**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**MÔN NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI:**

**SỬ DỤNG THUẬT TOÁN HỒI QUY ĐỂ DỰ ĐOÁN GIÁ HÀNH**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: NGUYỄN MẠNH CƯỜNG**  **: PHẠM ĐỨC HÙNG**  **: NGUYỄN MẠNH THỜI** | | **Giảng viên hướng dẫn** | **: PHẠM THỊ KIM DUNG** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | | **Lớp** | **: D14CNPM1** | | | **Khóa** | **: 2019-2024** | | |  |
| ***Hà Nội, tháng 5 năm 2022*** |  |

**PHIẾU CHẤM ĐIỂM**

**Sinh viên thực hiện**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi Chú** |
| Nguyễn Mạnh Cường |  |  |
| Phạm Đức Hùng |  |  |
| Nguyễn Mạnh Thời |  |  |

**Giảng viên chấm**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1 |  |  |
| Giảng viên chấm 2 |  |  |

**LỜI MỞ ĐẦU**

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là **Machine Learning** (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự 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, …, chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning.

**Machine Learning** là một tập con của AI. Theo định nghĩa của Wikipedia, Machine learning is the subfield of computer science that “gives computers the ability to learn without being explicitly programmed”. Nói đơn giản, Machine Learning 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.

Mối quan tâm về **Machine Learning** ngày càng tăng lên. Sự gia tăng dung lượng và loại dữ liệu có sẵn, việc xử lý tính toán ngày càng rẻ hơn và mạnh mẽ hơn, và lưu trữ dữ liệu có giá cả phải chăng. Tất cả những điều này có nghĩa là nó có thể nhanh chóng và tự động tạo ra các mô hình mà có thể phân tích dữ liệu lớn hơn và phức tạp hơn cũng như cung cấp các kết quả nhanh hơn, chính xác hơn - ngay cả trên một quy mô rất lớn. Sự can thiệp của con người ngày càng trở nên không cần thiết, những mô hình tự động ngày càng được chú trọng áp dụng.

Chính bởi tính ứng dụng rộng rãi mà Machine Learning mang lại, vậy nên trong quá trình học và nghiên cứu tại trường Đại học Điện Lực khi học và tìm hiểu bộ môn **“ Nhập môn học máy ”** của cô **“ Phạm Thị Kim Dung ”** , nhóm chúng em đã chọn đề tài **“ *Sử dụng thuật toán hồi quy để dự đoán giá hành* ”** làm đề tài nghiên cứu để tìm hiểu sâu hơn về Machine Learning.

**CHƯƠNG I : TÌM HIỂU VỀ MACHINE LEARNING**

**1. Lịch sử ra đời**

Một sự thật khá thú vị là khái niệm trí tuệ nhân tạo đã bắt nguồn từ thời cổ đại với những ý niệm về sinh vật nhân tạo, người máy và các máy tự động khác đã tồn tại. Nhờ những suy nghĩ ban đầu đó, trí tuệ nhân tạo càng trở nên hữu hình hơn trong suốt những năm 1700 trở về sau.

Các nhà triết học đã suy tưởng về cách cơ giới hóa và điều khiển suy nghĩ của con người vào trong những cỗ máy thông minh. Từ những quá trình nghiên cứu và suy nghĩ không ngừng đã thúc đẩy các nhà triết học, toán học và logic cổ điển quan tâm càng nhiều đến trí tuệ thông minh. Kết quả cuối cùng, những năm 1940 máy tính kỹ thuật số có thể lập trình ra đời – Máy tính Atanasoff Berry (ABC). Phát minh này đã truyền cảm hứng cho các nhà khoa học tiến lên phía trước với những ý tưởng: tạo ra bộ não điện tử, tên lửa hoặc một sinh vật mang trí tuệ nhân tạo.

Gần một thập kỷ trôi qua, trước khi những ứng dụng tiến bộ của AI xâm nhập và tác động không ngừng vào các lĩnh vực nghiên cứu lẫn đời sống của con người. Thì Alan Turing, một nhà toán học, logic học người Anh, đã đề xuất một bài kiểm tra đo khả năng của một cỗ máy để tái tạo hành động của con người đến mức độ không thể phân biệt được với hành vi của con người.

Vào giữa những năm 1950, lĩnh vực nghiên cứu AI được thành lập trong hội nghị mùa hè tại Đại học Dartmouth. Tại đó, nhà khoa học máy tính và nhận thức John McCarthy đã đặt ra thuật ngữ trí tuệ nhân tạo, đánh dấu cột mốc quan trọng trong lịch sử hình thành định nghĩa “Trí tuệ nhân tạo”.

Từ những năm 1950 trở đi, nhiều nhà khoa học, lý thuyết học, logic học, lập trình viên đã tích cực hỗ trợ để củng cố sự hiểu biết rõ ràng hơn, hiện đại hơn về trí tuệ nhân tạo.

Cứ mỗi thập kỉ mới lại có những sáng kiến và phát minh làm thay đổi kiến thức cơ bản của mọi người về lĩnh vực trí tuệ nhân tạo. Hơn thế nữa, con người càng hình dung rõ ràng hơn về cái cách mà những tiến bộ lịch sử đã biến những phát minh trí tuệ nhân tạo dường như “không tưởng” có thể hiện thực khóa được trong tương lai.

**2. Giới thiệu về Machine Learning**

**2.1. Khái niệm Machine Learning là gì ?**

Machine learning là một ứng dụng của trí tuệ nhân tạo (AI) cung cấp cho các hệ thống khả năng tự động học hỏi và cải thiện từ kinh nghiệm mà không cần được lập trình rõ ràng. Học máy tập trung vào việc phát triển các chương trình máy tính có thể truy cập dữ liệu và sử dụng nó để tự học.

Quá trình học tập bắt đầu với các quan sát hoặc dữ liệu, chẳng hạn như ví dụ, kinh nghiệm trực tiếp hoặc hướng dẫn, để tìm kiếm các mẫu trong dữ liệu và đưa ra quyết định tốt hơn trong tương lai dựa trên các ví dụ mà chúng tôi cung cấp. Mục đích chính là cho phép máy tính học tự động mà không cần sự can thiệp hoặc trợ giúp của con người và điều chỉnh các hành động cho phù hợp.

**2.2. Phân loại Machine Learning**

Có rất nhiều cách phân loại machine learning, thông thường thì machine learning sẽ được phân làm hai loại chính sau:

+ Supervised learning: học có giám sát

+ Unsupervised learning: học không giám sát

Ngoài ra, machine learning còn có thể phân làm các loại sau:

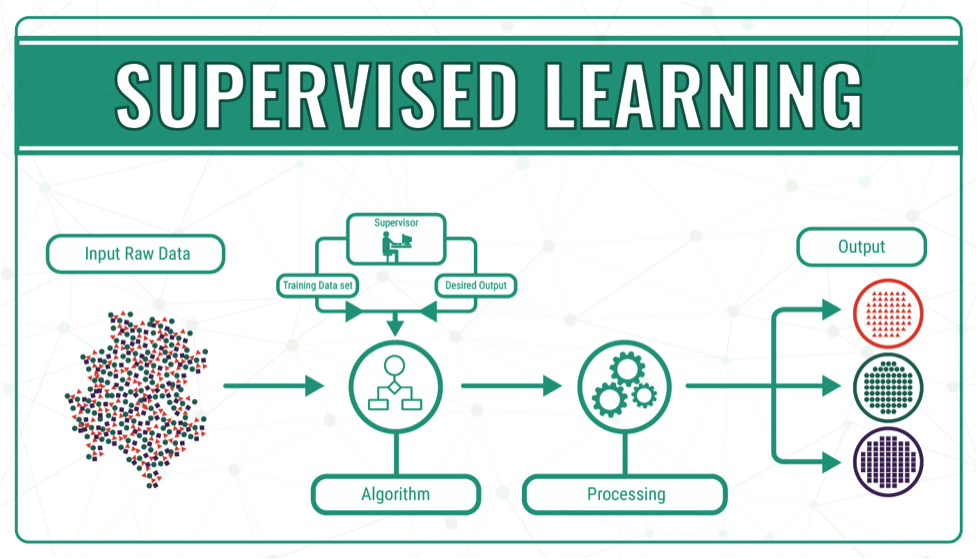
+ Semi-supervised learning: học bán giám sát

+ Deep learning: học sâu (về một vấn đề nào đó)

+ Reinforce learning: học củng cố/tăng cường

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

Học có giám sát là một kĩ thuật của ngành [học máy](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y) để xây dựng một hàm (*function*) từ dữ liệu huấn luyện. [Dữ liệu huấn luyện](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_hu%E1%BA%A5n_luy%E1%BB%87n) bao gồm các cặp gồm đối tượng đầu vào (thường dạng vec-tơ), và đầu ra mong muốn. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là [hồi qui](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_quy)), hay có thể là dự đoán một nhãn phân loại cho một đối tượng đầu vào (gọi là [phân loại](https://vi.wikipedia.org/wiki/Ph%C3%A2n_lo%E1%BA%A1i_b%E1%BA%B1ng_th%E1%BB%91ng_k%C3%AA)). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một đối tượng bất kì là đầu vào hợp lệ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hóa từ các dữ liệu sẵn có để dự đoán được những tình huống chưa gặp phải theo một cách "hợp lý".



Mô hình học có giám sát

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

Unsupervised Learning (Học không giám sát) là một nhóm thuật toán sử dụng dữ liệu không có nhãn. Các thuật toán theo cách tiếp cận này hướng đến việc mô hình hóa được cấu trúc hay thông tin ẩn trong dữ liệu. Hay nói cách khác, sử dụng các phương pháp này thiên về việc mô tả tính chất hay đặc tính của dữ liệu. Thông thường, các thuật toán này dựa trên những thông tin sau:

+ Mối quan hệ tương tự (similarity) giữa các ví dụ (được gọi là instance) trong dữ liệu như trong các thuật toán clustering (phân cụm)

+ Xác suất đồng xuất hiện của các đối tượng như trong Association mining

+ Các phép biến đổi ma trận để trích xuất các đặc trưng như PCA, SVD.

A picture containing text, sign

Description automatically generated

Mô hình học không giám sát

**2.2.3. Học bán giám sát ( Semi-supervised learning )**

Các bài toán khi chúng ta có một lượng lớn dữ liệu X nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên

**2.2.4. Học sâu ( Deep learning )**

**Deep Learning c**hỉ nhóm thuật toán nhỏ của Machine Learning lấy ý tưởng dựa trên Neural Network (mạng neuron) của con người. Deep Learning thường yêu cầu lượng dữ liệu lớn và nguồn tài nguyên sử dụng nhiều hơn các phương pháp thông thường, tuy nhiên cho độ chính xác cao hơn.

**2.2.5. Học củng cố/tăng cường ( Reinforce learning )**

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

**2.3. Ứng dụng của học máy ( Machine Learning )**

**2.3.1. Tự động phân loại**

Phân loại tin tức là một trong những ứng dụng của Machine Learning. Nhưng việc vận dụng này được thực hiện như thế nào? Thực tế, hiện nay, khối lượng thông tin trên Internet đã tăng lên rất nhiều. Tuy nhiên, mỗi người sẽ có sở thích hoặc lựa chọn thông tin khác nhau. Thu nhập được các thông tin phù hợp là một thách thức đối với người dùng từ vô số nội dung khác nhau trên web.



Machine Learning được ứng dụng trong việc phân loại tin tức

Machine Learning có thể phân loại danh mục một cách rõ ràng, dễ điều hướng, giúp cho người dùng tăng khả năng truy cập tìm kiếm các tin tức. Hơn nữa, người dùng có thể tìm kiếm được các tin tức cụ thể một cách hiệu quả và nhanh chóng.

### 2.3.2. Ứng dụng của Machine Learning trong các mạng xã hội

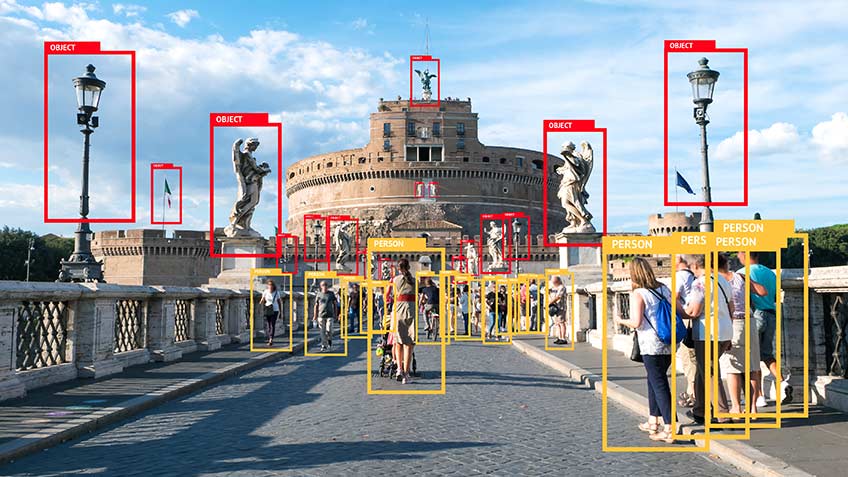
Machine Learning được sử dụng trong một loạt ứng dụng phổ biến hiện nay. Một trong những ví dụ thường thấy nhất là Facebook News Feed. Nguồn cấp của Facebook sử dụng máy họ” để cá nhân hóa các dữ liệu.

Nếu một người dùng thường xuyên dừng lại để học hoặc tương tác với một bài đăng của một người bạn, News Feed sẽ hiển thị nhiều hơn hoạt động của người bạn đó. Machine Learning đã thống kê và phân tích để xác định các mẫu trong dữ liệu người dùng, sử dụng chính các mẫu để điền vào News Feed. Ngược lại, nếu người dùng không dừng lại để đọc, tương tác trên bài đăng của bạn bè, dữ liệu đó cũng sẽ được điều chỉnh tương ứng.

Ngoài Facebook, bạn có thể bắt gặp những tính năng tương tự củ Machine Learning qua các trang mạng xã hội khác như Instagram, Google, Youtube,…

### 2.3.3. Khả năng nhận diện hình ảnh

Nhận diện hình ảnh là một trong những ví dụ phổ biến nhất về ứng dụng của Machine Learning. Về cơ bản, đây là một cách tiếp cận để phát hiện và xác định các đặc trưng của một đối tượng trong hình ảnh kỹ thuật số. Kỹ thuật này có thể được dùng để phân tích sâu hơn, ví dụ như nhận dạng khuôn mặt, nhân dạng mẫu, nhận dạng không, nhận dạng kí tự,…



Ứng dụng của Machine Learning trong nhận diện hình ảnh

Hiện nay, Machine Learning vẫn chưa đảm bảo đạt được độ chính xác 100%. Cũng như mọi lĩnh vực cần phân tích dữ liệu, dữ liệu do Machine Learning thu thập có lúc là dữ liệu tốt, có khả năng sử dụng, hoặc dữ liệu phải bỏ qua. Tuy nhiên, **ứng dụng của Machine Learning** vẫn khẳng định nó là một công cụ tuyệt vời, giúp AI có thể tiếp cận nhiều người dùng hơn.

# **CHƯƠNG 2: THUẬT TOÁN HỒI QUY SỬ DỤNG KỸ THUẬT LINEAR REGRESSION**

**2.1. Hồi quy tuyến tính ( Linear Regression )**

- Thuật toán Linear Regression hay được gọi là Hồi Quy Tuyến Tính thuộc nhóm Supervised learning ( Học có giám sát ). Hồi quy tuyến tính là một phương pháp rất đơn giản nhưng nó rất hữu ích vì nó đã được chứng minh cho một số lượng lớn các tình huống. Hồi quy chính là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập

Mục tiêu của giải thuật hồi quy tuyến tính là dự đoán giá trị của một hoặc nhiều *biến mục tiêu liên tục* (*continuous target variable*) y*,* dựa trên một vectơ đầu vào **x**.

- Công thức tổng quát của hồi quy tuyến tính:

A picture containing chart

Description automatically generated

Trong đó:

**Y** = dependent variable (biến phụ thuộc)

**X** = independent variable (biến độc lập)

**** = the intercept (hệ số chặn)

**** = the slope coefficient (hệ số góc)

**** = the error term (sai số)

* **Biến phụ thuộc (Y):**Là biến mà sự thay đổi giá trị được giải thích thông qua biến độc lập.
* **Biến độc lập (X):**Là biến được đưa vào mô hình để giải thích cho sự thay đổi giá trị của biến phụ thuộc.
* **Hệ số chặn (** **):**Thể hiện giá trị của biến phụ thuộc khi giá trị của biến độc lập bằng 0; tại giao điểm của đường hồi quy với trục biểu diễn giá trị của biến phụ thuộc (thường là trục tung).
* **Hệ số góc (** **):**Là số đơn vị thay đổi trong giá trị của biến phụ thuộc nếu như biến độc lập tăng hoặc giảm một đơn vị.
* **Sai số ():**Là khoảng chênh lệch giữa giá trị thực tế của biến phụ thuộc và giá trị ước lượng thông qua mô hình hồi quy.

**Chart, scatter chart

Description automatically generated**

*Minh họa mô hình hồi quy tuyến tính*

**CHƯƠNG III: ỨNG DỤNG THUẬT TOÁN DỰ ĐOÁN GIÁ HÀNH**