TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO BÀI TẬP LỚN MÔN**

**Machine Learning**

**ĐỀ TÀI:**

**Dự đoán giá xe Fiat 500 đã qua sử dụng**

**Nhóm sinh viên thực hiện** : Nhóm 12 - Lớp 60TH1

**Thành viên nhóm**:

Nông Văn Quân

Tạ Hồng Quân

Vũ Trung Kiên

**Giảng viên phụ trách môn học: Nguyễn Thị Kim Ngân**

**LỜI NÓI 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

Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, Machine Learning đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning (Học Sâu - *thực sự tôi không muốn dịch từ này ra tiếng Việt*). Deep Learning đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc .

**MỤC LỤC**

[**Phần 1. Giới Thiệu**](#_gwiki9lbpt6y) **4**

[Giới thiệu về môn Học máy](#_u0w12jq1n5io) 4

[1.1. Học máy là gì?](#_pnnlhk8tz2wa) 4

[1.2. Lịch sử phát triển của Machine Learning](#_cbv53gl6qlyj) 4

[1.3. Ứng dụng của học máy trong thực tiễn](#_q9mvxq2y3zom) 6

[1.4. Các loại thuật toán Học Máy](#_mi1nbt4hahtv) 6

[Trình bày thuật toán](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.t1g1kw31vmb7) [6](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.t1g1kw31vmb7)

[Phương pháp học máy](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.ciojbifi0cc7) [7](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.ciojbifi0cc7)

[**Phần 2: Thực Nghiệm**](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.dz22ozr0iq41)[**10**](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.dz22ozr0iq41)

[Mô tả tập huấn luyến](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.joujhznb41n8) [10](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.joujhznb41n8)

[Mô tả chương trình đề mô](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.ef0culm62bh) [11](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.ef0culm62bh)

[Mô tả kết quả dự đoán](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.ef0culm62bh) [11](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.ef0culm62bh)

[**Phần 3: Tổng kết**](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.m1uylgah2igi)[**11**](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.m1uylgah2igi)

[**Tài liệu tham khảo**](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.vfxdl2gyja3p)[**11**](https://docs.google.com/document/d/1MWcz7RctgS7N33L9MX5iUZiBR2ogCCMBtRc0icrcNbg/edit#heading=h.vfxdl2gyja3p)

# **Phần 1. Giới Thiệu**

## Giới thiệu về môn Học máy

### 1.1. Học máy là gì?

Học máy hay máy học trong tiếng anh là Machine Learning, viết tắt: ML

Học máy là một công nghệ từ lĩnh vực trí tuệ nhân tạo. Các thuật toán học máy là các chương trình máy tính có khả năng học hỏi về cách hoàn thành các nhiệm vụ và cách cải thiện hiệu suất theo thời gian.

Học máy vẫn đòi hỏi sự đánh giá của con người trong việc tìm hiểu dữ liệu cơ sở và lựa chọn các kĩ thuật phù hợp để phân tích dữ liệu. Đồng thời, trước khi sử dụng, dữ liệu phải sạch, không có sai lệch và không có dữ liệu giả.

Các mô hình học máy yêu cầu lượng dữ liệu đủ lớn để "huấn luyện" và đánh giá mô hình. Trước đây, các thuật toán học máy thiếu quyền truy cập vào một lượng lớn dữ liệu cần thiết để mô hình hóa các mối quan hệ giữa các dữ liệu. Sự tăng trưởng trong dữ liệu lớn (big data) đã cung cấp các thuật toán học máy với đủ dữ liệu để cải thiện độ chính xác của mô hình và dự đoán.

### 1.2. Lịch sử phát triển của Machine Learning

Lịch sử của Học máy là một quá trình dài dựa trên các thuật toán từ cả trăm năm trước, cũng như những phát triển về công nghệ máy tính trong 7 thập kỷ qua.

Trong đó, phải kể đến một bước đột phá lớn – công trình của Thomas Bayes vào thế kỷ thứ 18 mà sau đó dẫn tới Định lý Bayes (1812) của Pierre-Simon Laplace. Tương tự, những công trình khác như phép bình phương nhỏ nhất cho dữ liệu hồi quy (1805) của Adrien-Marie Legendre, và kỹ thuật phân tích “Chuỗi Markov” (1913) của Andrey Markov cũng đã góp vai trò to lớn trong sự phát triển của Học máy hiện đại.

* **Trước 1940:** Rất nhiều thuật toán đằng sau phép Học máy ngày nay đã xuất hiện từ trước khi chiếc máy tính đầu tiên ra đời và bắt nguồn từ phép thống kê.

Trong đó, phải kể đến một bước đột phá lớn – công trình của Thomas Bayes vào thế kỷ thứ 18 mà sau đó dẫn tới Định lý Bayes (1812) của Pierre-Simon Laplace. Tương tự, những công trình khác như phép bình phương nhỏ nhất cho dữ liệu hồi quy (1805) của Adrien-Marie Legendre, và kỹ thuật phân tích “Chuỗi Markov” (1913) của Andrey Markov cũng đã góp vai trò to lớn trong sự phát triển của Học máy hiện đại.

* **1948 Máy tính có chương trình lưu trữ sẵn:** Vào cuối những năm 1940, các máy tính với chương trình được lưu trữ sẵn đã có thể lưu trữ chương trình và dữ liệu trong cùng một bộ nhớ.
* **1950 Máy tính và trí thông minh**: Nghiên cứu “Computing Machinery and Intelligence” (Máy tính và Trí thông minh), được xuất bản năm 1950 bởi Alan Turing đã đưa ra một câu hỏi mà cho tới hiện tại vẫn chưa được giải đáp: “Liệu máy móc có thể suy nghĩ?” Nghiên cứu đã đi vào sức mạnh của máy tính, và là một trong những bài viết học thuật đầu tiên tập trung mô tả cách phát triển trí tuệ nhân tạo. Một trong những thử nghiệm nổi tiếng của nghiên cứu là “trò chơi mô phỏng” (imitation game), trong đó mức độ thông minh của máy tính sẽ được đánh giá bằng khả năng phân biệt người thật và máy tính khác qua tin nhắn. BBC đã mời Turing tới chia sẻ về thử nghiệm này trong radio “Third Programme” lần lượt vào năm 1951 và 1952.
* **1974 “Kỷ băng hà” AI đầu tiên**
* **1996 Deep Blue đánh bại Garry Kasparov:** Kasparov đã thắng vào năm 1996, nhưng lại bị đánh bại vào năm 1997 bởi một phiên bản cái tiến của Deep Blue. Để giành được chiến thắng này, Deep Blue đã phải sử dụng rất nhiều khả năng tính toán, với 480 con chip đặc biệt chỉ dùng để chơi cờ, giúp cỗ máy này tìm kiếm từ 6-20 nước cờ trước cho mỗi vị trí, “học hỏi” được qua quan sát hàng nghìn ván cờ.
* **2006 Truyền ngược**
* **2014 DeepMind:** DeepMind trở nên đặc biệt được chú ý sau khi thể hiện khả năng chơi game dựa vào hành vi của các đơn vị pixel trên màn hình, cũng như khả năng phát triển hệ nơron thần kinh có thể truy cập bộ nhớ ngoài – Máy nơron Turing (Neural Turing Machine). Tuy vậy, vẫn có khá nhiều hoài nghi xoay quanh DeepMind sau khi nó được cho là đã vi phạm quyền riêng tư của bệnh nhân Bệnh viện Royal Free khi đào tạo hệ thống y tế với dữ liệu bệnh nhân, mặc dù cáo buộc này sau đó đã bị bác bỏ.
* **2016 AlphaGo đánh bại Lee Sedol**

### 1.3. Ứng dụng của học máy trong thực tiễn

Học máy được sử dụng rất nhiều trong thực tiễn hiện nay, không chỉ trong ngành khoa học dữ liệu (lập trình) như mọi người nghĩ. Chúng có thể áp dụng vào trong kinh tế, dịch vụ, y tế, khoa học,...

Nhưng dù có được áp dụng ở lĩnh vực nào đi chăng nữa, mục đích chung nhất của học máy vẫn là đi tìm ý nghĩa của dữ liệu, từ đó định hướng phát triển cho doanh nghiệp hay tổ chức.

### 1.4. Các loại thuật toán Học Máy

Có 2 loại học máy chính bao gồm: học có giám sát (**supervised learning**) và học không có giám sát (**unsupervised learning**)

* Học có giám sát: Trong học có giám sát, máy tính học cách mô hình hóa các mối quan hệ dựa trên dữ liệu được gán nhãn (labeled data). Sau khi tìm hiểu cách tốt nhất để mô hình hóa các mối quan hệ cho dữ liệu được gắn nhãn, các thuật toán được huấn luyện được sử dụng cho các bộ dữ liệu mới.

Ứng dụng của kĩ thuật học có giám sát: Xác định tín hiệu hay biến số tốt nhất để dự báo lợi nhuận trong tương lai của cổ phiếu hoặc dự đoán xu hướng thị trường chứng khoán.

* Học không có giám sát: Trong học không giám sát, máy tính không được cung cấp dữ liệu được dán nhãn mà thay vào đó chỉ được cung cấp dữ liệu mà thuật toán tìm cách mô tả dữ liệu và cấu trúc của chúng.

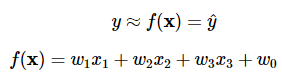
Ứng dụng của học không giám sát: Phân loại các công ty thành các nhóm công ty tương đồng dựa trên đặc điểm của chúng thay vì sử dụng tiêu chuẩn của các nhóm ngành hoặc các quốc gia.

## Phương pháp học máy

* Thuật toán: Linear Regression

"Hồi quy tuyến tính" là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục

* Công thức:

 (1)

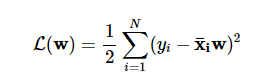
Trong đó: w1, w2, w3, w0 là các hằng số

w0 còn gọi là bias

* y là giá trị thực outcome (dựa trên số liệu thống kê chúng ta có trong tập training data)ŷ là giá trị mà mô hình Linear Regression dự đoán đượcy và ŷ là hai giá trị khác nhau do có sai số mô hình, tuy nhiên, chúng ta mong muốn rằng sự khác nhau này rất nhỏ.
* Công thức (1) có thể viết :
* **y≈ xw = ŷ**
* Nếu chúng ta đặt **w = [w0, w1, w2, w3]T** = là một vector (cột) hệ số cần phải tối ưu.Số
* X bar = [1, x1, x2, x3] là vector (hàng) dữ liệu đầu vào mở rộng. Khi đó, số 1 ở đầu được thêm vào để phép tính đơn giản hơn và thuận tiện cho việc tính toán
* Sai số dự đoán



* Hệ số là để thuận tiện cho việc tính toán (khi tính đạo hàm thì số 𝑠ẽ bị triệt tiêu).
* Chúng ta cần 𝑒𝑖2 vì ei =yi− ŷi có thể là một số âm, việc nói e𝑖 nhỏ nhất sẽ không đúng vì khi ei = −∞ là rất nhỏ nhưng sự sai lệch là rất lớn.
* Hàm mất mát
* Điều chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm w để hàm số sau đạt giá trị nhỏ nhất:



* Hàm số ℒ (w) được gọi là hàm mất mát (loss function)
* Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số w sao cho giá trị của hàm mất mát này càng nhỏ càng tốt:

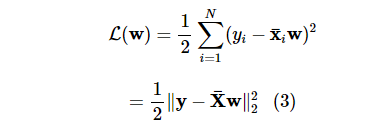


* Ta đặt:

y = [y1; y2;...; yn] là một vector cột chứa tất cả các output của training data

X = [x1; x2; … ; xn] là ma trận dữ liệu đầu vào (mở rộng) mà mỗi hàng của nó là một điểm dữ liệu.

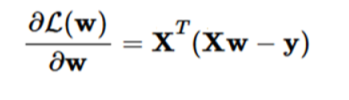
Khi đó, L(w) được viết dưới dạng ma trận đơn giản:

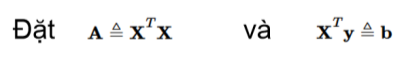


* Nghiệm cho bài toán Linear Regression

Cách phổ biến nhất để tìm nghiệm cho một bài toán tối ưu là giải phương trình đạo hàm (gradient) bằng 0! Nhưng chỉ trường hợp

* Tính đạo hàm
* Việc giải phương trình đạo hàm bằng 0 không quá phức tạp.
* Thật may, với các mô hình tuyến tính, hai việc này là khả thi đạo hàm theo w của hàm mất mát là:





Phương trình đạo hàm bằng 0 tương đương với:

XtXw = Xty (\*)

Nếu ma trận vuông A khả nghịch thì phương trình (\*) có nghiệm duy nhất: 𝑊=𝐴−1𝑏

# **Phần 2. Thực nghiệm**

### ***Mô tả bài toán***

Bài toán: Ứng dụng hồi quy tuyến tính để dự đoán giá xe Fiat 500 đã qua sử dụng.

Câu hỏi đặt ra: Có thể dự đoán giá xe Fiat 500 đã qua sử dụng qua các thông số của xe và người sở hữu hay không?

Công việc của bài toán: Ứng dụng phương pháp hồi quy tuyến tính, sử dụng các thông số kỹ thuật của xe, các chỉ số ảnh hưởng đến giá thành của xe để dự đoán giá của 1 chiếc xe khi đã biết các thông số của chiếc xe đó.

### ***Mô tả tập dữ liệu của bài toán***

Input: Đầu vào của bài toán là các thông số về chiếc xe Fiat 500 đã qua sử dụng như:

* Công suất của xe (Kw).
* Tuổi thọ của xe (Tính theo ngày).
* Số kilomet đã đi được (km).
* Số người sở hữu trước đây.
* Vĩ độ của người bán( giá ô tô ở Ý thay đổi từ Bắc vào Nam của đất nước)
* Kinh độ của người bán (giá ô tô ở Ý thay đổi từ Bắc vào Nam của đất nước)

Output: Đầu ra của bài toán là giá xe Fiat 500 đã qua sử dụng.

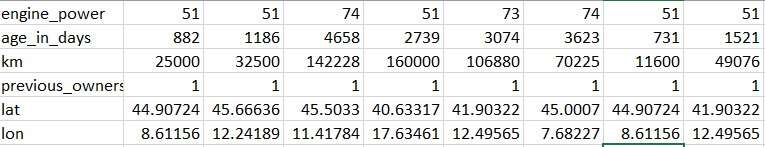
Mô tả: Mỗi chiếc xe có 1 bộ giá trị vector gồm các thông tin: Công suất của xe, tuổi thọ của xe, số kilomet đã đi, số người đã sở hữu trước đây, vĩ độ của người bán, kinh độ của người bán. Thông tin cần dự đoán là giá chiếc xe với mỗi bộ giá trị.

### ***Chia tập dữ liệu.***

Chia tập dữ liệu của bài toán ra làm 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra độ phù hợp của mô hình.

### ***Mô tả ma trận dữ liệu***

Ma trận dữ liệu huấn luyện (X)

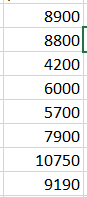


Demo 8 bộ dữ liệu trong ma trận dữ liệu huấn luyện (X)

ma trận dữ liệu huấn luyện X gồm 349 bộ dữ liệu đầu (bộ 0-348) trong file dữ liệu đã nộp đính kèm.

Link : [Ma tran X - Google Trang tính](https://docs.google.com/spreadsheets/d/18gQhjg0n9zZAozUt6zJvVtRLU6h6L_p5EuD_g9_Jv7w/edit#gid=0)

Vector đầu ra (Y)



Demo 8 giá trị đầu của vector đầu ra (Y)

vector đầu ra (Y) gồm 349 giá trị tương ứng với ma trận huấn luyện (X)

Link: [Ma tran Y - Google Trang tính](https://docs.google.com/spreadsheets/d/1NxIyBtOwmLtWRRdDv97TyH5bWfz0BYlDuCPHI1Qri10/edit#gid=0)

### ***Dùng phương pháp Hồi quy tuyến tính để xây dựng mô hình cho bài toán với tập dữ liệu huấn luyện.***

### ***Dùng tập dữ liệu kiểm tra để so sánh kết quả dự đoán của mô hình với kết quả thực tế, để đánh giá sự phù hợp của mô hình.***

Bộ dữ liệu thực tế, bộ dữ liệu dự đoán

8999, 10141.500890304698

5200, 5110.985369542541

10900, 10471.681157036743

10950, 10390.761047665583

8900, 9107.733002721172

10500, 10374.679035718014

7500, 7142.4955969418515

6200, 6819.321021037598

5490, 6615.928278294654

5699, 6389.526843773623

10500, 9682.60949675474

6990, 7308.037874902497

8579, 9552.226285069555

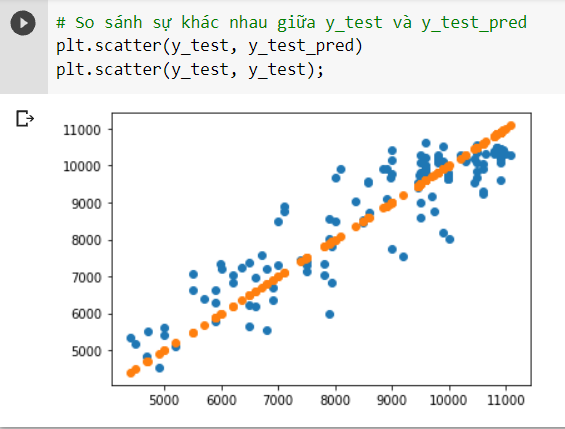
6350, 7226.545557424106

7800, 7051.234029580868

8600, 8727.281956085848

…

Đồ thị so sánh sự khác nhau giữa kết quả thực tế và kết quả dự đoán.



* Dựa vào đồ thị ta có thể thấy mô hình khá phù hợp vì phần lớn các điểm tập trung gần sát với đường hồi quy có sự phân bổ đều

# **Phần 3: Tổng kết**

Trong bài tập lớn lần này, chúng em áp dụng Học máy cho việc “ Dự đoán giá xe ô tô” bằng phương pháp Hồi quy tuyến tính - Linear Regression.

Do thời gian có hạn nên nhóm chúng em mới hoàn thành dự án ở mức cơ bản nhất.

Với đề tài và môn học này, nhóm chúng em xin chân thành cảm ơn sự giúp đỡ tận tình của cô Nguyễn Thị Kim Ngân. Mặc dù đã rất cố gắng để hoàn thành công việc, nhưng do thời gian có hạn và thiếu kinh nghiệm cũng như kỹ năng chưa cao nên việc hoàn thiện bài tập còn nhiều thiếu sót. Rất mong được cô đóng góp ý kiến để chúng em có thể hoàn thành tài liệu được tốt hơn.