**BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT**

**TRƯỜNG ĐẠI HỌC THỦY LỢI**

****

**BÀI TẬP LỚN**

**MÔN HỌC MÁY**

**ĐỀ TÀI:**

**ỨNG DỤNG THUẬT TOÁN ID3 VÀ THUẬT TOÁN PERCEPTRON LEARNING ALGORITHM(PLA) VÀO DỰ ĐOÁN GIÁ NHÀ Ở VIỆT NAM**

**Nhóm sinh viên thực hiện:**

1.Lâm Văn Thái - 60TH1

2.Phạm Thị Phương Nga - 60TH1

3.Đặng Thị Thúy - 60TH1

**Giảng viên hướng dẫn:**

TS. Nguyễn Thị Kim Ngân

HÀ NỘI, NĂM 2020

**PHẦN I:TỔNG QUAN**

# **1.1. Giới thiệu về học máy**

* Học máy là một lĩnh vực trong khoa học máy tính, là một trường con của trí tuệ nhân tạo (AI). Mục đích của việc học máy được sử dụng để phân tích và trích xuất kiến thức từ một tập dữ liệu lớn.
* Học máy có liên quan chặt chẽ với thống kê tính toán, có mối quan hệ chặt chẽ với tối ưu hóa toán học và khám phá các thuật toán có thể đưa ra dự đoán cao cấp về dữ liệu. Thuật toán học máy cho phép máy tính đào tạo trên đầu vào dữ liệu và sử dụng phân tích thống kê để xuất các giá trị nằm trong phạm vi cụ thể. Do vậy, học máy tạo điều kiện cho các máy tính trong việc xây dựng các mô hình từ dữ liệu mẫu để tự động hóa các quá trình ra quyết định dựa trên dữ liệu đầu vào.
* Các hệ thống học máy được phân thành bốn loại tùy thuộc vào bản chất của phương thức học tập bao gồm:
* Học có giám sát.
* Học không giám sát.
* Học bán giám sát.
* Học tăng cường.
* Hai trong số các phương pháp học máy được áp dụng rộng rãi nhất là học có giám sát và học không giám sát. Học có giám sát đào tạo các thuật toán dựa trên dữ liệu đầu vào và đầu ra mẫu được gắn nhãn bởi con người, còn đối với học không giám dữ liệu không có nhãn mà nó phải tự tìm cấu trúc dựa trên các dữ liệu đầu vào mà nó nhận được.

# **2.1.Giới thiệu về Cây quyết đinh (Decision Tree)**

## 2.1.1 Khái niệm

* Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

## 2.1.2 Thuật toán ID3

* ID3 (J. R. Quinlan 1993) sử dụng phương pháp tham lam tìm kiếm từ trên xuống thông qua không gian của các nhánh có thể không có backtracking. ID3 sử dụng Entropy và Information Gain để xây dựng một cây quyết định.

***Entropy*** *trong Cây quyết định (Decision Tree)*

* Entropy là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với công thức như sau:
* Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1,x2,…,xn.
* Giả sử rằng xác suất để x nhận các giá trị này là pi=p(x=xi).
* Ký hiệu phân phối này là p=(p1 ,p2 ,…,pn). Entropy của phân phối này được định nghĩa là:



***Information Gain*** *trong Cây quyết định (Decision Tree)*

* Information Gain dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Information gain cao nhất.
* Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Information Gain tại mỗi nút theo trình tự sau:
* **Bước 1**: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với Nc phần tử thuộc lớp c cho trước:



* **Bước 2**: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S1, S2, …, SK với số điểm trong mỗi child node lần lượt là m1, m2 ,…, mK , ta có:



* **Bước 3**: Chỉ số Gain Information được tính bằng:

**G(x, S) = H(S) – H(x,S)**

# **3.1.Giới thiệu Perceptron learning algorithm (PLA)**

## 3.1.1 Khái niệm:

* Là một thuật toán đơn giản giúp tìm một ranh giới siêu phẳng cho bài toán phân lớp nhị phân, với giả sử rằng tồn tại ranh giới phẳng đó. Nếu hai lớp dữ liệu có thể được phân chia hoàn toàn bằng một siêu phẳng, ta nói rằng hai lớp đó linearly separable.

## 3.1.2 Thuật toán Perceptron learning algorithm (PLA)

* Bước 1: Tại thời điểm t = 0, chọn ngẫu nhiên một vector hệ số w0.
* Bước 2: Tại thời điểm t, nếu không có điểm dữ liệu nào bị phân lớp lỗi, dừng thuật toán.
* Bước 3: Giả sử xi là một điểm bị phân lớp lỗi. Cập nhật wt+1 = wt + yixi 4. Thay đổi t = t + 1 rồi quay lại Bước 2.

**PHẦN II:THỰC NGHIỆM**

# **2.1. Mô tả bài toán**:

# Tên bài toán : Ứng dụng thuật toán ID3 và thuật toán Perceptron learning algorithm vào dự đoán giá nhà ở Việt Nam.

# Tóm tắt công việc thực hiện của bài toán:

# + Bước 1: Thu thập dữ liệu bài toán:Dataset: <https://www.kaggle.com> .

# + Bước 2: Mô tả tập dữ liệu của bài toán:

Input (11 trá trị ): gồm

* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (LỊCH SỬ LÂU ĐỜI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA
* FURNITURETATUS (NỘI THẤT)

Output: dự đoán giá nhà

* PRICE\_SEGMENT (TẦM GIÁ)

# + Bước 3: Mô tả ma trận dữ liệu (X), nhãn lớp (Y)

# + Bước 4: 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.

# + Bước 5: Dùng tập dữ liệu kiểm tra để so sánh tỷ lệ mẫu được dự đoán đúng của Perceptron và ID3 trên python.

# **2.2. Mô tả tập dữ liệu của bài toán**:

* Dữ liệu bài toán gồm: ma trận dữ liệu (X), nhãn lớp (Y)
* Ma trận x là vector đầu vào tập hợp các mẫu dữ liệu, mỗi mẫu dữ liệu là một vetor mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp y (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu:
* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (LỊCH SỬ LÂU ĐỜI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA
* FURNITURETATUS (NỘI THẤT)
* Bài toán có 150 mẫu dữ liệu.
* Dữ liệu bài toán:

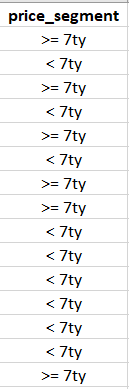
Dataset: <https://www.kaggle.com> .

# **2.3.Mô tả ma trận dữ liệu (X), nhãn lớp (Y)**

* Ma trận dữ liệu X: demo 15 mẫu dữ liệu ma trận X



* Ma trận nhãn lớp Y: demo 15 mẫu dữ liệu nhãn lớp Y



# **2.4. Chia tập dữ liệu thành 2 phần:**

* Tập dữ liệu có 150 mẫu dữ liệu chia: 105 mẫu dữ liệu dùng để huấn luyện mô hình, 45 mẫu dữ liệu dùng để kiểm tra sự phù hợp của mô hình.
* Dữ liệu tập huấn luyện (Training).
* Ma trận xtrain là vector đầu vào tập hợp các điểm dữ liệu, mỗi mẫu dữ liệu là một vetor mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp ytrain (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu:BEDROOMS ( PHÒNG NGỦ ),BATHROOMS ( PHÒNG TẮM ), STORIES (LỊCH SỬ LÂU ĐỜI), MAINROAD ( ĐƯỜNG CHÍNH ), GUESTROOM ( PHÒNG KHÁCH ), BASEMENT (TẦNG HẦM), HOTWATERHEATING (BÌNH NÓNG LẠNH), AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ), PARKING(BÃI ĐẬU XE), PREFAREA, FURNITURETATUS(NỘI THẤT).
* Tập dữ liệu huấn luyện gồm 105 mẫu dữ liệu.
* Dữ liệu tập kiểm tra (Testing)
* Ma trận xtest là vector đầu vào tập hợp các điểm dữ liệu, mỗi mẫu dữ liệu là một vector mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp ytest (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu: BEDROOMS ( PHÒNG NGỦ ), BATHROOMS (PHÒNG TẮM), STORIES (LỊCH SỬ LÂU ĐỜI), MAINROAD ( ĐƯỜNG CHÍNH ), GUESTROOM ( PHÒNG KHÁCH ), BASEMENT (TẦNG HẦM), HOTWATERHEATING (BÌNH NÓNG LẠNH), AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ), PARKING (BÃI ĐẬU XE), PREFAREA, FURNITURETATUS (NỘI THẤT).
* Tập dữ liệu kiểm tra gồm 45 mẫu dữ liệu.

# **2.5. Dùng thuật toán Perceptron và thuật toán ID3 để xây dựng mô hình phân lớp cho bài toán**

Demo chương trình

**PHẦN III:ĐÁNH GIÁ**

**3.1.Mô tả kết quả dự đoán bằng thuật toán ID3**

* Demo 15 kết quả dự đoán từ mô hình huấn luyện so với giá trị thực tế

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

...

Thuật toán ID3 dự đoán đúng được 31 giá trị trên tổng số 45.

# **3.2.Mô tả kết quả dự đoán bằng thuật toán Perceptron learning algorithm (PLA)**

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

...

Thuật toán Perceptron dự đoán đúng được 20 giá trị trên tổng số 45

# **3.3.Nhận xét**

* Từ vào tập dữ liệu kiểm tra ứng dụng hai thuật toán ID3 và Perceptron, dựa vào mô tả kết quả dự đoán của hai thuật toán ở trên ta thấy rằng thuật toán ID3 tỷ lệ mẫu dự đoán đúng với giá trị thực tế nhiều hơn so với sử dụng thuật toán Perceptron.

=> Vậy nên sử dụng thuật toán ID3 phù hợp hơn với mô hình dự đoán của bài toán.

**TÀI LIỆU THAM KHẢO**

* Tham khảo slide “Học máy” của cô TS.Nguyễn Thị Kim Ngân
* Tài liệu file: <https://github.com/tiepvupsu/ebookMLCB> của thầy Vũ Hữu Tiệp
* Dataset:[https://www.kaggle.com/](https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe/data?fbclid=IwAR0x2sXrpPkU5PBKFhvSrWrJL344T-Kwyx5Qtunm_z4TOkW3xHsPJ7Slg3s)