khác nhau. Cuối cùng, chúng tôi lấy điểm trung bình của điểm k làm ước tính hiệu suất.

Thông thường đối với các mô hình tuyến tính, bình phương tối thiểu thông thường là tiêu chí chính để xem xét để độ phù hợp với chúng vào dữ liệu. 3 phương pháp tiếp theo là các phương pháp thay thế có thể cung cấp độ chính xác dự đoán tốt hơn và tính dễ hiểu của mô hình để phù hợp với các mô hình tuyến tính.

4. Chọn tập hợp con

Cách tiếp cận này xác định một tập hợp con của các yếu tố dự đoán p mà chúng ta tin rằng có liên quan đến phản ứng. Sau đó, chúng ta làm mô hình cho phù bằng cách sử dụng bình phương tối thiểu của các tính năng tập hợp con.

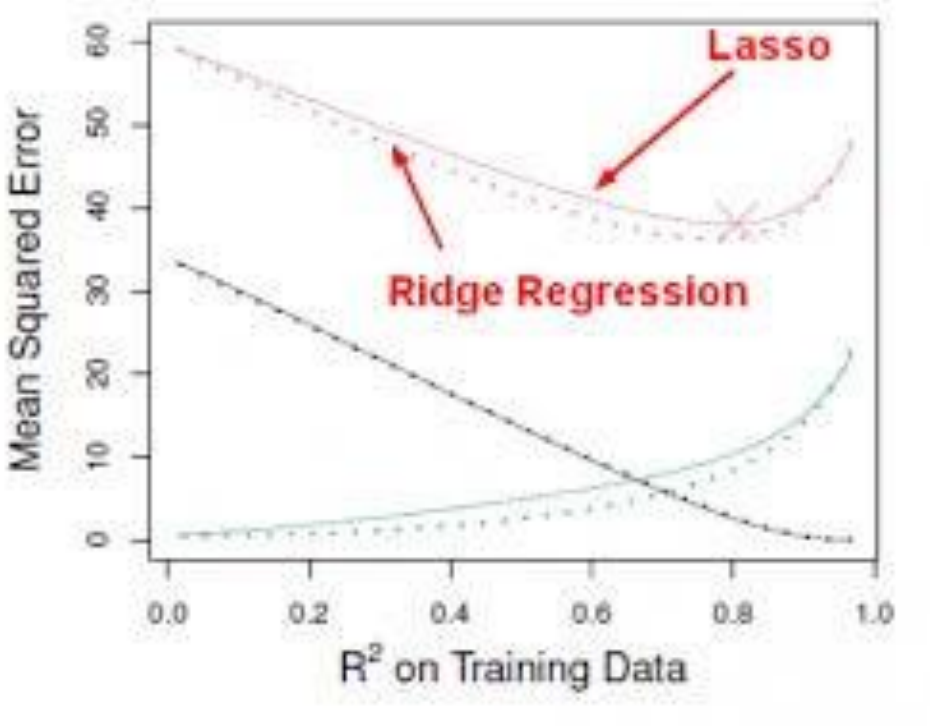


- Lựa chọn tập hợp con tốt nhất: Ở đây chúng tôi phù hợp với hồi quy OLS riêng cho từng kết hợp có thể có của p biến dự đoán và sau đó xem xét mô hình kết quả phù hợp. Thuật toán được chia thành 2 giai đoạn: (1) Phù hợp với tất cả các mô hình có chứa k dự đoán, trong đó k là độ dài tối đa của các mô hình, (2) Chọn một mô hình duy nhất sử dụng lỗi dự đoán xác thực chéo. Điều quan trọng là sử dụng thử nghiệm hoặc xác nhận lỗi và không luyện lỗi để đánh giá mức độ phù hợp của mô hình vì RSS và R^2 tăng khi số lượng biến. Cách tiếp cận tốt nhất là xác thực chéo và chọn mô hình có R^2 cao nhất và RSS thấp nhất trong các ước tính lỗi thử nghiệm.

- Lựa chọn forward stepwise xét một tập hợp con nhỏ hơn gồm p biến dự đoán. Nó bắt đầu với một mô hình không chứa các biến dự đoán, sau đó thêm các biến dự đoán vào mô hình, từng lần một cho đến khi tất cả các biến dự đoán nằm trong mô hình khoa học dữ liệu. Thứ tự của các biến được thêm vào là biến mà cải thiện sự phù hợp của mô hình nhiều nhất, cho đến khi không còn biến nào cải thiện hơn bằng cách sử dụng lỗi dự đoán xác thực chéo.
- Lựa chọn backward stepwise bắt đầu với tất cả p biến dự đoán trong mô hình, sau đó lặp đi lặp lại loại bỏ dự đoán ít hữu ích nhất tại một thời điểm.
- Phương thức hybrid theo phương pháp lựa chọn forward stepwise, tuy nhiên, sau khi thêm từng biến mới, phương thức cũng có thể loại bỏ các biến không đóng góp cho mô hình.

5. Shrinkage trong khoa học dữ liệu

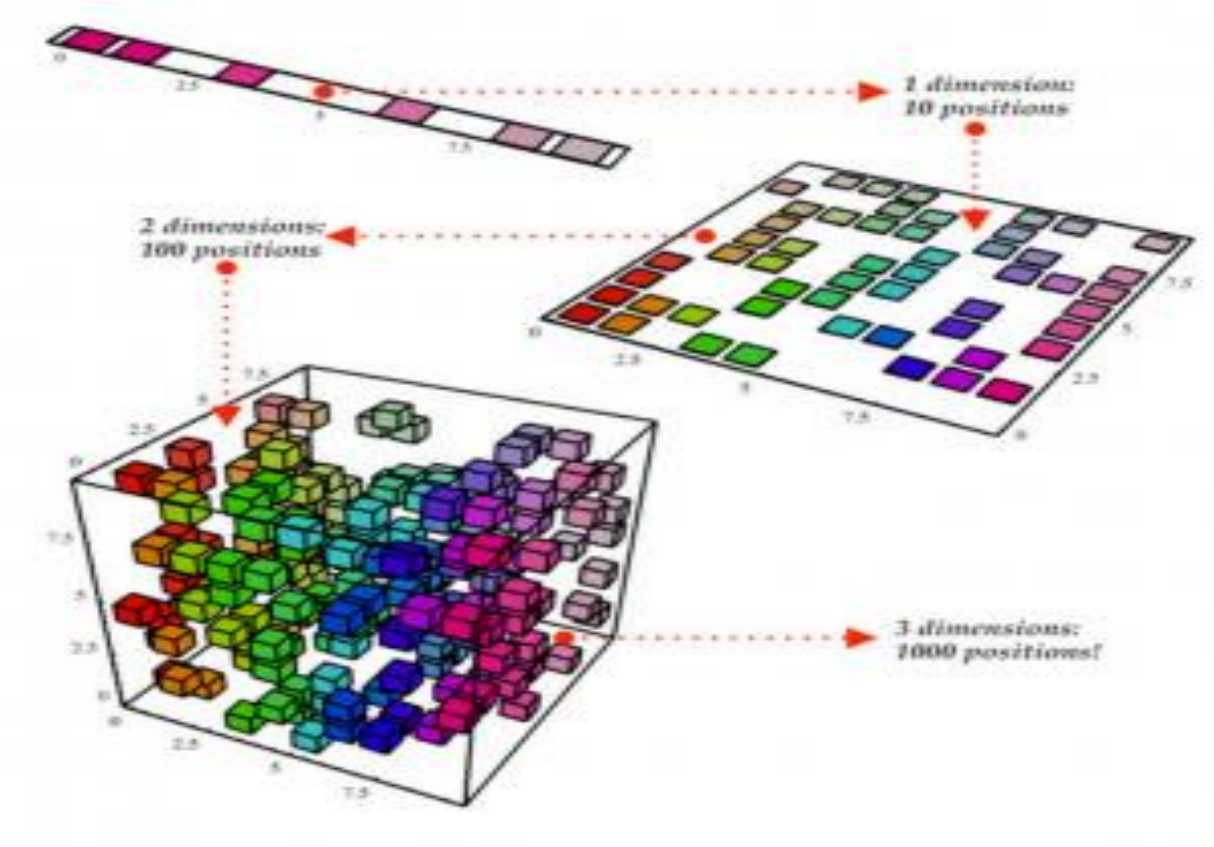
Cách tiếp cận này phù hợp với một mô hình liên quan đến tất cả p biến dự đoán, tuy nhiên, các hệ số ước tính được thu nhỏ về 0 so với ước lượng bình phương nhỏ nhất. Sự thu hẹp này, còn gọi là chính quy hóa có tác dụng làm giảm phương sai. Tùy thuộc vào loại shrinkage được thực hiện, một số hệ số có thể được ước tính là chính xác bằng không. Do đó phương pháp này cũng thực hiện lựa chọn biến. Hai kỹ thuật nổi tiếng nhất để thu hẹp các ước tính hệ số về 0 là hồi quy ridge và lasso.



- Hồi quy ridge tương tự như bình phương tối thiểu ngoại trừ các hệ số được ước tính bằng cách giảm thiểu một lượng hơi khác nhau. Hồi quy độ dốc, như OLS, tìm kiếm các ước tính hệ số làm giảm RSS, tuy nhiên chúng cũng có hình phạt co rút khi các hệ số tiến gần đến 0. Hình phạt này có tác dụng thu hẹp các ước tính hệ số về không. Không đi sâu vào toán học, sẽ hữu ích khi biết rằng hồi quy sườn thu nhỏ các tính năng với phương sai không gian cột nhỏ nhất. Giống như trong phân tích thành phần tiên đoán, hồi quy sườn dự án dữ liệu vào không gian hướng và sau đó thu nhỏ các hệ số của các thành phần phương sai thấp hơn các thành phần phương sai cao, tương đương với các thành phần chính lớn nhất và nhỏ nhất.
- Hồi quy ridge có ít nhất một nhược điểm; nó bao gồm tất cả các yếu tố dự đoán p trong mô hình cuối cùng. Thời hạn phạt sẽ đặt nhiều trong số chúng gần bằng 0, nhưng không bao giờ chính xác bằng không. Điều này nói chung là một vấn đề cho độ chính xác dự đoán, nhưng nó có thể làm cho mô hình khó diễn giải kết quả hơn. Lasso khắc phục nhược điểm này và có khả năng buộc một số hệ số về 0 với điều kiện là s đủ nhỏ. Vì s = 1 dẫn đến hồi quy OLS thông thường, khi s tiến đến 0, các hệ số co lại về 0. Do đó, hồi quy Lasso cũng thực hiện lựa chọn biến.

6. Giảm kích thước trong khoa học dữ liệu

Giảm kích thước làm giảm vấn đề ước tính các hệ số p + 1 thành vấn đề đơn giản của các hệ số M + 1, trong đó M <p. Điều này đạt được bằng cách tính M các kết hợp tuyến tính hoặc các phép chiếu khác nhau của các biến. Sau đó, các phép chiếu M này được sử dụng làm công cụ dự đoán để phù hợp với mô hình hồi quy tuyến tính theo bình phương tối thiểu. 2 cách tiếp cận cho nhiệm vụ này là hồi quy thành phần chính và bình phương nhỏ nhất một phần.



Người ta có thể mô tả Hồi quy các thành phần chính như một cách tiếp cận để tạo ra một tập hợp các tính năng chiều thấp từ một tập hợp lớn các biến. Hướng thành phần chính đầu tiên của dữ liệu là theo đó các quan sát thay đổi nhiều nhất. Nói cách khác, PC đầu tiên là một dòng phù hợp nhất có thể với dữ liệu. Chúng ta có thể phù hợp với p thành phần chính khác biệt. PC thứ hai là sự kết hợp tuyến tính của các biến không tương thích với PC thứ nhất và có phương sai lớn nhất. Ý tưởng là các thành phần chính thu được nhiều phương sai nhất trong dữ liệu bằng cách sử dụng kết hợp tuyến tính của dữ liệu theo các hướng trực giao sau đó. Theo cách này, chúng ta cũng có thể kết hợp các tác động của các biến tương quan để có thêm thông tin từ dữ liệu có sẵn, trong khi ở các bình phương tối thiểu thông thường, chúng ta sẽ phải loại bỏ một trong các biến tương quan.

Phương pháp PCR mà chúng tôi mô tả ở trên liên quan đến việc xác định các kết hợp tuyến tính của X đại diện tốt nhất cho các yếu tố dự đoán. Các kết hợp này được xác định theo cách không được giám sát, vì phản hồi Y không được sử dụng để giúp xác định các hướng thành phần chính. Nghĩa là, phản hồi Y không giám sát việc xác định các thành phần chính, do đó không có gì đảm bảo rằng các hướng giải thích tốt nhất cho các dự đoán cũng là tốt nhất để dự đoán phản hồi (mặc dù điều đó thường được giả định). Bình phương tối thiểu một phần (PLS) là một giám sát đối với PCR. Giống như PCR, PLS là phương pháp giảm kích thước, trước tiên xác định một bộ tính năng nhỏ

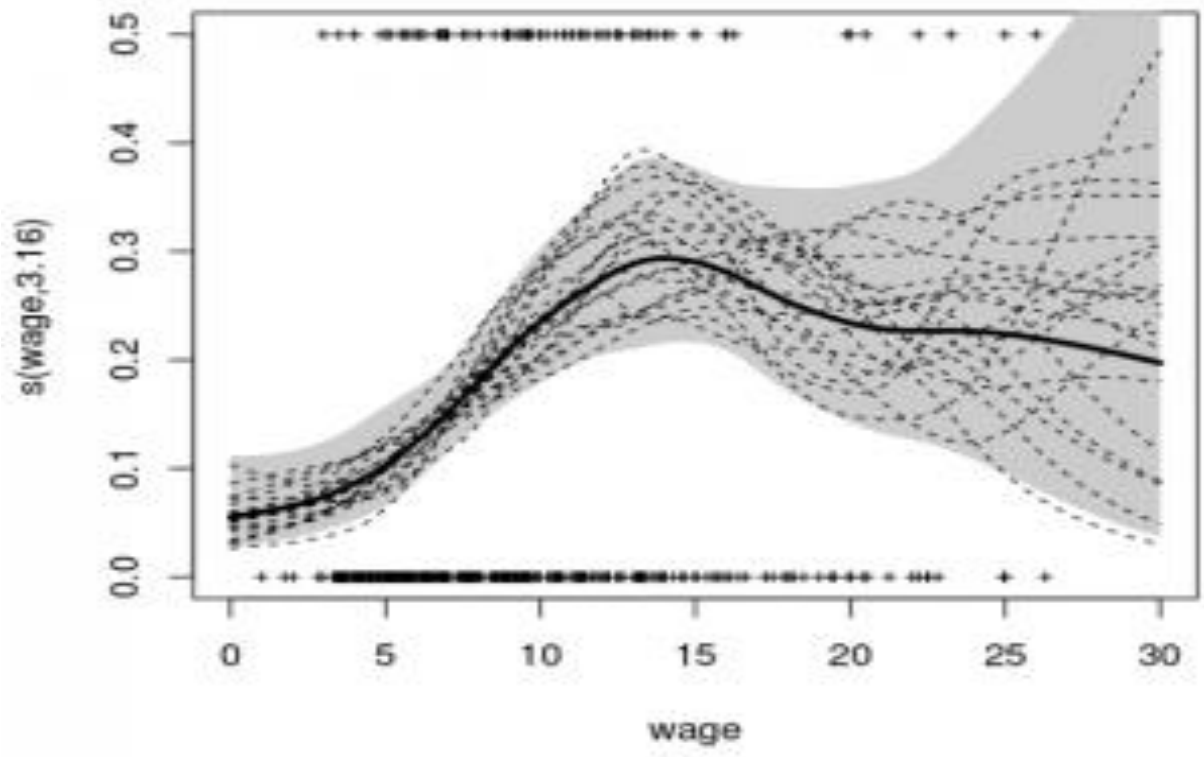
hơn mới là sự kết hợp tuyến tính của các tính năng gốc, sau đó khớp với mô hình tuyến tính thông qua các bình phương tối thiểu cho các tính năng M mới. Tuy nhiên, không giống như PCR, PLS sử dụng biến phản ứng để xác định các tính năng mới.

7. Mô hình phi tuyến

Trong thống kê, hồi quy phi tuyến là một dạng phân tích hồi quy trong đó dữ liệu quan sát được mô hình hóa bởi một hàm là sự kết hợp phi tuyến của các tham số mô hình và phụ thuộc vào một hoặc nhiều biến độc lập. Các dữ liệu được trang bị bởi một phương pháp xấp xỉ liên tiếp. Dưới đây là một số kỹ thuật quan trọng để đối phó với các mô hình phi tuyến:

Hàm trên các số thực được gọi là hàm bước nếu nó có thể được viết dưới dạng kết hợp tuyến tính hữu hạn của các hàm chỉ thị của các khoảng. Nói một cách không chính thức, một hàm bước là một hàm hằng số chỉ có nhiều phần.

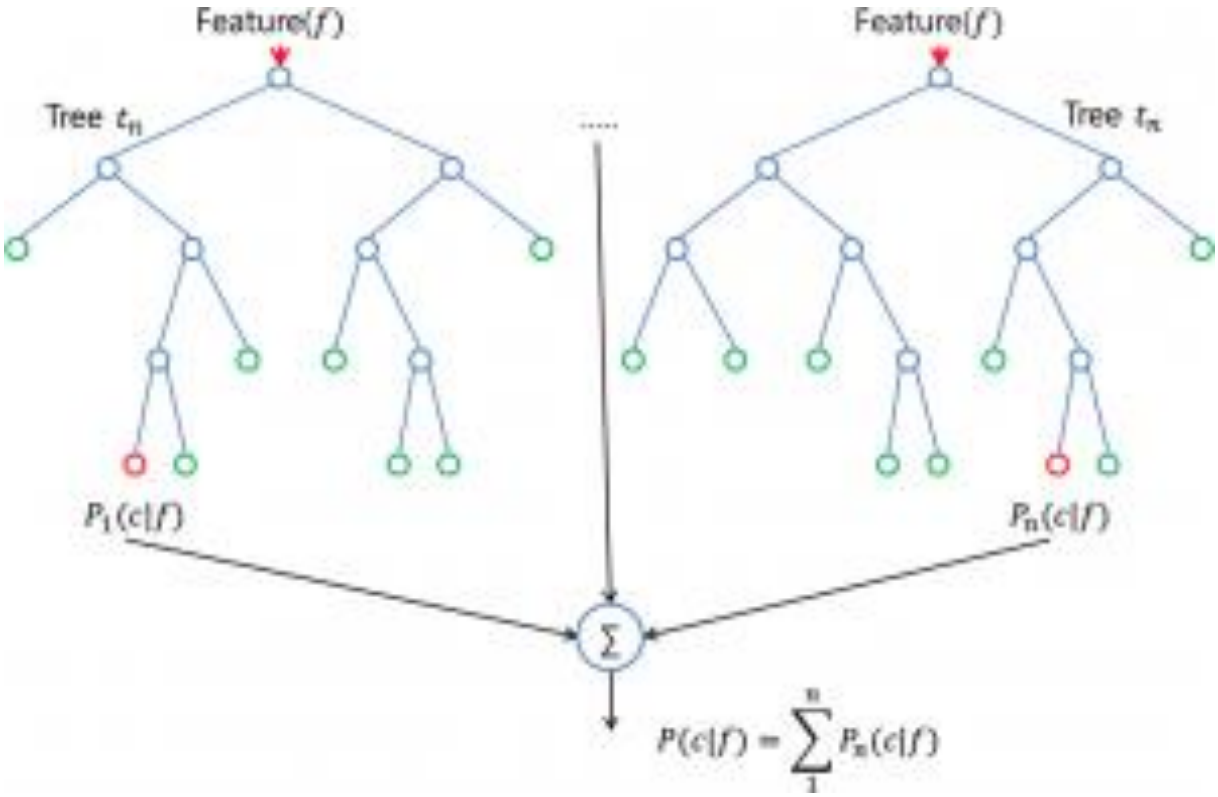
Hàm piecewise là một hàm được xác định bởi nhiều hàm phụ, mỗi hàm phụ áp dụng cho một khoảng nhất định của miền chính của hàm. Piecewise thực sự là một cách thể hiện chức năng, chứ không phải là một đặc tính của chính chức năng, nhưng với trình độ bổ sung, nó có thể mô tả bản chất của chức năng. Ví dụ, hàm đa thức piecewise là một hàm là đa thức trên mỗi miền con của nó, nhưng có thể là một hàm khác nhau trên mỗi miền.



- Một spline là một chức năng đặc biệt được xác định từng phần bởi các đa thức. Trong đồ họa máy tính, spline dùng để chỉ một đường cong tham số đa thức piecewise. Splines là những đường cong phổ biến vì sự đơn giản trong xây dựng của chúng, sự dễ dàng và chính xác của việc đánh giá và khả năng của chúng để xấp xỉ các hình dạng phức tạp thông qua khớp đường cong và thiết kế đường cong tương tác.
- Mô hình generalized addictive là mô hình tuyến tính tổng quát, trong đó bộ dự báo tuyến tính phụ thuộc tuyến tính vào các hàm trơn chưa biết của một số biến dự đoán và quan tâm tập trung vào suy luận về các hàm trơn này.

8. Phương pháp dựa trên cây

Các phương pháp dựa trên cây có thể được sử dụng cho cả các vấn đề hồi quy và phân loại. Chúng liên quan đến việc phân tầng hoặc phân đoạn không gian dự đoán thành một số vùng đơn giản. Do tập hợp các quy tắc phân tách được sử dụng để phân đoạn không gian dự đoán có thể được tóm tắt trong một cây, các kiểu tiếp cận này được gọi là các phương thức cây quyết định. Các phương pháp dưới đây trồng nhiều cây sau đó được kết hợp để đưa ra một dự đoán đồng thuận duy nhất.



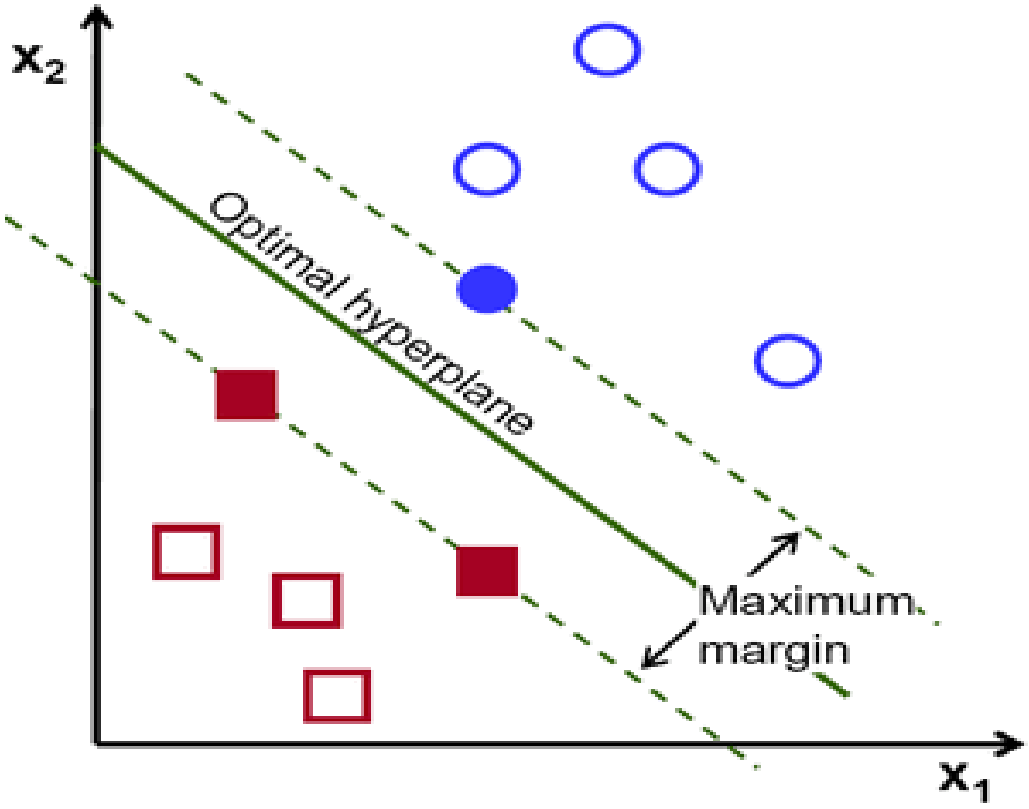
- Bagging là cách làm giảm phương sai dự đoán của bạn bằng cách tạo dữ liệu bổ sung để đào tạo từ bộ dữ liệu ban đầu của bạn bằng cách sử dụng kết hợp với các lần lặp lại để tạo ra nhiều mức độ giống nhau / kích thước như dữ liệu gốc của bạn. Bằng

cách tăng kích thước tập huấn luyện của bạn, bạn có thể cải thiện lực dự đoán mô hình, nhưng chỉ cần giảm phương sai, điều chỉnh hẹp dự đoán theo kết quả mong đợi.

- Boosting là một cách tiếp cận để tính toán đầu ra bằng cách sử dụng một số mô hình khác nhau và sau đó lấy trung bình kết quả bằng cách sử dụng phương pháp trung bình có trọng số. Bằng cách kết hợp các ưu điểm và cạm bẫy của các phương pháp này và thay đổi công thức tính trọng số của bạn, bạn có thể tạo ra một lực dự đoán tốt cho phạm vi dữ liệu đầu vào rộng hơn, sử dụng các mô hình được điều chỉnh hẹp khác nhau.
- Thuật toán random forest thực sự rất giống với bagging. Cũng ở đây, bạn vẽ các mẫu bootstrap ngẫu nhiên của tập huấn luyện của bạn. Tuy nhiên, ngoài các mẫu bootstrap, bạn cũng vẽ một tập hợp con các tính năng ngẫu nhiên để huấn luyện các cây riêng lẻ; trong bagging, bạn cung cấp cho mỗi cây đầy đủ các tính năng. Do lựa chọn tính năng ngẫu nhiên, bạn làm cho các cây độc lập với nhau hơn so với bagging thông thường, điều này thường mang lại hiệu suất dự đoán tốt hơn (do sự đánh đổi sai lệch phương sai tốt hơn) và nó cũng nhanh hơn, bởi vì mỗi cây chỉ học từ một tập hợp các tính năng.

9. Hỗ trợ Vector Machines

SVM là một kỹ thuật phân loại đưc liệt kê theo các mô hình học tập có giám sát trong Machine Learning. Theo thuật ngữ của giáo dân, nó liên quan đến việc tìm siêu phẳng (đường thẳng trong 2D, mặt phẳng trong 3D và siêu phẳng ở các chiều cao hơn. Chính thức hơn, một siêu phẳng là không gian con n-1 chiều của không gian n chiều) phân tách tốt nhất hai lớp điểm với lề tối đa. Về cơ bản, đây là một vấn đề tối ưu hóa bị ràng buộc trong đó lề được tối đa hóa theo ràng buộc mà nó phân loại hoàn hảo dữ liệu (lề cứng).

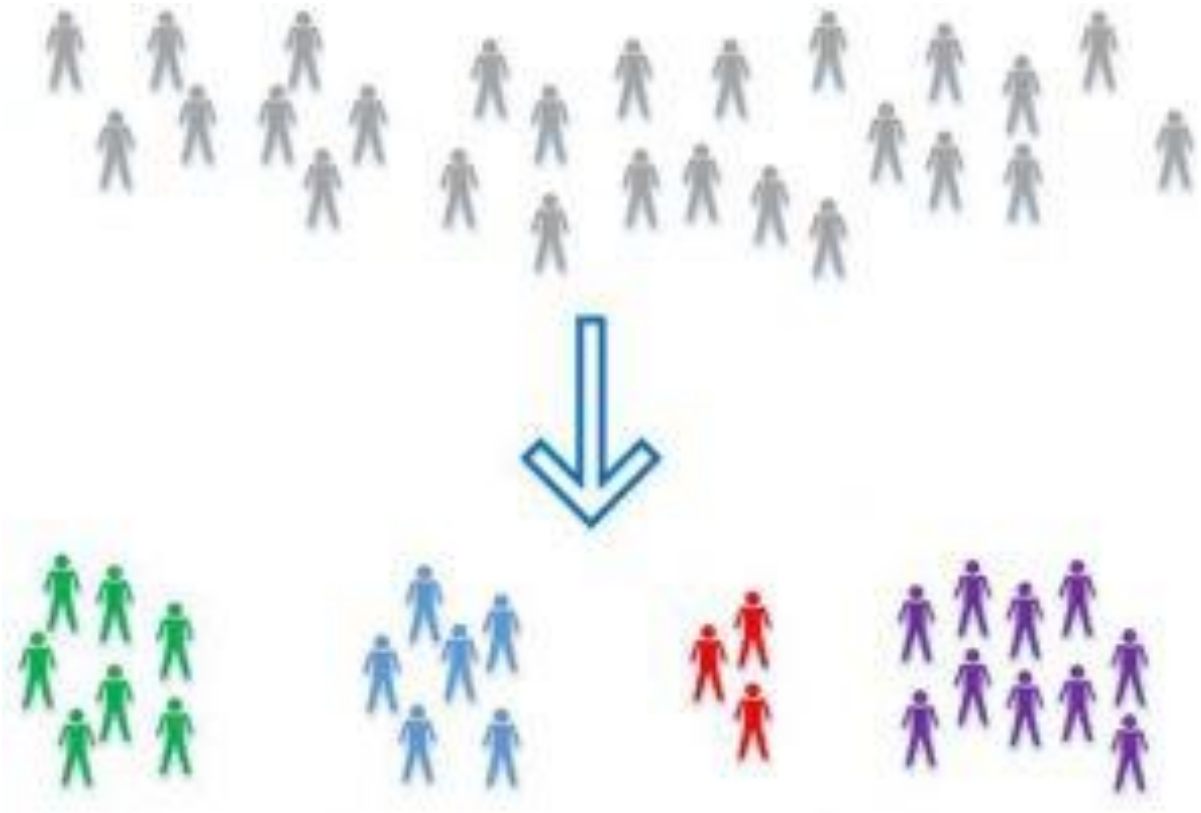


Khoa học dữ liệu

Các điểm dữ liệu mà loại hỗ trợ này ở hai bên được gọi là các vectơ hỗ trợ. Trong hình trên, vòng tròn màu xanh đầy và hai hình vuông đầy là các vectơ hỗ trợ. Đối với trường hợp hai lớp dữ liệu không thể phân tách tuyến tính, các điểm được chiếu đến một không gian bùng nổ (chiều không gian cao hơn) nơi có thể phân tách tuyến tính. Một vấn đề liên quan đến nhiều lớp có thể được chia thành nhiều vấn đề phân loại nhị phân một so với một hoặc một so với tất cả.

10. Học tập không giám sát

Cho đến nay, chúng ta chỉ thảo luận về các kỹ thuật học tập có giám sát, trong đó các nhóm được biết đến và kinh nghiệm cung cấp cho thuật toán là mối quan hệ giữa các thực thể thực tế và nhóm mà chúng thuộc về. Một tập hợp các kỹ thuật khác có thể được sử dụng khi các nhóm (danh mục) dữ liệu không được biết đến. Chúng được gọi là không giám sát vì nó còn lại trên thuật toán học tập để tìm ra các mẫu trong dữ liệu được cung cấp. Phân cụm là một ví dụ về học tập không giám sát, trong đó các tập dữ liệu khác nhau được nhóm thành các nhóm của các mục liên quan chặt chẽ. Dưới đây là danh sách các thuật toán học tập không giám sát được sử dụng rộng rãi nhất:



- Phân tích thành phần chính giúp tạo ra biểu diễn thứ nguyên thấp của bộ dữ liệu bằng cách xác định một tập hợp các tính năng tuyến tính có phương sai tối đa và không tương quan lẫn nhau. Kỹ thuật kích thước tuyến tính này có thể hữu ích trong việc tìm hiểu sự tương tác tiềm ẩn giữa biến trong một thiết lập không giám sát.
- k-Means clustering: phân vùng dữ liệu thành k cụm riêng biệt dựa trên khoảng cách đến tâm của cụm.
- Phân cụm theo phân cấp: xây dựng một hệ thống phân cấp nhiều cấp của các cụm bằng cách tạo một cây cụm.

Đây là bản tóm tắt cơ bản của một số kỹ thuật thống kê cơ bản trong cộng đồng IT có thể giúp người quản lý và điều hành chương trình khoa học dữ liệu hiểu rõ hơn về những gì đang chạy bên dưới đội ngũ khoa học dữ liệu của họ. Thực tế, một số nhóm khoa học dữ liệu hoàn toàn chạy các thuật toán thông qua các thư viện python và R. Hầu hết trong số họ thậm chí không phải suy nghĩ về toán học cơ bản. Tuy nhiên, việc có thể hiểu những điều cơ bản của phân tích thống kê giúp nhóm của bạn tiếp cận tốt hơn. Có cái nhìn sâu sắc vào các phần nhỏ nhất cho phép thao tác và trừu tượng dễ dàng hơn.

Thống kê là một lĩnh vực dựa trên toán học nhằm thu thập và diễn giải dữ liệu định lượng. Ngược lại, khoa học dữ liệu là một lĩnh vực đa ngành sử dụng các phương pháp, quy trình và hệ thống khoa học để trích xuất tri thức từ dữ liệu dưới nhiều hình thức khác nhau. Các nhà khoa học dữ liệu sử dụng các phương pháp từ nhiều lĩnh vực,

bao gồm cả thống kê. Tuy nhiên, các lĩnh vực này khác nhau về quy trình và những vấn đề mà chúng nghiên cứu.

Những thách thức các nhà khoa học dữ liệu phải đối mặt là gì?

Nhiều nguồn dữ liệu

Các loại ứng dụng và công cụ khác nhau tạo ra dữ liệu với nhiều định dạng khác nhau. Các nhà khoa học dữ liệu phải làm sạch và chuẩn bị dữ liệu để tạo sự nhất quán cho dữ liệu đó. Hoạt động này có thể rất nhàm chán và tốn thời gian.

Nắm rõ vấn đề kinh doanh

Các nhà khoa học dữ liệu phải làm việc với nhiều bên liên quan và các nhà quản lý doanh nghiệp để xác định vấn đề cần giải quyết. Điều này có thể rất khó khăn—đặc biệt là trong các công ty lớn với nhiều nhóm có các yêu cầu khác nhau.

Loại bỏ thiên kiến

Các công cụ máy học không hoàn toàn chính xác và do đó có thể tồn tại sự không chắc chắn hoặc thiên kiến. Thiên kiến là sự mất cân bằng trong dữ liệu đào tạo hoặc hành vi dự đoán của mô hình giữa các nhóm khác nhau, chẳng hạn như độ tuổi hoặc khung thu nhập. Ví dụ: nếu công cụ được đào tạo chủ yếu dựa trên dữ liệu từ các cá nhân trung niên thì công cụ này có thể kém chính xác hơn khi đưa ra các dự đoán liên quan đến những người trẻ tuổi và lớn tuổi hơn. Lĩnh vực máy học cung cấp cơ hội để giải quyết các thiên kiến bằng cách phát hiện và đo lường chúng trong dữ liệu và mô hình.

Trên đây là một số thuật toán phổ biến trong thống kê dữ liệu, phục vụ cho quá trình nghiên cứu dữ liệu phục vụ cho giai đoạn đầu để phát triển tri thức trong khoa học dữ liệu.