**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**Chuyên Ngành: Công Nghệ Thông Tin**

-------\*\*\*-------

****

**MACHINE LEARNING**

**ĐỀ TÀI: ỨNG DỤNG THUẬT TOÁNHỌC GIÁM SÁT**

**ĐỂ ĐÁNH GIÁ BẤT ĐỘNG SẢN**

*Giảng viên:* **Thầy Vũ Quang Huy**

*Sinh viên thực hiện:* **Phạm Thu Thảo**

**Nguyễn Thị Phi Vân**

**Cao Nguyễn Vũ Toàn**

Tp. HCM, tháng 12 năm 2018

# **CHƯƠNG 1. ĐẶT VẤN ĐỀ - GIỚI THIỆU ĐỀ TÀI**

## **Đặt vấn đề**

Mỗi người chúng ta thường sẽ thực hiện giao dịch bất động sản ít nhất một lần trong đời. Số tiền dành cho mua nhà là không nhỏ, vì vậy việc người mua quan tâm không chỉ ở việc lựa chọn được một ngôi nhà ưng ý mà còn xem giá cả có hợp lý hay không. Bên cạnh đó, hiện nay nhu cầu sống của người dân ngày càng tăng cao theo thu nhập bình quân đầu người. Xuất phát từ ý nghĩ “An cư lạc nghiệp” mà có ý định đầu tư một căn căn hộ chất lượng, cao cấp của người dân đã dần trở nên phổ biến. Nhưng không phải bất kỳ ai cũng có thể định giá chính xác cho ngôi nhà hoặc căn hộ của mình. Đôi khi những sự nhầm lẫn hoặc sự thiếu kinh nghiệm cũng có thể dẫn đến dự đoán sai và dẫn đến thua lỗ là điều không thể tránh khỏi. Ngay cả những doanh nghiệp về bất động sản thì đây cũng là một vấn đế khá nan giải. Vậy làm sao, dựa vào đâu để đưa ra một báo giá chính xác nhất, gần với thực tế nhất, cũng chính là lo nhóm em chọn đề tài “Dự báo giá nhà đất”.

## **Lý do chọn đề tài**

Xuất phát từ ý định mang đến không chỉ người mua mà còn người bán một công cụ có thể báo giá chính xác nhất có thể, từ đó góp phần giúp thị trường bất động sản chuyên nghiệp hơn, giảm tối thiểu tổn thất thua lỗ nhất có thể.

## **Ý nghĩa**

### **Ý nghĩa khoa học**

* Tìm hiểu sâu hơn về cơ sở lý thuyết, làm rõ tính ứng dụng của thuật toán được áp dụng trong đề tài.
* Đề xuất phương pháp dự đoán giá nhà đất ổn định và chính xác bằng cách áp dụng thuật toán trong Machine Learning.

### **Ý nghĩa thực tiễn**

* Kết quả sau khi nghiên cứu giúp người dùng dự đoán giá nhà đất tiện lợi và chính xác, tránh thua lỗ do thiếu kinh nghiệm hoặc sai sót.
* Định hướng sự phát triển của ngành bất động sản dựa trên kết quả nghiên cứu.

## **Mục tiêu**

* Mang đến cho người dùng một công cụ báo giá gần thực tế nhất, đáng tin cậy.
* Dữ liệu cần cho việc dự đoán phải tinh gọn, không quá phức tạp. Nghĩa là bất kỳ ngôi nhà cũng đáp ứng đầy đủ những yêu cầu cho việc dự đoán.
* Mang ứng dụng phổ biến đến người quan tâm đến mua bán nhà nói riêng và bất động sản nói chung.
* Đáp ứng được nhu cầu phát triền đối với dữ liệu toàn cầu.

# **CHƯƠNG 2. LÝ THUYẾT – LINEAR REGRESSION**

## **Linear regression**

### **Linear là gì?**

Vì ta cần tìm ra 1 đường thẳng, đường thẳng trong tiếng Anh là line, linear là “có thể biểu diễn bằng 1 đường thẳng” — trong toán học Việt Nam còn gọi là “tuyến tính” (tuyến trong tiếng Hán Việt nghĩa là đường thẳng — tuyến tính là có tính chất của đường thẳng).

### **Regression là gì?**

Là một cái tên lấy từ bộ môn [thống kê (statistics)](https://medium.com/pymi/d%C3%B9ng-python-%C4%91%E1%BB%83-h%E1%BB%8Dc-th%E1%BB%91ng-k%C3%AA-8e41dfdaaf97), một thuật toán/khái niệm mang từ bên statistics về dùng.  
Trong ngành thống kê: regression là mối quan hệ giữa giá trị đầu ra (y) với các giá trị (biến) đầu vào (x, t …).

### **Linear regression là gì?**

Linear regression là một phương pháp (thuật toán) thuộc loại đơn giản nhất trong Machine Learning, trong tiếng Việt gọi là *hồi quy tuyến tính*, với mục đích TÌM (vẽ) ra một đường thẳng, sao cho nó đi qua hoặc đi gần nhất với các điểm cho trước. Từ một tập dữ liệu cho trước, khi ta vẽ được một đường thẳng như vậy, ta có thể đoán xem các điểm khác sẽ nằm ở đâu.

Có thể nói 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 véc-to đầu vào **x**.

Linear regression đóng một vai trò quan trọng trong lĩnh vực [trí thông minh nhân tạo](https://en.wikipedia.org/wiki/Artificial_intelligence) như Machine Learning . Thuật toán hồi quy tuyến tính là một trong những thuật toán [học máy được giám sát](https://en.wikipedia.org/wiki/Supervised_learning) cơ bản do tính đơn giản tương đối và các thuộc tính nổi tiếng của nó.

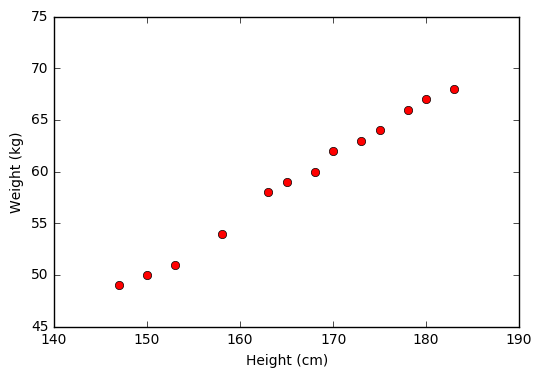
Xét ví dụ về chiều cao và cân nặng, ta có bảng dữ liệu như sau:

| **Chiều cao (cm)** | **Cân nặng (kg)** | **Chiều cao (cm)** | **Cân nặng (kg)** |
| --- | --- | --- | --- |
| 147 | 49 | 168 | 60 |
| 150 | 50 | 170 | 72 |
| 153 | 51 | 173 | 63 |
| 155 | 52 | 175 | 64 |
| 158 | 54 | 178 | 66 |
| 160 | 56 | 180 | 67 |
| 163 | 58 | 183 | 68 |
| 165 | 59 |  |  |

*Bảng 2.1.*

Chúng ta có thể thấy là cân nặng sẽ tỉ lệ thuận với chiều cao (càng cao càng nặng), nên có thể sử dụng Linear Regression model cho việc dự đoán này.

Ta thể hiện lên đồ thị để thấy rõ hơn sự tuyến tính trong í dụ này:

*Hình 2.1.*

Từ đồ thị này ta thấy rằng dữ liệu được sắp xếp gần như theo 1 đường thẳng, vậy mô hình Linear Regression nhiều khả nSăng sẽ cho kết quả tốt:

(cân nặng) = w\_1\*(chiều cao) + w\_0

### **Các giả định trong phân tích hồi qui tuyến tính**

Phân tích hồi qui tuyến tính không chỉ là việc mô tả các dữ liệu quan sát được trong mẫu (sample) nghiên cứu mà cần phải suy rộng cho mối liên hệ trong dân số (population). Vì vậy, trước khi trình bày và diễn dịch mô hình hồi qui tuyến tính cần phải dò tìm vi phạm các giả định. Nếu các giả định bị vi phạm thì các kết quả ước lượng không đáng tin cậy được.

Các giả định cần thiết trong hồi qui tuyến tính:

* là biến số cố định, không có sai sót ngẫu nhiên trong đo lường.
* Phần dư (trị số quan sát trừ cho trị số ước đoán) phân phối theo luật phân phối chuẩn
* Phần dư có trị trung bình bằng 0 và phương sai không thay đổi cho mọi trị
* Không có tương quan giữa các phần dư

### **Một vài tính chất của hồi quy tuyến tính**

* Đường hồi quy luôn luôn đi qua trung bình của biến độc lập (x) cũng như trung bình của biến phụ thuộc (y)
* Đường hồi qui tối thiểu hóa tổng của "Diện tích các sai số". Đó là lý do tại sao phương pháp hồi quy tuyến tính được gọi là "Ordinary Least Square (OLS)"
* Β1 giải thích sự thay đổi trong Y với sự thay đổi X bằng một đơn vị. Nói cách khác, nếu chúng ta tăng giá trị của X bởi một đơn vị thì nó sẽ là sự thay đổi giá trị của Y

### **Nhược điểm của Linear regression**

* Hạn chế đầu tiên của Linear Regression là nó rất **nhạy cảm với nhiễu** (sensitive to noise)
* Hạn chế thứ hai của Linear Regression là nó **không biễu diễn được các mô hình phức tạp**

## **K-nearest neighbor**

### **Định nghĩa**

K-nearest neighbor (KNN) là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này *không học* một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning)), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là [Classification](https://machinelearningcoban.com/2016/12/27/categories/#classification-phan-loai) và [Regression](https://machinelearningcoban.com/2016/12/27/categories/#regression-hoi-quy). KNN còn được gọi là một thuật toán [Instance-based hay Memory-based learning](https://en.wikipedia.org/wiki/Instance-based_learning).

### **Ý tưởng của thuật toán**

Thuật toán có 2 đầu vào, một là tập các dữ liệu đã biết trước kiểu(loại) của từng dữ liệu (hay còn gọi là tập huấn luyện - training set), đầu vào thứ 2 là dữ liệu, chúng ta chưa biết kiểu(loại) dữ liệu đó. Đầu ra của thuật toán kNN là kiểu dữ liệu của đầu vào thứ 2.

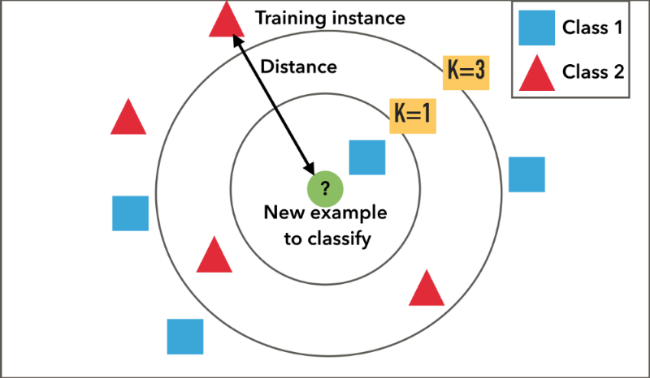
Vậy ta có thể hiểu KNN là thuật toán tìm label của một input dựa vào K điểm dữ liệu lân cận gần nó nhất (tức là điểm input).

### **Cách hoạt động của KNN**

Dưới đây trình bày từng bước cách sử dụng KNN trong việc dự đoán với biến phụ thuộc định lượng

* Xác định tham số K (số láng giềng gần nhất)
* Tính khoảng cách (Distance) giữa Query point và tất cả training samples
* Sắp xếp khoảng cách và xác định K láng giềng gần nhất với Query point
* Lấy giá trị của biến phụ thuộc Y tương ứng của K láng giềng gần nhất
* Sử dụng giá trị trung bình (average) của biến phụ thuộc Y của K láng giềng gần nhất là giá trị dự đoán của Query point.

Ta có ví dụ đơn giản, như hình dưới đây:



*Hình 2.2.*

Nhìn vào hình, hình tròn màu xanh lá chính là đối tượng cần được phân loại, xung quanh nó có các hình vuông (Category = class 1) và hình tam giác (Category = class 2) khác.

* Khi k=1, category của đối tượng mà tương đồng với nó nhất sẽ được chọn, đó chính là hình vuông;
* Khi k=3, trong số 3 đối tượng tương đồng với nó nhất có đến 2 hình tam giác, do đó category tam giác sẽ được chọn.
* Với mỗi k láng giềng ta sẽ thu được nhãn phú hợp dữ liệu đầu vào

### **Ưu điềm của KNN**

* Độ phức tạp tính toán của quá trình training là bằng 0. Do thuật toán này đơn giản gần như không phải học bất cứ điều gì.
* Việc dự đoán kết quả của dữ liệu mới rất đơn giản.
* Không cần giả sử gì về phân phối của các class.

### **Nhược điểm của KNN**

* KNN rất nhạy cảm với nhiễu khi K nhỏ.
* Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

# **CHƯƠNG 3. MODEL**

## **Flowchart**

*Hình 3.1. Flowchart*

## **Block diagram**

*Hình 3.2. Block diagram*

# **CHƯƠNG 4. GIẢI QUYẾT VẤN ĐỀ**

## **Khảo sát**

Để đánh giá chính xác giá của một căn nhà, người ta không chỉ đòi hỏi một sự hiểu biết chuyên môn về thị trường bất động sản (một thị trường rất biến động) mà còn đòi hỏi một sự hiểu biết thật sự tường tận về bản thân các thuộc tính của bất động sản đó (Mu et al., 2014). Những kiến thức này thường chỉ được lưu trữ bởi các đại lý kinh doanh bất động sản. Nếu chúng ta có thể nắm bắt kiến thức này bằng cách thu thập dữ liệu, sử dụng các dữ liệu mở, tận dụng sự giúp sức của các thuật toán, chương trình máy tính, các kiến thức này trở nên dễ tiếp cận hơn với các người dân bình thường, giúp đưa ra quyết định mà không cần dựa vào chuyên gia vì không may vị chuyên gia đó có thể tư vấn theo chiều hướng có lợi cho họ.

## **Phân tích, chọn lọc dữ liệu phù hợp**

Hiện nay, người ta chủ yếu ước lượng giá bất động sản dựa trên các phương pháp truyền thống như phương pháp so sánh trực tiếp, chiết trừ, thu nhập, thặng dư, hệ số điều chỉnh. Các phương pháp này chủ yếu nhờ sự phân tích và can thiệp của nhân viên định giá nên rất khó tránh khỏi sai lầm do chủ quan hoặc không minh bạch (Quỳnh và cs., 2015).

Ngoài các phương pháp truyền thống, trên thế giới đã và đang nghiên cứu và áp dụng rộng rãi các phương pháp có sử dụng đến các mô hình toán học để xác định giá trị bất động sản. Mới nhất là công trình (Król, 2015) sử dụng mô hình hodenic để mô hình hóa giá bất động sản ở Ba Lan.

Một cách tổng quát, trong mô hình hoderic, hàm giá của bất động sản phụ thuộc vào các thuộc tính của nó như vị trí so với trung tâm, gần đường, gần các khu tiện ích, diện tích nhà, số phòng ngủ, số tầng, số phòng tắm...

Các mô hình để xác định hàm giá có thể là các mô hình đơn giản như mô hình tuyến tính hay các mô hình phức tạp hơn như mô hình mũ, mô hình logarit,...

## **Định hướng giải quyết vấn đề**

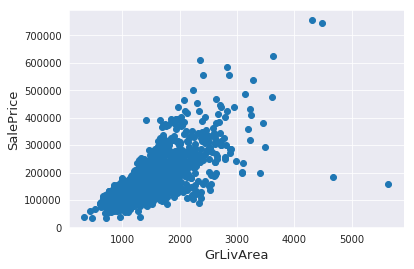
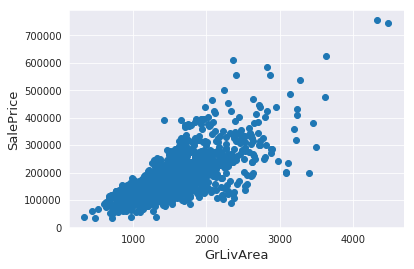
Mô hình tuyến tính là một mô hình đơn giản và được sử dụng nhiều trong bài toán xác định giá bất động sản. Trong các nghiên cứu về giá bất động sản có sử dụng đến mô hình tuyến tính chúng ta có thể kể đến các nghiên cứu của (Christian el al., 2009; Richard, 2009). Hồi quy tuyến tính xác định một đường thẳng hay một mặt phẳng qua các điểm dữ liệu trong không gian thuộc tính. Giả sử giá của bất động sản là y và các thuộc tính ảnh hưởng đến giá của nó như diện tích, độ rộng mặt tiền, độ rộng đường vào nhà, tình trạng pháp lý của khu đất, tiện ích của khu dân cư (điều kiện vệ sinh, điều kiện trường học, y tế), khoảng cách đến trung tâm phường, quận, thành phố...

## **Quá trình huấn luyện**

Dưới đây là hình ảnh minh họa của data:

*Hình 4.1.*

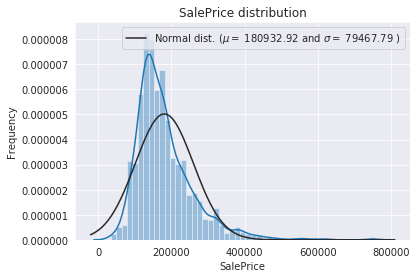
Bắt đầu huấn luyện

* Kích thước dữ liệu trước khi bỏ thuộc tính Id là: (1460, 81)
* Kích thước dữ liệu sau khi bỏ thuộc tính Id là: (1460, 80)

*Hình 4.2.*

* Dự đoán khả năng mua củ căn nhà trong tầm giá

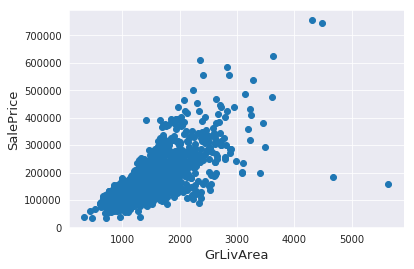
Ở đây ta dùng phương pháp K-nearest neighbor để giải quyết vấn đề này.

Giả sử ta thu thập được giá dự đoán của căn nhà từ bước trên, sau đó ta cần xác định mức giá này có phù hợp hay không, xác suất người mua sẽ mua là bao nhiêu.

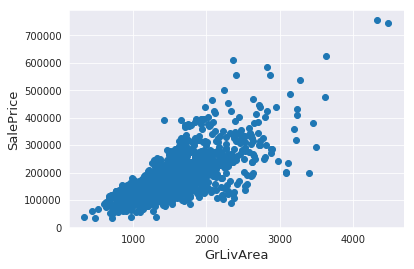
*Hình 4.3.*

## **Kiểm tra mức độ chính xác của thuật toán**

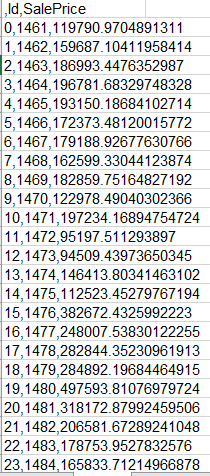
Cùng với các bước huấn luyện ta có các bước kiểm tra tương ứng:

* Kích thước dữ liệu trước khi bỏ thuộc tính Id là: (1459, 80)

*Hình 4.4.*

* Kích thước dữ liệu sau khi bỏ thuộc tính Id là: (1459, 79)

*Hình 4.5.*

Ta có thể thấy ở 2 bước này mức độ chính xác của thuật toán là rất cao. Sau quá trình huấn luyện ta thu được tập kết quả giá nhà đất như sau:

*Hình 4.6.*

# **CHƯƠNG 5. KẾT LUẬN**

Qua quá trình tìm hiều và thực hiện đề tài, chúng em thu được một số kết quả sau:

* Hiểu về thuật toán và nguyên lý hoạt động của thuật toán Linear regression và K-nearest neighbor.
* Ứng dụng thuật toán vào bài toán dự đoán giá nhà đất một cách phù hợp
* Hình thành phương pháp/ công cụ dự đoán giá nhà đất cho người dùng.

# **TÀI LIỆU THAM KHẢO**

https://machinelearningcoban.com/2017/01/08/knn/

https://codelungtung.com/2018/08/22/image-classification-k-nearest-neighbor-classification/