**Practical Econometrics and Data Science**

**Andrius Buteikis**

<http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/1-intro.html>

**Phân tích dữ liệu (Data Analysis)**. Tập trung vào các phương pháp thống kê và kinh tế lượng để phân tích dữ liệu. Sử dụng các phương pháp này, các mô hình dựa trên dữ liệu được tạo ra giúp hiểu rõ hơn và giải thích các liên kết giữa các hiệu ứng xã hội, kinh tế và tài chính khác nhau. Những mô hình này cũng giúp đưa ra các quyết định khác nhau, vì hiệu ứng của chúng có thể được đánh giá và định lượng dựa trên các mô hình được tạo. Một nhược điểm khác là các mô hình thường dễ diễn giải và có thể phân biệt các hiệu ứng cụ thể. **Kinh tế lượng (Econometrics**) là việc áp dụng các phương pháp toán học và thống kê vào dữ liệu kinh tế. Đây là một nhánh của kinh tế học sử dụng dữ liệu thực nghiệm để phân tích tính hợp lệ của các quan hệ kinh tế. Thông thường chúng ta có thể gọi phân tích dữ liệu là kinh tế lượng mà không mất tính tổng quát.

**Khoa học dữ liệu (Data Science)**. Không giống như phân tích dữ liệu, khoa học dữ liệu tập trung vào độ phức tạp của mô hình bằng cách sử dụng các thuật toán thống kê và học máy dựa trên số lượng lớn dữ liệu khác nhau (không nhất thiết là tài chính và kinh tế). Do sự phức tạp của các phương pháp này và khối lượng dữ liệu lớn có sẵn, các mô hình được đánh giá không phải lúc nào cũng có các giải thích rõ ràng cho các yếu tố riêng lẻ, so với các mô hình phân tích dữ liệu. Mặc dù điều này làm cho việc đánh giá mô hình trở nên khó khăn hơn, tuy nhiên, chúng cung cấp các dự đoán rất chính xác và được sử dụng thường xuyên khi làm việc với các tập dữ liệu lớn và phức tạp.

Nói cách khác - **phân tích dữ liệu** tập trung vào việc tìm kiếm và giải thích mối quan hệ nhân quả giữa các hiệu ứng khác nhau, trong khi **khoa học dữ liệu** tập trung vào việc dự đoán các kết quả có thể sử dụng dữ liệu có sẵn.

Một điểm khác biệt quan trọng - ngôn ngữ và thuật ngữ được sử dụng để mô tả các đặc điểm hoặc phương pháp nhất định. Điều này chủ yếu là do Khoa học dữ liệu được liên kết chặt chẽ với Khoa học máy tính. Dưới đây chúng tôi cung cấp một vài ví dụ:

Table 1.1: A table generated by the longtable package.

|  |  |  |
| --- | --- | --- |
| **Statistics** | **Data Science** | **Meaning** |
| estimation | learning | use data to estimate an unknown parameter (mean, variance, model coefficients, etc.) |
| classification | supervised learning | Predict a discrete value of YY using values of other variables X1X1, X2X2, …, XKXK |
| regression | supervised learning | Predict a continuous value of YY using values of other variables X1X1, X2X2, …, XKXK |
| clustering | unsupervised learning | Group the data based on some variables X1X1, X2X2, …, XKXK |
| data | training sample | (Y1,X1,1,X2,1,...,XK,1)(Y1,X1,1,X2,1,...,XK,1), …, (YN,X1,N,X2,N,...,XK,N)(YN,X1,N,X2,N,...,XK,N) |
| Covariates (factor) | features | variables X1,i, X2,i…, XK,i, which are collected for each observation  i=1,...,N |
| response | label | the variable of interest YY |
| classifier | hypothesis | a map ff of covariates to outcomes which describes their relationship, i.e. f:X→Yf:X→Y |
| hypothesis | - | a subset of a parameter space |

Phải nói rằng, có những phương pháp áp dụng cho cả phân tích dữ liệu và khoa học dữ liệu và trong một số trường hợp, ranh giới giữa nhà phân tích dữ liệu và nhà khoa học dữ liệu có thể trở nên mờ nhạt. Như vậy, cuốn sách này cung cấp một cái nhìn tổng quan thực tế về các phương pháp và ứng dụng khác nhau khi xử lý dữ liệu kinh tế với các chương được chọn dành riêng cho phương pháp giới thiệu cho khoa học dữ liệu. Mục tiêu là cung cấp một hộp công cụ rộng cho các phương thức cho các loại dữ liệu khác nhau. Các phương pháp này sau đó có thể được kết hợp theo nhiều cách khác nhau để sử dụng khi làm việc trên các ứng dụng thực tế.

**1.2 Kiểu dữ liệu thống kê Statistical Data types**

Phần này giới thiệu ngắn gọn về ba loại dữ liệu phổ biến nhất được sử dụng trong phân tích dữ liệu (cụ thể là phân tích thống kê / kinh tế lượng và mô hình hóa).

**1.2.1 Dữ liệu cắt ngang Cross-sectional Data**

Dữ liệu cắt ngang được thu thập trong một khoảng thời gian duy nhất và được đặc trưng bởi các đơn vị riêng lẻ - người dân, công ty, quốc gia, v.v ... Một số ví dụ bao gồm:

Lớp học sinh vào cuối học kỳ hiện tại;

Dữ liệu hộ gia đình của năm trước - chi tiêu cho thực phẩm, thất nghiệp, thu nhập, v.v.

Dữ liệu xe - tốc độ trung bình, mã lực, màu sắc, v.v.

**Với dữ liệu cắt ngang, việc sắp xếp dữ liệu không thành vấn đề**. Nói cách khác, chúng ta có thể sắp xếp thứ tự dữ liệu theo thứ tự tăng dần, giảm dần hoặc thậm chí ngẫu nhiên và điều này sẽ không ảnh hưởng đến kết quả mô hình hóa.

Mẫu dữ liệu sau đây cho tốc độ của ô tô và khoảng cách thực hiện để dừng lại. Dữ liệu được ghi lại trong 1920 1920.

**1.2.2 Dữ liệu chuỗi thời gian Time Series Data**

Dữ liệu được thu thập tại một số điểm cụ thể theo thời gian được gọi là dữ liệu chuỗi thời gian. Các ví dụ này bao gồm giá cổ phiếu, lãi suất, tỷ giá hối đoái cũng như giá sản phẩm, GDP, v.v. Dữ liệu chuỗi thời gian có thể được quan sát ở nhiều tần số khác nhau (hàng giờ, hàng ngày, hàng tuần, hàng tháng, hàng quý, hàng năm, v.v.).

**Không giống như dữ liệu cắt ngang**, thứ tự của dữ liệu rất quan trọng trong dữ liệu chuỗi thời gian. Mỗi điểm đại diện cho các giá trị tại các thời điểm cụ thể. Như vậy, dữ liệu chuỗi thời gian thường được trình bày theo thứ tự thời gian. Thay đổi thứ tự của dữ liệu bỏ qua tính thời gian của dữ liệu.

Mẫu dữ liệu sau đây là các quan sát hàng quý về Tiền, GDP và Lãi suất ở Canada, trong đó

m là nhật ký của cung tiền thật,

y là nhật ký GDP tính theo đô la năm 1992, theo mùa, được điều chỉnh;

p là nhật ký của mức giá và

r là kho bạc 3 tháng cho đến khi lãi suất.

**1.2.3 Dữ liệu bảng (hoặc dọc) Panel (or Longitudinal) data**

Dữ liệu bảng điều khiển kết hợp dữ liệu mặt cắt ngang và chuỗi thời gian: cùng một cá nhân (người, công ty, thành phố, v.v.) được quan sát tại một số thời điểm (ngày, năm, trước và sau khi điều trị, v.v.). Dữ liệu bảng điều khiển cho phép bạn kiểm soát các biến bạn không thể quan sát hoặc đo lường như:

yếu tố văn hóa (như quốc gia hoặc khu vực cụ thể);

sự khác biệt trong thực tiễn kinh doanh giữa các công ty;

Nếu chúng ta có cùng số lượng quan sát trong khoảng thời gian cho mỗi cá nhân, thì chúng ta có một bảng cân bằng.

Mẫu dữ liệu sau đây là của Dữ liệu Đầu tư Grunfeld - một bảng gồm 10 quan sát từ 1935 đến 1954 tại Hoa Kỳ, trong đó công ty là ID công ty, năm là ngày, đầu tư là tổng đầu tư, giá trị là giá trị của công ty và vốn là cổ phiếu của nhà máy và thiết bị.

**1.3 Các loại dữ liệu khác**

Dữ liệu hình ảnh

Các bộ dữ liệu khác nhau dựa trên dữ liệu hình ảnh (vị trí sản phẩm và sản phẩm, điều kiện ánh sáng, màu sắc của quảng cáo, v.v ... Có thể xem ví dụ về sử dụng thực tế hình ảnh / video và các loại dữ liệu khác trong chuỗi cửa hàng Amazon Go (liên kết khác).

Dữ liệu văn bản phân tích tình cảm

Còn được gọi là khai thác ý kiến. Được sử dụng trong việc trích xuất và phân tích đánh giá, trả lời khảo sát, diễn đàn trực tuyến và bài đăng trên phương tiện truyền thông xã hội. Nó được sử dụng để xác định thái độ (tích cực, tiêu cực, cảm xúc, v.v.) của đối tượng (khách hàng, người đánh giá, người trả lời, v.v.) của dữ liệu.

**1.4 Mô hình phù hợp và học tập mô hình Model Fitting and Model Learning**

Trong kinh tế lượng chuẩn, chúng tôi tận dụng dữ liệu có sẵn và phù hợp với một số loại mô hình cho dữ liệu. Sự phù hợp mô hình thường được thực hiện bằng cách ước tính một số tham số mô hình chưa biết.

Trong học máy, mô hình phù hợp “**fitting**” và thường được gọi là xây dựng mô hình **building** và ước lượng tham số của thường được gọi là học mô hình **learning**. Tùy thuộc vào loại mô hình được sử dụng và các nhiệm vụ trong tay, việc học mô hình **learning** có thể được chia thành một số loại khác nhau.

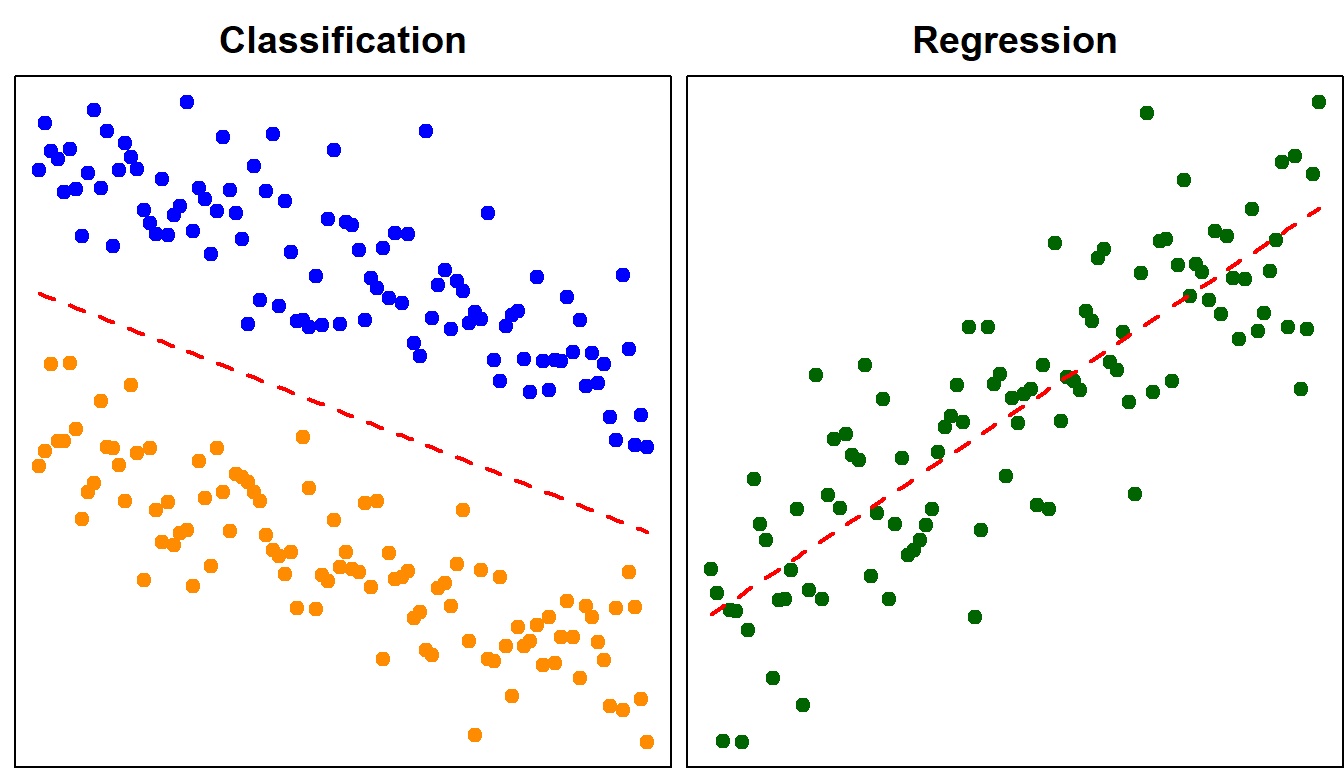
**1.4.1 Supervised Learning**

Học tập có giám sát liên quan đến dữ liệu được gắn nhãn **labeled data** - đó là dữ liệu (lịch sử), với một loại biến phụ thuộc (thường được ký hiệu là Y) mà chúng ta muốn mô hình hóa và chúng ta quan sát các giá trị. Ngoài ra, chúng tôi thường có số lượng biến K bổ sung (hy vọng đủ lớn) (thường được ký hiệu là Xi, i = 1, ..., K), mô tả một số tính năng **features** (thuật ngữ thường được sử dụng trong học máy) của dữ liệu. Trong kinh tế lượng, điều này thường được gọi là các biến giải thích **explanatory variables**.

Có hai lĩnh vực chính mà việc học có giám sát là hữu ích:

• Phân loại **Classification**, trong đó mục đích là để dự đoán một số giá trị phân loại **categorical** riêng biệt bằng cách sử dụng dữ liệu lịch sử để xác định lớp hoặc nhóm mà nó thuộc về. Số lượng các nhóm được biết trước từ dữ liệu lịch sử. Một số ví dụ bao gồm phân loại xem ai đó có khả năng mặc định cho khoản vay của họ, mua sản phẩm, bỏ công việc, chọn sản phẩm của đối thủ cạnh tranh, v.v. Một ví dụ khác bao gồm phân loại hình ảnh (ví dụ: loại động vật nào được hiển thị trong hình).

• Hồi quy **Regression**, trong đó mục đích là dự đoán một số giá trị liên tục Một số ví dụ bao gồm mô hình hóa: tiền lương, giá sản phẩm, số lượng khách hàng (còn được gọi là dữ liệu đếm **count data**), v.v.



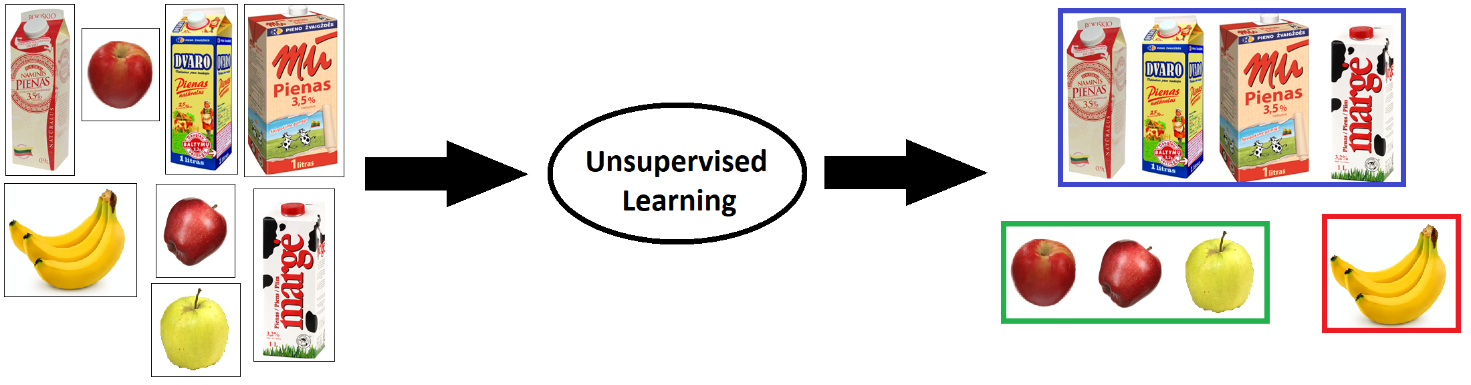
Cả hồi quy và phân loại là hai nhiệm vụ thường gặp nhất trong phân tích kinh tế lượng.

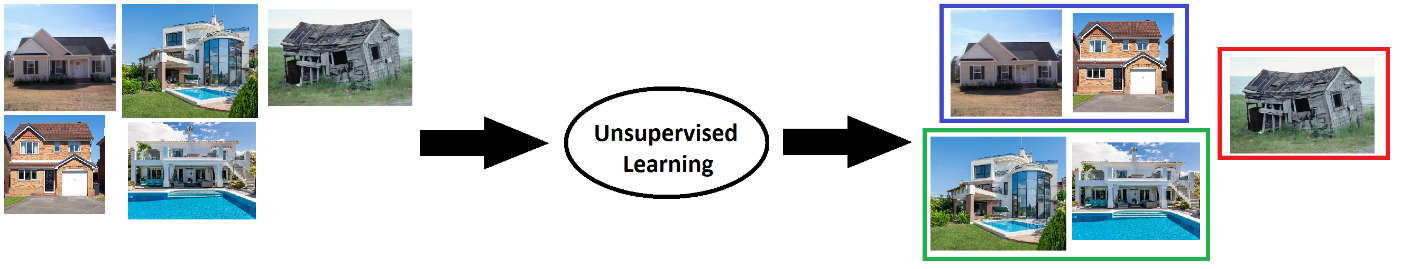
**1.4.2 Học tập không giám sát Unsupervised Learning**

Thật không may, dữ liệu trong thế giới thực thường không đầy đủ. Điều này có nghĩa là bây giờ chúng ta có thể biết số lượng danh mục thực sự trong dữ liệu. Trong những trường hợp như vậy, chúng tôi sử dụng việc học tập không giám sát - chúng tôi sử dụng các biến đầu vào có sẵn Xi, i = 1, ..., K. Mục tiêu của việc học tập không giám sát là mô hình hóa cấu trúc cơ bản của dữ liệu, tuy nhiên, không giống như học có giám sát, không có câu trả lời đúng “correct”, vì không có biến đầu ra tương ứng Y **no corresponding output variable Y**.

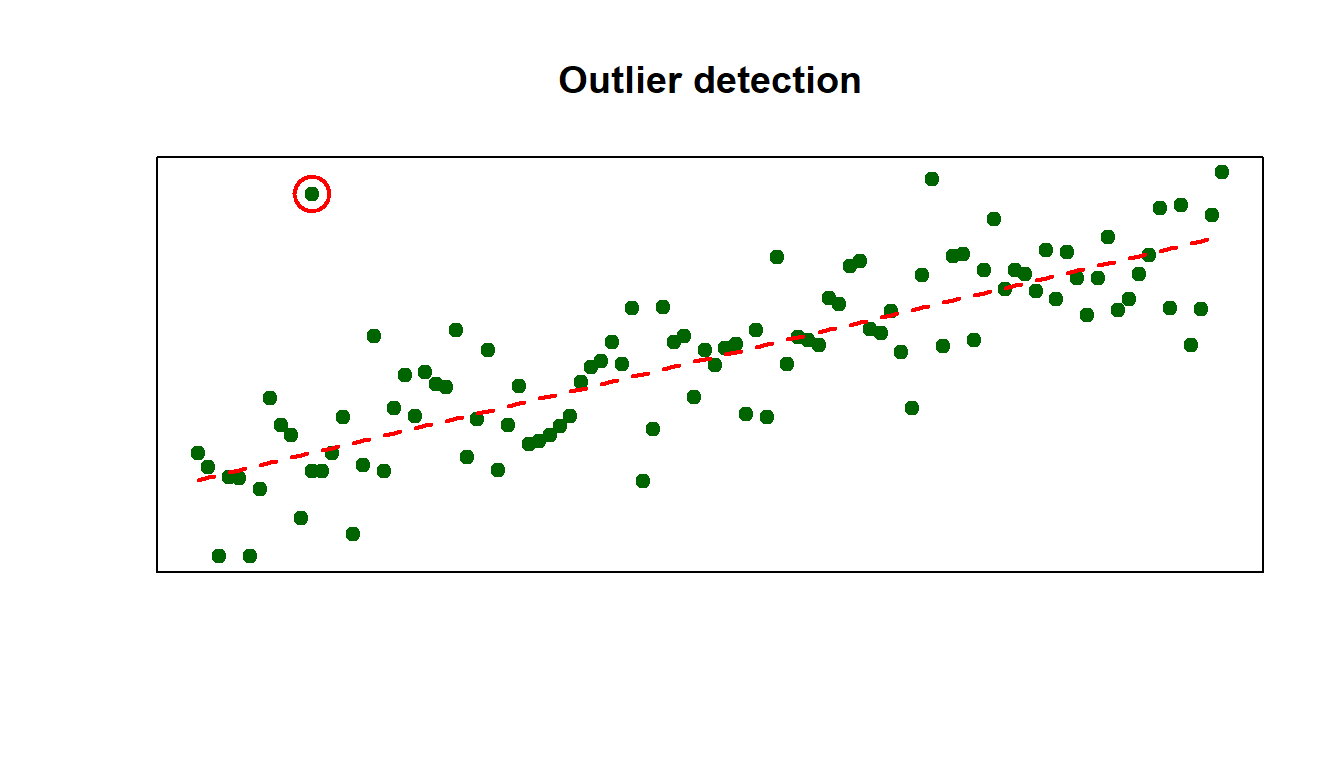
Các mô hình học tập không giám sát có thể được phân loại thành như sau:

• Phân cụm **Clustering** - nhóm dữ liệu có sẵn thành các cụm / danh mục / nhóm riêng biệt, dựa trên các biến đầu vào có sẵn. Không giống như phân loại, số lượng danh mục không xác định và không có biến, cho biết, nhóm dữ liệu lịch sử thuộc về nhóm nào.





• Phát hiện bất thường **Anomaly detection**- còn được gọi là Phát hiện ngoại lệ **Outlier detection**- liên quan đến việc phát hiện các quan sát đáng ngờ, khác biệt đáng kể so với dữ liệu còn lại. Theo định nghĩa, các ngoại lệ là các giá trị bất thường và bất ngờ, có thể do nhiều lý do (dị thường (ví dụ: có một sản phẩm cao cấp duy nhất trong số nhiều sản phẩm trung bình không có cách ghi nhãn như vậy); lỗi đo lường).



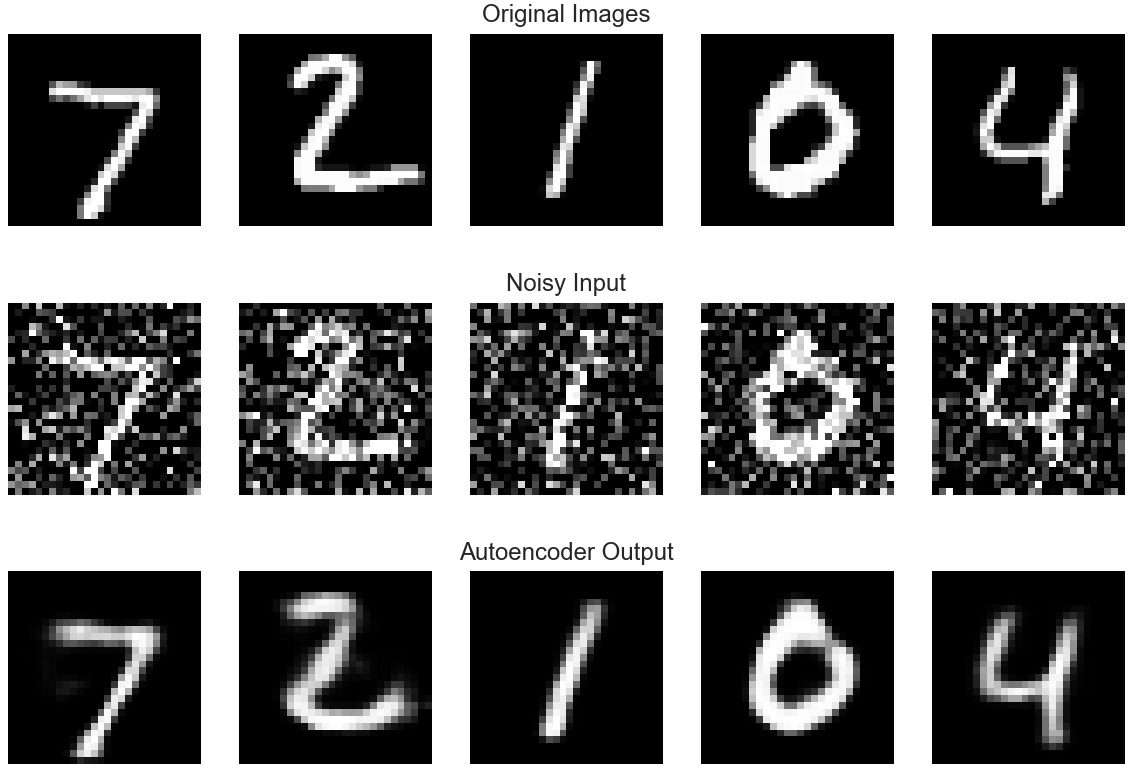
Các ngoại lệ có thể gây ra vấn đề nghiêm trọng trong phân tích thống kê và chúng thường gặp trong dữ liệu thực nghiệm. Trong một số trường hợp, dữ liệu ngoại lệ có thể bị loại bỏ, những lần khác có thể hữu ích để phân loại dữ liệu đó thành các trường hợp riêng biệt hoặc (trong trường hợp có lỗi đo lường) thay thế các giá trị này bằng các quan sát tương tự (ví dụ: bằng cách tìm dữ liệu với các biến giải thích tương tự Xi ).

• Hiệp hội **Association** - sử dụng các phương pháp học máy dựa trên quy tắc [rule-based machine learning](https://en.wikipedia.org/wiki/Rule-based_machine_learning) để xác định các mối quan hệ trong dữ liệu, trong đó các tính năng nhất định của mẫu dữ liệu tương quan với các tính năng khác. Nó được áp dụng phổ biến nhất trong phân tích giỏ thị trường để tìm hiểu hành vi mua hàng của khách hàng. Thuật toán học tập không giám sát được áp dụng sau đó có thể được sử dụng để bán lên, hoặc giới thiệu, các sản phẩm bổ sung, có thể bổ sung cho khách hàng mua. Ví dụ: nếu ai đó đang đặt hàng thực phẩm trực tuyến và giỏ mua hàng của họ chứa {sữa, cà rốt, hành tây}, thì mô hình có thể được sử dụng để giới thiệu một số sản phẩm bổ sung (-s), thường xuyên nhất mua trong các giỏ tương tự, ví dụ:

|  |  |
| --- | --- |
| ID | Giỏ hàng |
| 1 | {Bánh mì, Soda, Khoai tây} |
| 2 | {Rượu, Bánh mì} |
| 3 | {Rượu, Bánh mì, Khoai tây, Sữa} |
| 4 | {Soda, khoai tây, sữa} |
|  | Các mặt hàng thường xuyên mua Các mặt hàng được mua thường xuyên: {Sữa, Khoai tây} |
|  | Quy tắc kết hợp Quy tắc tổ chức: {Sữa} → {Khoai tây} {Sữa} → {Khoai tây} (nếu sữa được mua, thì khoai tây cũng thường được mua). |

Xem và ví dụ trong bài viết này trên [DataCamp](https://www.datacamp.com/community/tutorials/market-basket-analysis-r)  cũng như wikipedia.

• Bộ tự động mã hóa **Autoencoders** - lấy dữ liệu đầu vào, nén nó thành một mã, sau đó thử tạo lại dữ liệu đầu vào từ mã được tóm tắt đó. Mục đích là để tìm hiểu một đại diện (mã hóa encoding) cho một tập hợp dữ liệu bằng cách đào tạo mô hình để bỏ qua tín hiệu. Ví dụ: sử dụng cả phiên bản nhiễu và sạch của hình ảnh để huấn luyện mô hình, bộ mã hóa tự động sau đó có thể được sử dụng để tái tạo hình ảnh bằng cách loại bỏ nhiễu hình ảnh khỏi hình ảnh và cải thiện chất lượng hình ảnh.



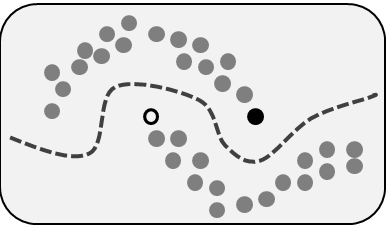
Xem ở đây [here](https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798?gi=78f58f143406) cho một ví dụ của hình ảnh trên.

Lưu ý rằng, không giống như học có giám sát, việc không có biến phụ thuộc sẽ gây khó khăn cho việc đo lường độ chính xác của mô hình không giám sát. Mặt khác, trong trường hợp dữ liệu không được gắn nhãn, lựa chọn duy nhất khác là xem xét thủ công từng điểm dữ liệu và quyết định nhãn theo cách thủ công (nếu chúng tôi đang làm việc với các danh mục) hoặc cố gắng loại bỏ nhiễu bằng tay (ví dụ: sử dụng một số quy tắc hiện có, hoặc kinh nghiệm trong ngành). Làm điều này bằng tay không chỉ tốn thời gian mà còn khó khăn.

Mặc dù trong những trường hợp như vậy, việc học tập không giám sát có thể chứng minh là nhanh hơn và nhất quán hơn so với việc thực hiện thủ công.

**1.4.3 Học bán giám sát Semi-supervised Learning**

Học bán giám sát là một phương tiện giữa học có giám sát và học không giám sát - nó sử dụng cả dữ liệu được dán nhãn và không nhãn cho đào tạo mô hình. Thông thường, một mẫu nhỏ của dữ liệu được dán nhãn (có thể tốn thời gian hoặc yêu cầu thuê thêm chuyên gia và do đó không khả thi cho toàn bộ dữ liệu) và một mẫu lớn dữ liệu chưa được gắn nhãn (có thể là dữ liệu thô chưa được gắn nhãn sẵn) được sử dụng.



Ví dụ trên được lấy từ wikipedia. Đường đứt nét hiển thị ranh giới quyết định cho hai loại mà chúng tôi có thể áp dụng nếu mẫu dữ liệu của chúng tôi có hai điểm được gắn nhãn (một vòng tròn trắng đen) cũng như tập hợp dữ liệu chưa được gắn nhãn (vòng tròn màu xám).

**1.4.4 Học tăng cường Reinforcement Learning**

Học tăng cường nhằm mục đích tối đa hóa một số khái niệm về phần thưởng tích lũy. Đây là một trong ba mô hình học máy cơ bản, bên cạnh việc học có giám sát và học không giám sát. Mục đích tổng thể của việc học tăng cường là dự đoán bước tiếp theo tốt nhất cần thực hiện để tối đa hóa phần thưởng cuối cùng.

Học tăng cường hoạt động theo nguyên tắc giống như trò chơi điện tử - hoàn thành một cấp độ và tăng điểm cao của bạn; mất tất cả cuộc sống của bạn và có được một trò chơi trên. Như vậy, trò chơi điện tử là một môi trường thử nghiệm phổ biến cho các loại mô hình này.

Xem Chơi Atari với Học tập tăng cường sâu [Playing Atari with Deep Reinforcement Learning](https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf) và các trò chơi Retro trong môi trường phòng tập thể dục để học tăng cường [Retro Games in Gym environments for reinforcement learning](https://github.com/openai/retro) (cũng ở [here](https://pypi.org/project/gym-super-mario-bros/)).

Một ví dụ khác mà việc học tăng cường có thể rất hữu ích là giao dịch chứng khoán. Để biết một số kết quả sơ bộ của các ứng dụng như vậy, hãy xem Phương pháp học tập củng cố sâu thực tế cho giao dịch chứng khoán và học tập củng cố sâu đối thủ trong quản lý danh mục đầu tư.

**1.4.5 Học sâu Deep Learning**

Còn được gọi là học tập phân cấp. Nó là một phần của một nhóm các phương pháp học máy rộng hơn dựa trên các mạng thần kinh nhân tạo sử dụng nhiều lớp để trích xuất dần các tính năng cao hơn từ dữ liệu đầu vào thô.

Không đi sâu vào chi tiết (các chương về các chủ đề này có thể được thêm vào trong tương lai) Không cần đi sâu vào chi tiết (các chương về các chủ đề này có thể được thêm vào trong tương lai), nhóm này bao gồm:

• Mạng nơ ron sâu [Deep neural networks](https://en.wikipedia.org/wiki/Deep_learning#Deep_neural_networks) (DNN) - mạng nơ ron nhân tạo [artificial neural network](https://en.wikipedia.org/wiki/Artificial_neural_network)  (ANN) có nhiều lớp giữa các lớp đầu vào và đầu ra, có thể mô hình các mối quan hệ phi tuyến tính phức tạp. Mạng nơ-ron nhân tạo là một mạng gồm các phần tử đơn giản gọi là nơ-ron nhân tạo, nhận đầu vào, thay đổi trạng thái bên trong (kích hoạt activation) theo đầu vào đó và tạo ra đầu ra tùy thuộc vào đầu vào và kích hoạt.

• Mạng niềm tin sâu sắc [Deep belief networks](https://en.wikipedia.org/wiki/Deep_belief_network) (DBN) - một thành phần của các mạng đơn giản, không bị giám sát, chẳng hạn như bộ thu phát tự động, trong đó mỗi lớp ẩn mạng con phụ đóng vai trò là lớp hiển thị cho lớp kế tiếp. Thành phần này dẫn đến một quy trình đào tạo nhanh, không giám sát từng lớp. Việc quan sát rằng các DBN có thể được đào tạo một cách tham lam, từng lớp một, dẫn đến một trong những thuật toán học sâu hiệu quả đầu tiên.

• Mạng thần kinh tái phát [Recurrent neural networks](https://en.wikipedia.org/wiki/Recurrent_neural_network)  (RNN) - một lớp mạng thần kinh nhân tạo nơi các kết nối giữa các nút tạo thành một biểu đồ có hướng dọc theo một chuỗi thời gian. Một số ứng dụng bao gồm nhận dạng chữ viết tay, nhận dạng giọng nói.

• Mạng nơ ron kết hợp [Convolutional neural networks](https://en.wikipedia.org/wiki/Convolutional_neural_network)  (CNN) - phiên bản chính quy của các tri giác đa lớp multilayer perceptrons. Các tri giác đa lớp thường đề cập đến các mạng được kết nối đầy đủ, nghĩa là, mỗi nơ-ron trong một lớp được kết nối với tất cả các nơ-ron ở lớp tiếp theo. Các mạng được kết nối đầy đủ, các mạng của các mạng này khiến chúng dễ bị quá tải dữ liệu. Các CNN có một cách tiếp cận khác đối với việc chính quy hóa: họ tận dụng mô hình phân cấp trong dữ liệu và lắp ráp các mẫu phức tạp hơn bằng cách sử dụng các mẫu nhỏ hơn và đơn giản hơn. Do đó, trên quy mô kết nối và độ phức tạp, CNN ở mức cực thấp. Áp dụng phổ biến nhất để phân tích hình ảnh trực quan.

• Các mạng đối nghịch tạo ra [Generative adversarial networks](https://en.wikipedia.org/wiki/Generative_adversarial_network) (GAN) là các kiến trúc mạng nơ-ron sâu bao gồm hai lưới, rỗ một lưới với nhau (do đó, đối thủ trực tuyến, xem bài báo). Một bài báo tương tự về Few-Shot Adversarial Learning về các mô hình đầu nói chuyện thần kinh thực tế cho phép tạo hiệu ứng hoạt hình trên khuôn mặt của các hình ảnh tĩnh.

**1.4.6 Học tập không bắn Zero-shot Learning**

Học tập không bắn mở rộng việc học có giám sát trong trường hợp, trong đó dữ liệu đào tạo không có sẵn cho một số lớp. Các thuật toán sử dụng một số kết nối giữa các thông tin có sẵn và các lớp không nhìn thấy. Xem thêm Học tập Featur (hoặc Đại diện) [Featur (or Representation) Learning](https://en.wikipedia.org/wiki/Feature_learning).

Ví dụ: nếu bạn đã đọc mô tả rất chi tiết về ngựa vằn, bạn có thể biết ngựa vằn là gì trong một bức ảnh trong lần đầu tiên bạn nhìn thấy nó. Điều này có thể được coi là một nhiệm vụ xử lý ngôn ngữ tự nhiên [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) task, sử dụng các từ nhúng [word embeddings](https://en.wikipedia.org/wiki/Word_embedding) - trong đó các từ hoặc cụm từ từ vựng được ánh xạ tới các vectơ của các số thực. Những từ có nghĩa tương tự sẽ có từ nhúng tương tự. Tiếp tục ví dụ, nếu dữ liệu huấn luyện không có hình ảnh ngựa vằn, nhưng có nhiều hình ảnh khác nhau về động vật bị tước (hổ), động vật giống ngựa (ngựa, lừa, ngựa non), động vật đen / trắng (gấu trúc, chim cánh cụt, v.v.) , sau đó các tính năng của chúng (có thể rút ra các tính năng của chúng, có thể được trích xuất và các từ nhúng giống như ngựa, có thể được trích xuất và có thể tạo ra các từ nhúng. Sau đó, chúng tôi có thể mô tả những gì một con ngựa vằn trông giống như và tạo ra một từ trong từ điển thích hợp bằng cách sử dụng các tính năng có sẵn. Cuối cùng, chúng ta có thể đọc một hình ảnh của một con ngựa vằn, trích xuất các tính năng của hình ảnh mới này, tạo ra một từ nhúng và so sánh với các từ nhúng gần nhất từ ​​từ điển của chúng ta, có thể sẽ là một con ngựa vằn.

Kỹ thuật này, được gọi là học tập không bắn zero-shot learning (ZSL), vẫn còn ở giai đoạn sơ khai và hiện đang là một chủ đề nghiên cứu tích cực.