TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: DỰ ĐOÁN GIÁ VÀNG**

Giáo viên hướng dẫn: TS.Nguyễn Thị Kim Ngân

Sinh viên/nhóm sinh viên thực hiện:

1. Phạm Huy Hoàng, lớp 63CNTT1

2. Nguyễn Đình Lâm, lớp 63CNTT1

3. Nguyễn Gia Long, lớp 63CNTT1

4. Nguyễn Như Nhất Vũ, lớp 63CNTT1

**Hà Nội, năm 2023**

**MỤC LỤC**

**Phần 1. Tổng quan**

1. ***Giới thiệu về học máy***
2. ***Trình bày phương pháp học máy được sử dụng trong bài tập lớn***
   1. ***Phương pháp Hồi quy tuyến tính (Linear Regression)***
   2. ***Phương pháp Neural Network***

**Phần 2. Thực nghiệm**

***1.Mô tả tập dữ liệu***

***2. Mô tả cách giải bài toán bằng phương pháp học máy***

***3. Đánh giá mô hình***

***4. Mô tả các chức năng của chương trình***

**Phần 3. Kết luận**

**Tài liệu tham khảo**

**-Hết-**

**Lời cảm ơn**

Để hoàn thành được bài tập lớn này, chúng em xin gửi lời cảm ơn đến cô Nguyễn Thị Kim Ngân dã tận tình hướng dẫn trong suốt quá trình học tập

Mặc dù đã có nhiều cố gắng trong suốt quá trình làm bài tập, nhưng kiến thức còn hạn chế, cũng như thời gian có hạn nên không thể tránh khỏi những sai sót. Vì vậy, nhóm em mong cô góp ý để có thể chỉnh sửa, hoàn thiện bài làm tốt hơn.

Chúng em chân thành cảm ơn cô đã dành thời gian để xem xét về đề tài nhóm em.

# **Lời mở đầu**

**Học máy** ([tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh) : *machine learning*) là một lĩnh vực của [trí tuệ nhân tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

Học máy có liên quan lớn đến [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA), vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán. Nhiều bài toán suy luận được xếp vào loại bài toán [NP-khó](https://vi.wikipedia.org/wiki/NP-kh%C3%B3), vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được.

Học máy có hiện nay được áp dụng rộng rãi bao gồm [máy truy tìm dữ liệu](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u), [chẩn đoán y khoa](https://vi.wikipedia.org/wiki/Ch%E1%BA%A9n_%C4%91o%C3%A1n), phát hiện [thẻ tín dụng giả](https://vi.wikipedia.org/w/index.php?title=Th%E1%BA%BB_t%C3%ADn_d%E1%BB%A5ng_gi%E1%BA%A3&action=edit&redlink=1), phân tích [thị trường chứng khoán](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n), phân loại các [chuỗi DNA](https://vi.wikipedia.org/w/index.php?title=Chu%E1%BB%97i_DNA&action=edit&redlink=1), [nhận dạng tiếng nói](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i) và [chữ viết](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt_tay), [dịch tự động](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng)…

Những năm gần đây, sự phát triển của các hệ thống tính toán cùng với lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn đã giúp machine learning tiến thêm một bước dài.

Vì vậy, chúng em đã áp dụng hai thuật toán trong học máy - thuật toán để làm bài phương pháp hồi quy tuyến tính (Linear Regression) và mạng nơron (Neural Network) để dự đoán giá vàng.

**Phần 1. Tổng quan**

**1. Giới thiệu về học máy:**

* ***Lịch sử Machine learning****:*
* 1950 - Nhà bác học Alan Turing đã tạo ra "Turing Test (phép thử Turing)" để xác định xem liệu một máy tính có trí thông minh thực sự hay không. Để vượt qua bài kiểm tra đó, một máy tính phải có khả năng đánh lừa một con người tin rằng nó cũng là con người.
* 1952 - Arthur Samuel đã viết ra chương trình học máy (computer learning) đầu tiên. Chương trình này là trò chơi cờ đam, và hãng máy tính IBM đã cải tiến trò chơi này để nó có thể tự học và tổ chức những nước đi trong chiến lược để giành chiến thắng.
* 1957 - Frank Rosenblatt đã thiết kế mạng nơron (neural network) đầu tiên cho máy tính, trong đó mô phỏng quá trình suy nghĩ của bộ não con người.
* 1967 - Thuật toán "nearest neighbor" đã được viết, cho phép các máy tính bắt đầu sử dụng những mẫu nhận dạng (pattern recognition) rất cơ bản. Nó được sử dụng để vẽ ra lộ trình cho một người bán hàng có thể bắt đầu đi từ một thành phố ngẫu nhiên nhưng đảm bảo anh ta sẽ đi qua tất cả các thành phố khác theo một quãng đường ngắn nhất.
* 1979 - Sinh viên tại trường đại học Stanford đã phát minh ra giỏ hàng "Stanford Cart" có thể điều hướng để tránh các chướng ngại vật trong một căn phòng.
* 1981 - Gerald Dejong giới thiệu về khái niệm Explanation Based Learning (EBL), trong đó một máy tính phân tích dữ liệu huấn luyện và tạo ra một quy tắc chung để nó có thể làm theo bằng cách loại bỏ đi những dữ liệu không quan trọng.
* 1985 - Terry Sejnowski đã phát minh ra NetTalk, nó có thể học cách phát âm các từ giống như cách một đứa trẻ tập nói.
* 1990s - Machine Learning đã dịch chuyển từ cách tiếp cận hướng kiến thức (knowledge-driven) sang cách tiếp cận hướng dữ liệu (data-driven). Các nhà khoa học bắt đầu tạo ra các chương trình cho máy tính để phân tích một lượng lớn dữ liệu và rút ra các kết luận - hay là "học" từ các kết quả đó.
* 1997 - Deep Blue của hãng IBM đã đánh bại nhà vô địch cờ vua thế giới.
* 2006 - Geoffrey Hinton đã đưa ra một thuật ngữ "deep learning" để giải thích các thuật toán mới cho phép máy tính "nhìn thấy" và phân biệt các đối tượng và văn bản trong các hình ảnh và video.
* 2010 - Microsoft Kinect có thể theo dõi 20 hành vi của con người ở một tốc độ 30 lần mỗi giây, cho phép con người tương tác với máy tính thông qua các hành động và cử chỉ.
* 2011 - Máy tính Watson của hãng IBM đã đánh bại các đối thủ là con người tại Jeopardy.
* 2011 - Google Brain đã được phát triển, và mạng deep nơron (deep neural network) của nó có thể học để phát hiện và phân loại nhiều đối tượng theo cách mà một con mèo thực hiện.
* 2012 - X Lab của Google phát triển một thuật toán machine learning có khả năng tự động duyệt qua các video trên YouTube để xác định xem video nào có chứa những con mèo.
* 2014 - Facebook phát triển DeepFace, một phần mềm thuật toán có thể nhận dạng hoặc xác minh các cá nhân dựa vào hình ảnh ở mức độ giống như con người có thể.
* 2015 - Amazon ra mắt nền tảng machine learning riêng của mình.
* 2015 - Microsoft tạo ra Distributed Machine Learning Toolkit, trong đó cho phép phân phối hiệu quả các vấn đề machine learning trên nhiều máy tính.
* 2015 - Hơn 3.000 nhà nghiên cứu AI và Robotics, được sự ủng hộ bởi những nhà khoa học nổi tiếng như Stephen Hawking, Elon Musk và Steve Wozniak (và nhiều người khác), đã ký vào một bức thư ngỏ để cảnh báo về sự nguy hiểm của vũ khí tự động trong việc lựa chọn và tham gia vào các mục tiêu mà không có sự can thiệp của con người.
* 2016 - Thuật toán trí tuệ nhân tạo của Google đã đánh bại nhà vô địch trò chơi Cờ Vây, được cho là trò chơi phức tạp nhất thế giới (khó hơn trò chơi cờ vua rất nhiều). Thuật toán AlphaGo được phát triển bởi Google DeepMind đã giành chiến thắng 4/5 trước nhà vô địch Cờ Vây.
* ***Vai trò: Machine Learing***đóng vai trò quan trọng trong nhiều lĩnh vực và mang lại nhiều lợi ích. Dưới đây là 1 số vai trò chính:
* Dự đoán và dự báo: Machine learning có khả năng học từ dữ liệu quá khứ để dự đoán và dự báo các sự kiện tương lai.
* Phân loại và nhận dạng: Machine learning có khả năng phân loại và nhận dạng các đối tượng và mẫu dữ liệu.
* Tư vấn và đề xuất: Machine learning có khả năng tư vấn và đề xuất dựa trên thông tin và hành vi người dùng.
* Xử lý và phân tích dữ liệu lớn: Machine learning giúp xử lý và phân tích dữ liệu lớn (big data) để tìm ra các mẫu, xu hướng và thông tin hữu ích.
* Tự động hóa và tối ưu hóa: Machine learning có thể tự động hóa quy trình và tối ưu hóa các hệ thống và quy trình phức tạp.
* Hiểu và xử lý ngôn ngữ tự nhiên: Machine learning giúp máy tính hiểu và xử lý ngôn ngữ tự nhiên, từ việc nhận dạng và tự động sinh văn bản đến dịch máy và phân tích ý kiến.
* ***Ưu điểm và hạn chế của học có giám sát và học không giám sát***:

|  | Học **có** giám sát | Học **không có** giám sát |
| --- | --- | --- |
| Ưu điểm | * Có khả năng dự đoán chính xác và phân loại các mẫu dữ liệu mới. * Đưa ra giải thích và diễn giải dễ hiểu * Có thể tận dụng các phương pháp đánh hiệu suất chuẩn mực | * Có khả năng khám phá cấu trúc ẩn và mẫu dữ liệu không biết trước. * Không yêu cầu dữ liệu huấn luyện được gán nhãn |
| Hạn chế | * Yêu cầu dữ liệu phải gán nhãn đầy đủ * Dễ bị ảnh hưởng bởi nhiễu và sai sót trong dữ liệu huấn luyện | * Khó khăn trong việc đánh giá hiệu suất mô hình * Cần nhiều dữ liệu hơn để học hiệu quả * Khó khăn trong việc diễn giải kết quả |

**2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn**

***2.1. Phương pháp Hồi quy tuyến tính***

* **Mục đích của phương pháp :**

Phân tích hồi quy tuyến tính được sử dụng để dự đoán giá trị của một biến dựa trên giá trị của biến khác.

Biến bạn muốn dự đoán được gọi là biến phụ thuộc. Biến bạn đang sử dụng để dự đoán giá trị của biến khác được gọi là biến độc lập.

* **Input :** Cho tập dữ liệu huấn luyện gồm N mẫu. Mỗi mẫu là một cặp (xi,yi):

+ xi: vector đặc trưng

+ yi: giá trị của vector đặc trưng x

* **Output :** + Hàm tuyến tính có dạng f(xi) = wxi + w0
* **Method (Cách thực hiện) :** Cần tìm hệ số w của hàm f(xi ) sao cho trung bình sai số giữa yi và f(xi ) là nhỏ nhất. Nghĩa là, tìm w để hàm số:

**f(x)=w1x1+w2x2+w3x3** đạt giá trị nhỏ nhất

* **Dạng của linear regression**

Trong phương trình : **f(x)=w1x1+w2x2+w3x3**

x = : vector đặc trưng chứa dữ liệu đầu vào.

* + = [,,…,]
* w = : vector trọng số.
* Giá trị dự đoán: y = x\*w

*+ Sai số dự đoán*

**e = y –**   được gọi là sai số dự đoán. Giá trị này sẽ được tối ưu sao cho gần 0 nhất.

Chúng ta muốn sự sai khác e giữa giá trị thực y và giá trị dự đoán là nhỏ nhất :

+ Hệ số là để thuận tiện cho việc tính toán (khi tính đạo hàm thì số sẽ bị triệt tiêu)

+ Sự sai khác giữa **y** và không đồng nghĩa với e nhỏ nhất. Vì **e = 𝑦 −**  có thể là một số âm, khi e = −∞ là rất nhỏ nhưng sự sai lệch là rất lớn. Do đó chúng ta cần

*+ Hàm mất mát*

Dùng để đánh giá độ lỗi, độ sai của mô hình. Kết quả của hàm này có giá trị càng lớn thì mô hình của chúng ta càng sai.

**L(w) =**

+ Tìm nghiệm cho bài toán: Giải phương trình đạo hàm của hàm mất mát bằng 0:

**=0**      ⬄   **w =**

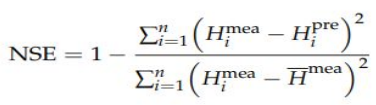
+ Nếu ma trận khả nghịch thì phương trình trên có nghiệm duy nhất là:

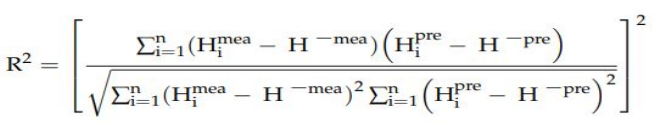
**w** =

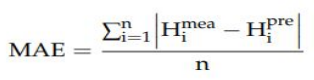
+ Nếu ma trận không khả nghịch thì nghiệm của phương trình có thể xác định dựa vào giả nghịch đảo:

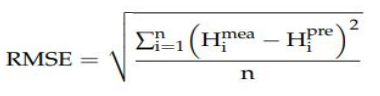
**w** =

* **Các độ đo để đánh giá mô hình : NSE , R2 , MAE , RMSE:**



****

****

****

***2.2. Phương pháp Neural network***

* **Mục đích của phương pháp**:

Mục đích chính của mô hình học máy neural network là học và nhận biết các mẫu, mối quan hệ và cấu trúc trong dữ liệu.

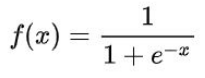
Tạo ra các mô hình tính toán mạnh mẽ để giải quyết các vấn đề phức tạp trong học máy, xử lý thông tin và trí tuệ nhân tạo.

* **Input** : Tập dữ liệu (Xtrain, ytrain), kiến trúc mạng nơ ron (số lớp ẩn, số nơ ron của mỗi lớp ẩn), hàm kích hoạt (activation function), hàm mất mát (loss function)
* **Output** : Bộ vector trọng số của các liên kết giữa các nơ ron (W) để hàm mất mát đạt giá trị tối ưu
* **Method** (Cách thực hiện) :  
  + Phương pháp phổ biến nhất để tối ưu MLP vẫn là Gradient Descent (GD)

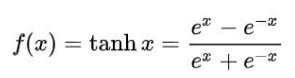
+ Để áp dụng GD, chúng ta cần tính được gradient của hàm mất mát theo từng ma trận trọng số Wl, và vector bias bl

Một số hàm kích hoạt phổ biến:

Hàm sigmoid :



Hàm tanh :



Hàm ReLu :



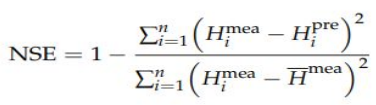
+ Hàm lỗi (tổng chêch lệch giữa đầu ra thu được và đầu ra mong muốn) là 1 hàm f(w) của các trọng số liên kết

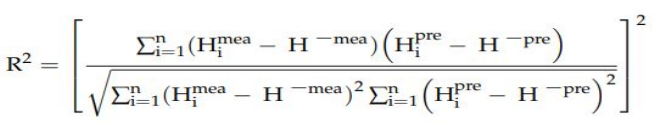
+ Cần tìm ra 1 trọng số w mà tại đó hàm lỗi là nhỏ nhất

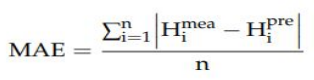
+ Hàm lỗi sẽ giảm dần mỗi khi học 1 dữ liệu mẫu. Tức là đầu ra thu được của mạng sẽ tiến sát dần đến đầu ra mong muốn

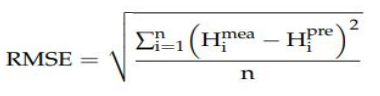
Wi+1 = Wi-f’(Wi , x)

* **Các độ đo để đánh giá mô hình** :



****

****

****

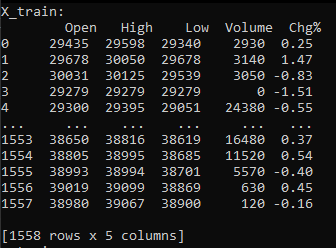
**Phần 2. Thực nghiệm**

**1.Mô tả tập dữ liệu**

* Dữ liệu gồm những chiều thông tin : *Date , Price , Open , High , Low , Volume, Chg%*
* Có 2227 mẫu dữ liệu
* Mô tả nhãn lớp của dữ liệu :
* Ma trận dữ liệu (X) gồm các thông tin :

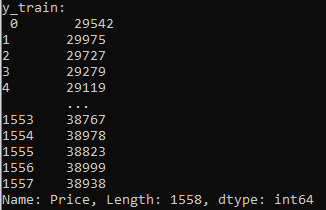
+Open : Giá ( tại thời điểm phiên giao dịch mở đầu ngày )

+High : Giá cao nhất trong ngày   
+Low : Giá thấp nhất trong ngày  
+Volume : Khối lượng giao dịch  
+Chg% : % giá thay đổi so với giá cũ

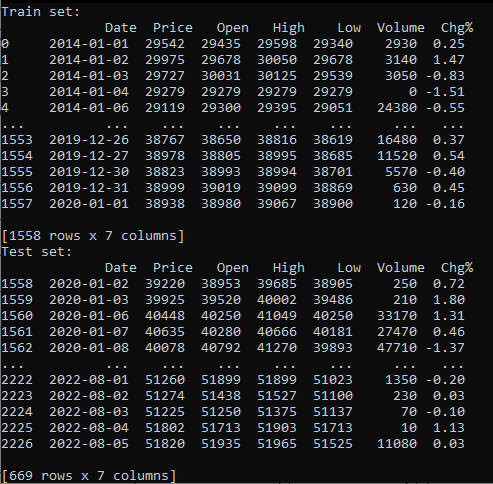


* Nhãn lớp (Y) gồm 1 chiều thông tin:

+ Price : Giá ( có thể coi là giá cuối cùng )



* Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình.



* Mô tả bài toán dùng thông tin của những thuộc tính: ***Open , High , Low , Volume, Chg%*** để dự đoán giá trị của thuộc tính***Price***

**2. Mô tả cách giải bài toán bằng phương pháp học máy**

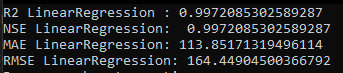
Dựa theo độ đo R2, có thể kết luận được độ đo R2 của mô hình Linear Regression cao hơn độ đo R2 của mô hình Neural network nên với tập dữ liệu này sử dụng mô hình Linear Regression là tốt nhất.

**3. Đánh giá mô hình**

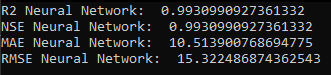
***Đánh giá chất lượng của mô hình được lựa chọn theo các độ***:

* Mô hình hồi quy tuyến tính:

Linear Regression:



* Mô hình Neural network:



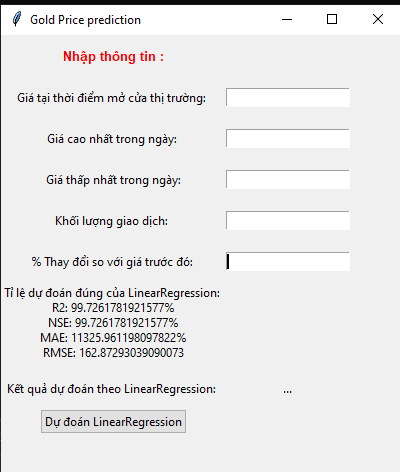
* Mô hình tốt nhất cho bài toán là mô hình Linear Regression (mô hình có độ đo R2 cao nhất).

**4. Mô tả các chức năng của chương trình**

- Chương trình bao gồm các chức năng:

Cho phép người dùng nhập thông tin và đưa ra dự đoán

- Giao diện của chương trình:

**Phần 3. Kết luận**

1. ***Xây dựng mô hình dự đoán***: Bài tập đã sử dụng thuật toán học có giám sát, để xây dựng một mô hình dự đoán giá vàng. Điều này bao gồm việc gán nhãn trên dữ liệu đã được cho sẵn và đánh giá hiệu suất của mô hình để đảm bảo tính chính xác và khả năng dự đoán.
2. ***Dự đoán giá vàng***: Bài tập đã sử dụng mô hình đã huấn luyện để đưa ra dự đoán về giá vàng trong ngày dựa trên thông tin nhập từ form. Điều này cung cấp cho chúng ta thông tin quan trọng để đánh giá và đưa ra quyết định để đầu tư đúng chỗ, đúng thời điểm.
3. ***Tìm ra tỷ lệ dự đoán đúng, tỉ lệ dự đoán sai***

**Tài liệu tham khảo**

- Bài giảng của giảng viên

- Một số tài liệu thêm: [LinearRegression](https://machinelearningcoban.com/2016/12/28/linearregression/), [Neural Network](https://machinelearningcoban.com/2017/02/24/mlp/)

- Các thư viện trong Python: [sklearn.linear\_model.LinearRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html), [sklearn.neural\_network.MLPRegressor](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html)

- Data [Gold-Price.csv](https://www.kaggle.com/datasets/nisargchodavadiya/daily-gold-price-20152021-time-series)