TON DUC THANG UNIVERSITY

UNI]]

MINING OF

MASSIVE

DATASETS

NGUYEN HAI LONG

**MỤC LỤC**

[Chương I: Tìm hiểu về Machine Learning 1](#_Toc70512240)

[**1.** **Giới thiệu** 1](#_Toc70512241)

[***1.1*** ***Trí tuệ nhân tạo (Artificial Intelligence/AI)*** 1](#_Toc70512242)

[***1.2*** ***Big Data*** 2](#_Toc70512243)

[***1.3*** ***Dự đoán tương lai*** 2](#_Toc70512244)

[**2.** **Phân loại thuật toán Machine learning** 2](#_Toc70512245)

[***2.1*** ***Học có giám sát (Supervised learning)*** 3](#_Toc70512246)

[***2.1.1*** ***Phân loại (Classification)*** 3](#_Toc70512247)

[***2.1.2*** ***Hồi quy (Regression)*** 3](#_Toc70512248)

[***2.2*** ***Học không giám sát (Unsupervised learning)*** 4](#_Toc70512249)

[**3.** **Các thuật toán Machine learning** 4](#_Toc70512250)

[***3.1*** ***Support Vector Machines*** 4](#_Toc70512251)

[***3.2*** ***Mô hình xác suất (Probabilistic Models)*** 4](#_Toc70512252)

[***3.3*** ***Học sâu(Deep learning)*** 5](#_Toc70512253)

[**4.** **Lưu ý quan trọng trong Machine learning** 5](#_Toc70512254)

[***4.1*** ***Feature engineering*** 5](#_Toc70512255)

[***4.1.1*** ***Trích xuất đặc trưng (Feature extraction)*** 5](#_Toc70512256)

[***4.1.2*** ***Trích chọn đặc trưng (Feature selection)*** 6](#_Toc70512257)

[***4.1.3*** ***Dữ liệu huấn luyện (Training data)*** 6](#_Toc70512258)

[***4.1.4*** ***Dữ liệu kiểm thử và các độ đo*** 6](#_Toc70512259)

[***4.1.5*** ***Hiệu suất*** 7](#_Toc70512260)

[**5.** **Ứng dụng của Machine learning trong thực tiễn** 7](#_Toc70512261)

[***5.1*** ***Tự động phân loại*** 7](#_Toc70512262)

[***5.2*** ***Ứng dụng trong các mạng xã hội*** 7](#_Toc70512263)

[***5.3*** ***Nhận diện hình ảnh*** 8](#_Toc70512264)

[Tài liệu tham khảo 9](#_Toc70512265)

# Chương I: Tìm hiểu về Machine Learning

1. **Giới thiệu**

Machine Learning (Học Máy hoặc Máy Học) là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể. Trong những năm gần đây machine learning thu hút rất nhiều người quan tâm, từ các công ty lớn như Google, Facebook, Microsoft đến các công ty khởi nghiệp đều đầu tư vào machine learning. Hàng loạt các ứng dụng sử dụng machine learning ra đời trên mọi lĩnh vực của cuộc sống, tiêu biểu như xe tự vận hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, ngoài ra còn nhiều ứng dụng khác bao trùm từ khoa học máy tính đến những ngành ít liên quan hơn như vật lý, hóa học, y học, chính trị.

Machine learning có 3 mối quan hệ sau:[3]

* Machine learning và trí tuệ nhân tạo.
* Machine learning và Big data.
* Machine learning và dự đoán tương lai.

Sau đây mình sẽ đi sâu tìm hiểu các mối quan hệ này.

* 1. ***Trí tuệ nhân tạo (Artificial Intelligence/AI)***

Trong khoa học máy tính, trí tuệ nhân tạo hay AI đôi khi được gọi là trí thông minh nhân tạo, là trí thông minh được thể hiện bằng máy móc, trái ngược với trí thông minh tự nhiên của con người. Thông thường, thuật ngữ "trí tuệ nhân tạo" thường được sử dụng để mô tả các máy móc (hoặc máy tính) có khả năng bắt chước các chức năng "nhận thức" mà con người thường phải liên kết với tâm trí, như "học tập" và "giải quyết vấn đề".[1]

Theo hiểu biết chung, AI là một ngành khoa học được sinh ra với mục đích làm cho máy tính có được trí thông minh. Tuy nhiên mục tiêu này vẫn khá mơ hồ vì không phải ai cũng đồng ý với một định nghĩa thống nhất về trí thông minh. Các nhà khoa học phải định nghĩa một số mục tiêu cụ thể hơn, một trong số đó là việc làm cho máy tính lừa được Turing Test. Turing Test được tạo ra bởi Alan Turing (1912-1954), người được xem là cha đẻ của ngành khoa học máy tính hiện đại, nhằm phân biệt xem người đối diện có phải là người hay không.

AI thể hiện một mục tiêu của con người. Machine learning là một phương tiện được kỳ vọng sẽ giúp con người đạt được mục tiêu đó. Machine learning và AI có mối quan hệ chặt chẽ với nhau nhưng không hẳn là trùng khớp vì một bên là mục tiêu (AI), một bên là phương tiện (machine learning). Chinh phục AI mặc dù vẫn là mục đích tối thượng của machine learning, nhưng hiện tại machine learning chỉ tập trung vào hai mục tiêu:[3]

* Làm cho máy tính có những khả năng nhận thức cơ bản của con người như nghe, nhìn, hiểu được ngôn ngữ, giải toán, lập trình, …
* Hỗ trợ con người trong việc xử lý một khối lượng thông tin khổng lồ mà chúng ta phải đối mặt hàng ngày, hay còn gọi là Big Data.
  1. ***Big Data***

Dữ liệu lớn (Big data) là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Những tập dữ liệu lớn này có thể bao gồm các dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc.[4]

**Big Data** thực chất không phải là một ngành khoa học chính thống. Đó là một cụm từ tung hô để ám chỉ thời kì bùng nổ của dữ liệu hiện nay. Nó cũng không khác gì với những cụm từ như "cách mạng công nghiệp", "kỉ nguyên phần mềm". Big Data là một kết quả tất yếu của việc mạng Internet ngày càng phát triển. Với sự ra đời của các mạng xã hội nhưng Facebook, Instagram, Twitter đã khiến cho nhu cầu chia sẻ thông tin của con người tăng trưởng một cách chóng mặt. Youtube cũng có thể được xem là một mạng xã hội, nơi mọi người chia sẻ video và comment về nội dung của video. Machine Learning và Big Data có một quan hệ tương hỗ với nhau: machine learning phát triển hơn nhờ sự gia tăng của khối lượng dữ liệu của Big Data; ngược lại, giá trị của Big Data phụ thuộc vào khả năng khai thác tri thức từ dữ liệu của machine learning.[3]

* 1. ***Dự đoán tương lai***

Khi nói đến machine learning thì phải nói đến “dự đoán”, từ việc dự đoán nhãn phân loại đến dự đoán hành động cần thực hiện trong bước tiếp theo. Machine learning có thể dự đoán được tương lai chỉ khi tương lai có mối liên hệ mật thiết với hiện tại.

Ví dụ về “dự đoán tương lai”: giả sử bạn được đưa cho một đồng xu, rồi được yêu cầu tung đồng xu một số lần. Vấn đề đặt ra là: dựa vào những lần tung đồng xu đó, bạn hãy tiên đoán ra kết quả lần tung tiếp theo. Chỉ cần dựa vào tỉ lệ sấp/ngửa của những lần tung trước đó, bạn có thể đưa ra một dự đoán khá tốt. Nhưng nếu mỗi lần tung, người ta đưa cho bạn một đồng xu khác nhau thì mọi chuyện sẽ hoàn toàn khác. Các đồng xu khác nhau có xác suất sấp ngửa khác nhau. Lúc này việc dự đoán gần như không thể vì xác suất sấp ngửa của lần tung sau không hề liên quan gì đến lần tung trước. Điều tương tự cũng xảy ra với việc dự đoán tương lai bằng machine learning, nếu ta xem như mỗi ngày có một "đồng xu" được tung ra để xem một sự kiện có diễn ra hay không. Nếu "đồng xu" của ngày mai được chọn một cách tùy ý không theo phân bố nào cả thì machine learning sẽ thất bại. Tuy nhiên luôn có một hoạt động theo những quy luật nhất định và machine learning có thể nhận ra được những quy luật đó bằng cách tổng quát hóa những kinh nghiệm, những gì đã được học từ dữ liệu.[3]

1. **Phân loại thuật toán Machine learning**

Machine learning có 2 thuật toán cơ bản: Học có giám sát (Supervised learning) và Học không giám sát (Unsupervised learning). Sự khác nhau lớn nhất giữa 2 thuật toán này là cách mà chúng ta cung cấp tập dữ liệu huấn luyện cho mô hình, cách thuật toán sử dụng dữ liệu và loại vấn đề mà chúng giải quyết.

* 1. ***Học có giám sát (Supervised learning)***

Trong Supervised learning thuật toán machine learning thực hiện quá trình chuyển dữ liệu đầu vào thành đầu ra mong muốn. Đầu vào được gọi là input, đầu ra còn được gọi là label (nhãn). Tập dữ liệu huấn luyện được gọi là training set. Mỗi mẫu dữ liệu trong tập huấn luyện được gọi là training example.[2]

Thuật toán machine learning cần học cách để chuyển đổi mỗi input (đầu vào) thành label (kết quả) tương ứng. Do vậy, mọi mẫu dữ liệu huấn luyện trong tập dữ liệu huấn luyện cần phải biết trước label của nó. Ví dụ: hệ thống nhận vào các reviews về một sản phẩm đồ ăn hay một chương trình nào đó, cần xác định các reviews đó có nội dung tích cực hay tiêu cực. Giả sử trường hợp này chúng ta cần xác định nhãn cho mỗi review: đầu vào của tập dữ liệu huấn luyện sẽ là các review về món ăn hay chương trình nào đó, và mỗi review đó đã được chỉ định rõ nội dung là tích cực hay tiêu cực.

Tùy thuộc vào loại đầu ra mong muốn, chúng ta tiếp tục chia nhỏ supervised learning gồm: phân loại (Classification) và hồi quy (Regression).[2]

* + 1. ***Phân loại (Classification)***

Khi đầu ra mong muốn của chúng ta là một tập hữu hạn và rời rạc. Khi đó bài toán của chúng ta được gọi là bài toán phân loại. Ví dụ ở phía trên có thể xếp vào bài toán phân loại, các đầu ra mong muốn của chúng ta là: tích cực, tiêu cực và trung tính. Khi đó, tập dữ liệu huấn luyện sẽ như thế này:[2]

|  |  |
| --- | --- |
| **Text** | **Label** |
| “Chương trình này hay phết, nó giúp người xem vui vẻ và có thêm kiến thức nhưng mỗi tuần chỉ chiếu một tập.” | Tích cực |
| “Chờ lâu quá thể đáng.” | Tiêu cực |

* + 1. ***Hồi quy (Regression)***

Khi đầu ra mong muốn là một dải các giá trị liên tục, chẳng hạn như giá trị xác suất, khi đó bài toán sẽ thuộc loại hồi quy. Ví dụ một hệ thống nhận vào hình ảnh/thông tin của một người. Đánh giá số điểm đo khả năng người đó sẽ trả một khoản vay tín dụng. Đây là một bài toán dự đoán điểm tin cậy trong [0; 1], nó thể hiện xác suất một người sẽ trả các khoản vay của mình. Khi đó, dữ liệu sẽ giống như sau:[2]

|  |  |  |  |
| --- | --- | --- | --- |
| **Nghề nghiệp** | **Thu nhập** | **Tuổi** | **Điểm tin cậy** |
| Lập trình viên | > 20 triệu | 25 | 0.85 |
| Sinh viên | < 5 triệu | 18 | 0.3 |

Supervised learning là thuật toán phổ biến nhất trong các thuật toán machine learning. Hạn chế khi sử dụng thuật toán này là chúng ta cần cung cấp dữ liệu có gán nhãn. Trong nhiều trường hợp, để có được dữ liệu gán nhãn này rất tốn rất nhiều chi phí. Chẳng hạn trong ví dụ về phân loại, nếu ta cần 10.000 review có nhãn (tích cực, tiêu cực, trung tính) để huấn luyện mô hình thì việc này sẽ cần con người đọc từng review và gán nhãn thủ công. Điều này rất tốn thời gian và công sức. Đây cũng là một rào cản của machine learning: xây dựng các tập dữ liệu gán nhãn chất lượng.

* 1. ***Học không giám sát (Unsupervised learning)***

Unsupervised learning cũng là một nhánh trong machine learning. Các mẫu dữ liệu trong học không giám sát chỉ cần input (đầu vào) mà không cần label (đầu ra). Nó được sử dụng nhiều trong việc khám phá cấu trúc và mối quan hệ của dữ liệu. Một thuật toán điển hình là bài toán phân cụm (clustering algorithm), nó học cách để tìm các mẫu dữ liệu tương tự nhau và nhóm vào thành các cụm (cluster). Một số thuật toán phân cụm như K-means học cách phân cụm chỉ học từ tập dữ liệu đầu vào.[2]

1. **Các thuật toán Machine learning**

Để thực hiện chuyển đổi từ input thành output mong muốn, chúng ta có thể sử dụng các mô hình khác nhau. Machine learning không phải là một loại thuật toán duy nhất, có thể bạn đã nghe tới Support vector machine (SVM), Naive Bayes, Cây quyết định (Decision Trees) hay Học sâu (Deep learning). Các thuật toán này đều cố gắng giải quyết một bài toán: học cách chuyển đổi mọi input thành output chính xác của mà nó thuộc về.[2]

Những thuật toán machine learning này sử dụng các mô hình/kỹ thuật khác nhau để thực hiện quá trình học tập và thể hiện kiến thức về những gì nó được học. Có một nguyên tắc chung: các thuật toán machine learning luôn cố gắng khái quát hóa. Nghĩa là, nó sẽ tìm và giải thích theo cách đơn giản nhất. Nguyên tắc đó được biết đến là Occam’s razor - là một lý thuyết triết học nổi tiếng của nhà triết học người Anh William xứ Ockham, nó được ông đưa ra vào năm 1324. William đã viết rằng lời giải thích đơn giản nhất thường là lời giải thích xác đáng nhất, hoặc “Điều gì có thể được giải thích bằng ít giả thuyết hơn thì lại được giải thích một cách vô ích bằng nhiều giả định hơn”. Tức là nếu một vấn đề khoa học được giải thích bằng ít giả thuyết nhất thì sẽ đúng đắn nhất.[5]

Có rất nhiều thuật toán machine learning khác nhau và đây là ba thuật toán trong số đó:

* 1. ***Support Vector Machines***

Đây là một thuật toán cố gắng xây dựng một siêu mặt phẳng trong không gian nhiều chiều để phân biệt các đối tượng ở các lớp khác nhau sao cho khoảng cách giữa 2 đối tượng khác label gần nhau nhất có khoảng cách cực đại. Ý tưởng của thuật toán cực kỳ đơn giản, nhưng mô hình này lại rất phức tạp và có hiệu quả. Thực tế, ở một số bài toán Support Vector Machines là một mô hình machine learning cho hiệu quả tốt nhất.[2]

* 1. ***Mô hình xác suất (Probabilistic Models)***

Đây là mô hình cố gắng giải quyết bài toán bằng phân bổ xác suất. Một thuật toán phổ biến nhất là phân loại Naive Bayes, nó sử dụng lý thuyết Bayes và giả thiết các đặc trưng là độc lập. Điểm mạnh của mô hình xác suất là đơn giản nhưng hiệu quả. Đầu ra của nó không chỉ là label mà còn đi kèm xác suất thể hiện độ chính xác cho kết quả đó.[2]

* 1. ***Học sâu(Deep learning)***

Hiện đang là xu hướng trong machine learning dựa trên các mô hình mạng nơ ron nhân tạo (Artificial Neural Networks). Mạng nơ ron có cách tiếp cận kết nối và sử dụng ý tưởng theo cách bộ não con người làm việc. Chúng bao gồm số lượng lớn các nơ ron liên kết với nhau được tổ chức thành các lớp (layers). Học sâu liên tục được phát triển với các cấu trúc mới sâu hơn, nó không chỉ cố gắng học mà còn xây dựng các cấu trúc biểu diễn các đặc trưng quan trọng một cách tự động.[2]

1. **Lưu ý quan trọng trong Machine learning**
   1. ***Feature engineering***

Feature engineering là quá trình chúng ta thực hiện trích xuất và trích chọn các đặc trưng (thuộc tính) quan trọng từ dữ liệu thô để sử dụng làm đại diện cho các mẫu dữ liệu huấn luyện. Một tập dữ liệu huấn luyện có thể có rất nhiều thuộc tính, nhưng không phải cái nào cũng cần thiết và quan trọng. Feature engineering là kỹ thuật giúp loại bỏ các thuộc tính dư thừa, làm đơn giản hóa quá trình biểu diễn dữ liệu nhưng không làm ảnh hưởng tới kết quả cuối cùng. Nếu không có bước này, mô hình học sẽ cực kỳ phức tạp và thậm chí còn giảm độ chính xác khi có những thuộc tính nhiễu.[2]

Feature engineering là một bước quan trọng trong machine learning (trừ khi dữ liệu của bạn đã sạch sẽ hoặc là bài toán chưa đủ lớn). Nếu bạn không thực hiện bước này, kết quả sẽ có thể rất tệ, cho dù bạn có dùng một thuật toán tốt nhất đi nữa.

* + 1. ***Trích xuất đặc trưng (Feature extraction)***

Để nạp dữ liệu huấn luyện vào mô hình học máy, bạn cần phải đưa dữ liệu thô về cấu trúc nào đó mà thuật toán có thể “hiểu”, công việc này được gọi là trích xuất đặc trưng. Thông dụng nhất, chúng ta sẽ chuyển dữ liệu thô về dữ liệu số là vector của các đặc trưng.[2]

Ví dụ: làm sao ta có thể truyền vào thuật toán machine learning một hình ảnh?

Một cách đơn giản là chuyển ảnh đó thành một vector, mỗi phần tử trong vector đó tương ứng là giá trị màu xám của từng pixel trong ảnh. Khi đó, mỗi đặc trưng/thành phần sẽ là một giá trị số từ 0 – 255, trong đó 0 là màu đen, 255 là trắng và 1 – 254 là các giá trị mức độ xám. [2]

Giải pháp trên có thể cho kết quả, nhưng kết quả sẽ cải thiện hơn nếu ta cung cấp nhiều đặc trưng có giá trị hơn:

* Hình ảnh đó có chứa ảnh khuôn mặt người?
* Màu da là gì?
* Màu mắt là gì?
* Khuôn mặt có tóc hay không?

Đó là những đặc trưng ở mức cao hơn. Những đặc trưng này có giá trị hơn nhiều so với đặc trưng màu của các pixel ảnh. Các đặc trưng này có thể là kết quả của một thuật toán machine learning khác. Cung cấp các đặc trưng mức cao hơn giúp mô hình học máy của ta học tập và đưa ra dự đoán chính xác hơn. Nếu chúng ta cung cấp các đặc trưng tốt hơn:

* Thuật toán sẽ có khả năng cung cấp những kết quả chính xác hơn.
* Có thể sẽ cần ít dữ liệu huấn luyện hơn.
* Có thể giảm đáng kể thời gian huấn luyện mô hình.
  + 1. ***Trích chọn đặc trưng (Feature selection)***

Đôi khi, các đặc trưng chúng ta cung cấp cho thuật toán machine learning có thể vô dụng. Chẳng hạn với bài toán phân loại review, chúng ta cung cấp chiều dài của review, ngày tạo và người tạo review đó,… chúng có thể hữu ích hoặc không. Sẽ rất hữu ích nếu có phương pháp tự động phát hiện điều này. Đó là trích chọn đặc trưng, thuật toán này sử dụng kỹ thuật đánh trọng số cho từng đặc trưng. Và chỉ lựa chọn các đặc trưng có trọng số cao.[2]

Một lưu ý nữa: cái gì nhiều quá cũng không tốt, tránh sử dụng số lượng đặc trưng quá lớn. Bạn có thể bị cám dỗ và thêm tất cả các đặc trưng có thể hữu ích. Nhưng đó không phải cách tốt, thêm đặc trưng cũng đồng nghĩa không gian bộ nhớ tăng lên, làm cho dữ liệu cách xa nhau hơn. Vấn đề phổ biến này được biết đến với cái tên curse of dimensionality. Khi số lượng mẫu huấn luyện tăng theo cấp số nhân, đó thực sự là một vấn đề.[2]

* + 1. ***Dữ liệu huấn luyện (Training data)***

Thuật toán machine learning cần truyền vào các mẫu dữ liệu huấn luyện. Tùy thuộc vào bài toán cần giải quyết, chúng ta có thể cần vài trăm, vài nghìn, triệu hoặc hàng tỷ mẫu dữ liệu huấn luyện. Cân nhắc, cung cấp các mẫu huấn luyện tốt là rất quan trọng, nếu bạn đưa vào các mẫu dữ liệu sai, cơ hội để có kết quả tốt sẽ giảm đi.[2]

Thu thập một số lượng lớn dữ liệu liệu có chất lượng tốt để huấn luyện các mô hình học máy thường tốn nhiều chi phí. Trừ khi bạn đã có sẵn dữ liệu gán nhãn rồi. Bạn có thể thực hiện công việc này thủ công bằng sức người hay một số công cụ hỗ trợ tăng tốc độ gán nhãn có thể giúp bạn. Nguyên tắc chung về dữ liệu huấn luyện là: chất lượng của dữ liệu huấn luyện của bạn càng tốt, khả năng bạn thu được kết quả tốt sẽ cao hơn.

* + 1. ***Dữ liệu kiểm thử và các độ đo***

Sau khi chúng ta đào tạo một thuật toán máy học, chúng ta cần kiểm tra hiệu quả của nó. Đây là một bước đặc biệt quan trọng, nếu không bạn sẽ không thể biết mô hình của mình có học được gì hay không. Cách thường dùng là kiểm thử và đo độ chính xác.

Ý tưởng thực hiện rất đơn giản, chúng ta sử dụng dữ liệu kiểm thử (testing data). Dữ liệu này có cấu trúc giống với dữ liệu huấn luyện nhưng là một tập dữ liệu không có ở trong tập dữ liệu huấn luyện. Chúng ta sẽ thử từng mẫu dữ liệu kiểm thử, kiểm tra mô hình hoạt động có như mong đợi không. Nếu là bài toán học có giám sát, chúng ta quan sát những mẫu mô hình cho ra kết quả đúng. Nếu độ chính xác trả ra là 90% trên tập kiểm thử, chúng ta kết luận mô hình này có độ chính xác là 90%.[2]

Điều quan trọng phải nhớ rằng dữ liệu huấn luyện và kiểm thử phải tách biệt. Đây là cách duy nhất để kiểm tra hiệu suất của mô hình học máy. Bạn có thể có kết quả tốt trên tập dữ liệu huấn luyện, nhưng có kết quả tệ trên tập kiểm thử. Vấn đề này được gọi là overfiting, mô hình học máy quá khớp với dữ liệu huấn luyện và có một kết quả dự đoán tệ. Cách thường dùng để tránh overfiting là sử dụng ít số lượng đặc trưng hơn, đơn giản hóa mô hình, sử dụng tập huấn luyện lớn hơn và không gian dữ liệu bao quát hơn.[2]

* + 1. ***Hiệu suất***

Trong thực tế, nếu bạn phát triển giải pháp cho nhu cầu thực tiễn thì độ chính xác và tốc độ xử lý là những yếu tố bạn cần quan tâm. Ứng dụng của bạn dù có hiệu quả tốt nhưng xử lý chậm thì cũng chẳng sài được.

Để thực hiện được những yếu tố kể trên thì việc đầu tiên là bạn cần lựa chọn một Machine learning Framework, các framework này thường có ở một số ngôn ngữ lập trình nhất định. Python và thư viện Scikit-learn là một sự lựa chọn hoàn hảo.[2]

Tuy nhiên, vấn đề hiệu năng vẫn có thể xảy ra. Nó phụ thuộc vào số lượng dữ liệu huấn luyện, độ phức tạp và thuật toán bạn sử dụng. Điều này có thể cần những máy tính với bộ nhớ lớn và hiệu năng cao để huấn luyện. Có thể bạn phải huấn luyện nhiều lần để có được kết quả tốt. Bạn cũng có thể đào tạo lại mô hình để phủ hết những không gian dữ liệu mới và tối ưu độ chính xác. Nếu muốn huấn luyện dữ liệu lớn với tốc độ nhanh, chúng ta sẽ cần những máy tính với cấu hình cao, chúng cho phép chạy đa luồng tốt và tính toán song song.[2]

1. **Ứng dụng của Machine learning trong thực tiễn**
   1. ***Tự động phân loại***

Phân loại tin tức là một ứng dụng của machine learning. Một vấn đề thực tế là bây giờ khối lượng thông tin đã tăng lên rất nhiều trên web. Tuy nhiên, mỗi người có sở thích hoặc lựa chọn cá nhân của mình. Vì vậy, để chọn hoặc thu thập một phần thông tin phù hợp trở thành một thách thức đối với người dùng từ vô số nội dung trên trang web.[6]

Phân loại các danh mục một cách rõ ràng, dễ điều hướng giúp cho các khách hàng mục tiêu chắc chắn sẽ tăng khả năng truy cập các trang tin tức. Hơn nữa, độc giả hoặc người dùng có thể tìm kiếm tin tức cụ thể một cách hiệu quả và nhanh chóng. Có một số thuật toán machine learning có thể giải quyết vấn đề này như Support Vector Machines, Naive Bayes, K-nearest neighbor,...[6]

* 1. ***Ứng dụng trong các mạng xã hội***

Machine learning đang được sử dụng trong một loạt các ứng dụng ngày nay. Một trong những ví dụ nổi tiếng nhất là Facebook News Feed. Nguồn cấp tin tức sử dụng học máy để cá nhân hóa từng nguồn cấp dữ liệu thành viên.

Nếu một thành viên thường xuyên dừng lại để đọc hoặc thích một bài đăng của một người bạn cụ thể, News Feed sẽ bắt đầu hiển thị nhiều hơn về hoạt động của người bạn đó trước đó trong nguồn cấp dữ liệu. Đằng sau hệ thống ấy, phần mềm sử dụng phân tích thống kê và phân tích dự đoán để xác định các mẫu trong dữ liệu người dùng và sử dụng các mẫu đó để điền vào News Feed. Nếu thành viên không còn dừng lại để đọc, thích hoặc bình luận trên các bài đăng của bạn bè, dữ liệu mới đó sẽ được bao gồm trong tập dữ liệu và News Feed sẽ điều chỉnh tương ứng.[6]

Không chỉ riêng facebook, ta có thể bắt gặp những tính năng tương tự đó qua các mạng xã hội khác như google, instagram,...

* 1. ***Nhận diện hình ảnh***

Nhận dạng hình ảnh là một trong những ví dụ về machine learning và trí tuệ nhân tạo phổ biến nhất. Về cơ bản, nó là một cách tiếp cận để xác định và phát hiện các đặc trưng của một đối tượng trong hình ảnh kỹ thuật số. Hơn nữa, kỹ thuật này có thể được sử dụng để phân tích sâu hơn, chẳng hạn như nhận dạng mẫu, nhận diện hình khuôn, nhận dạng khuôn mặt, nhận dạng ký tự quang học và nhiều hơn nữa,...[6]

# Tài liệu tham khảo

1. <URL: <https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o>>
2. <URL: <https://nguyenvanhieu.vn/machine-learning-la-gi/>>
3. <URL: <https://khanh-personal.gitbook.io/ml-book-vn/machine-learning-la-gi>>
4. <URL: <https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u_l%E1%BB%9Bn>>
5. <URL: <https://vi.wikipedia.org/wiki/Dao_c%E1%BA%A1o_Ockham>>
6. <URL: <https://longvan.net/machine-learning-la-gi-ung-dung-cua-machine-learning.html>>