**MÔ HÌNH HỌC MÁY**

**LƯU Ý:** **Phần code cho các bước được trình bày trong phần báo cáo này được thực hiện trong file notebook đính kèm trong phần nộp bài của nhóm.**

**I. BÀI TOÁN ĐẶT RA**

**I.1. Bài toán**

Dự đoán giá bán nhà ở, chung cư tại thủ đô Hà Nội, Việt Nam?

**I.2. Giới thiệu chung**

- Trong học máy, **học có giám sát** là một nhóm các thuật toán phổ biến trong lĩnh vực này và một trong những vấn đề quan trọng của học có giám sát là hồi quy(regression). Hồi quy là các bài toán liên quan đến việc dự đoán đầu ra có giá trị liên tục (predicting continous valued output).

- Và trong bài toán mà nhóm đề ra thì từ những cột thuộc tính đầu vào như diện tích căn hộ, số tầng, số phòng, .... Nhóm tiến hành dự đoán cột mục tiêu là giá bán (trên m2) của chung cư, nhà ở bằng thuật toán hồi quy tuyến tính (linear regression).

**II. TIỀN XỬ LÝ DỮ LIỆU**

**II.1. Loại những thuộc tính không có ý nghĩa cho việc mô hình hóa**

- Khi nhìn vào bộ dữ liệu ta có thể thấy ngay có những cột không có ý nghĩa cho việc mô hình hóa khi chúng hoàn toàn không có mối liên quan đến biến mục tiêu. Loại bỏ sớm các thuộc tính như vậy sẽ giúp việc tiền xử lý dữ liệu được dễ dàng hơn.

- Ta có thể thấy:

+ Cột ‘Ngày’, ‘Địa chỉ’ và ‘Huyện’ **chứa các giá trị quá riêng biệt**, không có ý nghĩa cho việc trực quan hay phân tích, chúng được đăng để cung cấp đầu đủ thông tin rõ ràng cho một căn nhà cần được bán, hoặc theo format của trang web dùng để đăng bán, do đó các cột này sẽ không được lựa chọn làm thuộc tính đầu vào.

+ Thuộc tính ‘Diện tích’ được tính bằng công thức ‘Dài’\*’Rộng’, bên cạnh đó có nhiều mẫu dữ liệu không có thông tin về chiều dài và chiều rộng căn nhà mà chỉ có tổng diện tích, do đó có thể loại bỏ hai thuộc tính này để giảm độ phức tạp của mô hình.

+ Các cột của bộ dữ liệu được giữ lại bao gồm: **‘Quận’, ‘Loại hình nhà ở’, ‘Giấy tờ pháp lý’, ‘Số tầng’, ‘Số phòng’, ‘Diện tích’, ‘Giá/m2’**.

- **Đây là bước loại những thuộc tính có thể thấy ngay về mặt ý nghĩa, sau khi hoàn thành các bước tiền xử lý tiếp theo sẽ thực hiện tính correlation (hệ số tương quan) giữa từng thuộc tính với biến đầu ra để có thể đưa ra những lựa chọn chính xác hơn nữa.**

**II.2. Chuyển đổ các cột không phải dạng số về dạng số**

- Dữ liệu đầu vào cho các mô hình học máy đều phải ở dang số, do đó cần chuyển đổi các cột không phải dạng số về dạng số

- Có nhiều cách để biến đổi cột dạng danh mục về dạng số như one-hot vector, ordinal label, sử dụng một giá trị số làm đại diện, hoặc đơn giản là thay đổi trực tiếp kiểu dữ liệu của cột dự liệu đó. Việc chuyển đổi của nhóm được thực hiện như sau:

+ Nhận thấy tất cả các thuộc tính dạng danh mục của bộ dữ liệu đều có nhiều hơn 2 loại giá trị. Đối với các thuộc tính có nhiều hơn 2 loại giá trị, chúng ta sẽ má hõa bằng one-hot vector, lý do sử dụng one-hot mà không dùng ordinal hay label là để **tránh xảy ra hiện tượng bias do mã hóa thành các giá trị lớn nếu số lượng các giá trị trong cột dữ liệu là lớn.** Ví dụ đơn giản như cột ‘Quận’, có hơn 30 quận ở Hà Nội và có những Quận mà mức giá đất của chúng ngang nhau. Khi sử dụng ordinal hay label để chuyển hóa thành dạng số, các quận sẽ được đánh số từ 1 đến hơn 30, khi đó hiện tượng bias sẽ xảy ra trong trường hợp ví dụ như quận thứ 1 và quận thứ 30 có mức ảnh hưởng đến giá đất như nhau nhưng những mẫu thuộc quận thứ 30 sẽ bị phạt nặng hơn vì giá trị lên đến 30.

+ Ngoài ra các giá trị của các cột 'Số tầng', 'Số phòng ngủ', 'Diện tích', 'Giá/m2' có thể giữ nguyên lại giá trị số của nó mà không cần dùng one-hot vì số tầng hay số phòng ngủ càng nhiều thì càng ảnh hưởng đến giá của căn nhà. 'Diện tích' và 'Giá/m2' mang kiểu categorical vì chỉ có thêm phần đơn vị.

**II.3. Tính Correlations và tiếp tục chọn những thuộc tính thật sự có giá trị**

- Correlation là một thuật ngữ thống kê được sử dụng phổ biến đề cập đến mức độ liên quan của hai biến để có mối quan hệ tuyến tính với nhau hay không.

- Correlation cao nhất có giá trị là 1 (hai biến hoàn toàn có quan hệ tuyến tính) và thấp dần nếu hai biến càng không có quan hệ tuyến tính.

- Nhóm sẽ chọn ra những thuộc tính hoàn toàn không có liên quan gì đến ‘Giá/m2’ bằng cách loại các cột cho ra điểm correlation với ‘Giá/m2’ là không.

**Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động**

**-** Correlations của ‘Giá/m2’ với hầu hết các cột khác đều rất nhỏ nhưng không bằng 0 chứng tỏ đây là dấu hiệu cho thấy mực độ phân tán cực kỳ lớn của bộ dữ liệu này, có thể mô hình hồi quy sẽ gặp khó khăn.

**II.4. Xử lý các giá trị NaN**

**-** Xóa đi những mẫu có cột ‘Giá/m2’ là Nan.

- Đối với bộ dữ liệu này mỗi thuộc tính có số lượng giá trị thiếu khá nhiều do đó việc bỏ đi các dòng chứa giá trị nan có thể gây ảnh hưởng lớn đến tính chính xác khi tiến hành học trên bộ dữ liệu do thiếu thông tin.

- Giải pháp có thể sử dụng là thay thế giá trị NaN bằng các giá trị đặc biệt của cột dữ liệu chẳng hạn: trung bình, trung vị, most, ....

- Nhóm sẽ sử dụng giá trị median để thay thế các giá trị NaN cho thuộc tính 'Diện tích', do dữ liệu về thuộc tính diện tích có thể có tồn tại điểm ngoại lệ có giá trị quá lớn nên việc dùng giá trị mean có thể gây ra sai sót. Và dùng giá trị most\_frequent để thay thế các giá trị NaN cho các thuộc tính định lượng rời rạc còn lại (trừ thuộc tính 'Số tầng').

- Đặc biệt với thuộc tính 'Số tầng' do có quá nhiều mẫu bị thiếu - đến gần phân nửa số mẫu bị thiếu thuộc tính 'Số tầng' do đó không đủ thông tin để thực hiện chọn chiến lược điền giá trị thiếu, do đó nhóm sẽ tạo một giá trị mới là -1 để giữ nguyên tính toàn vẹn của dữ liệu.

**III. XÂY DỰNG MÔ HÌNH HỌC MÁY**

**III.1. Hồi quy tuyến tính đa biến (Linear Regression with Multi Variables)**

**III.1.1. Mô tả bài toán**

- Hồi quy tuyến tính đa biến áp dụng nhiều biến thuộc tính để xây dựng hypothesis cho mô hình hồi quy.

- Hypothesis chỉ có thể được xây dựng gần với giá trị thật của biến đầu ra nhất khi thuộc tính được chọn phải có mối quan hệ tuyến tính thật sự với biến y và quan trọng hơn hết nó có vai trò ảnh hưởng phải lớn đến giá trị của biến y. Đối vối bài toán dự đoán giá bán của căn hộ/chung cư thì có thể chấp nhận việc thuộc tính diện tích ngôi nhà là có vai trò ảnh hưởng lớn đến giá nhà nhưng vẫn có khả năng số phòng ngủ hay số phòng vệ sinh sẽ cũng ảnh hưởng nhiều đến giá bán.

**III.1.2. Regularization**

**-** Khi dùng càng nhiều biến thì độ phức tạp của hypothesis càng tăng lên, dẫn đến có thể tạo ra những dạng đường ngoằn ngòe quá mức cần thiết và không thực tế, gọi là vấn đề over fitting.

- Có nhiều cách để khắc phục vấn đề over fitting này như là các thuật toán lựa chọn mô hình (Model Selection) hay Regularization.

- Trong đồ án lần này nhóm sẽ sử dụng phương pháp Regularization.

**III.1.2. Cost Function**

- Ý tưởng thực hiện regularization đó chính là thêm vào cost function một đại lượng nữa là tổng các tích của một hằng số với vector tham số của hypothesis.

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, màu trắng

Mô tả được tạo tự động

**III.1.3. Gradient Descent**

- các tham số 𝜃 sẽ là những giá trị mà chúng ta cần phải thay đổi để tối ưu hóa Cost Function, và một trong những cách để thực hiện việc này là thuật toán Gradient Descent.

- Thuật toán được thực hiện như sau:

**Trong mỗi lần lặp cập nhật một cách đồng thời các tham số 𝜃𝑗 theo công thức như sau:**

**Ảnh có chứa văn bản, Phông chữ, màu trắng, đại số

Mô tả được tạo tự động**

Trong đó: alpha là 'learning rate' giúp việc học được tối ưu hơn.

**III.1.4. Normal Equation**

- Ngoài việc đạo hàm để cập nhật tham số, đối với các bộ dữ liệu lớn để tiết kiệm thời gian chúng ta có thể sử dụng một cách khác để tìm nghiệm tối ưu nhất đó chính là ‘Normal Equation’.



- Với A là ma trận chứa các giá trị 0 và chỉ có các giá trị trên đường chéo chính có giá trị là 1 nhưng trừ phần tử ở vị trí [0, 0] cũng bằng 0.

**III.3. Cài đặt**

**Các bước cài đặt được trình bày chi tiết trong file notebook source code mà nhóm nộp kèm theo.**

**III.4 Đánh giá mô hình:**

- Nhóm sử dụng độ đo R2 score để đánh giá hiệu suất mô hình hồi quy.

Ảnh có chứa văn bản, biên lai, Phông chữ, màu trắng

Mô tả được tạo tự động

- Độ đo này đánh giá độ phù hợp của mô hình. Điểm số tốt nhất có thể có là 1.0 và có thể có giá trị âm (mô hình cho ra kết quả quá tệ). Nó sẽ cho biết tỷ lệ các điểm dữ liệu nằm gần đường tuyến tính như thế nào. Nếu càng nhiều điểm dữ liệu trong bộ dữ liệu nằm gần đường tuyến tình thì điểm càng cao và ngược lại.

- **R2 score thu được cho bộ dữ liệu này khi áp dụng hồi quy tuyến tính là 0.15, bên cạnh đó giá trị Cost trên tập Train là 283.5 và trên tập Validation là 288.6.**

- Normal Equation cho kết quả tốt hơn đôi chút khi Cost trên tập Validation là 249.3.

- **NHẬN XÉT:**

+ R2 score chỉ đạt 0.15, từ đó có thể thấy được tính không phù hợp của mô hình hồi quy tuyến tính trên bộ dữ liệu này.

+ Lý do không phù hợp có thể là:

\* Nếu chỉ xét riêng về vị trí địa lý thì những quận ở trung tâm như Cầu Giấy, Hoàn Kiếm, Ba Đình, Hai Bà Trưng, Đống Đa sẽ có giá đất rất mắc hơn so với các quận vùng ven khác như Long Biên, Thanh Trì.

\* Tuy nhiên ở các vùng trung tâm thì mật độ dân số lại đông và diện tích xây dựng được phép sẽ hẹp hơn, ta có thể thấy diện tích ngôi nhà thường lớn ở những quận không phải trung tâm trong bộ dữ liệu cũng sẽ ảnh hưởng về giá.

\* Mô hình có thể dự đoán giá của các căn hộ có giá bán từ thấp đến trung bình, nhưng không dự đoán tốt với các căn hộ có giá bán cao.

\* Còn nhiều yếu tố ngoại cảnh ảnh hưởng đến giá bán như chất lượng xây dựng, chất lượng nội thất, giá thầu, ...

\* Có thể nhiều chủ căn hộ đăng giá theo cảm tính, như ở phần xem xét sự phân bố của một số thuộc tính thì khi nhìn vào bộ dữ liệu có rất nhiều điểm dữ liệu xếp chồng lên nhau, điều này chứng tỏ với những kích thước gần nhau nhưng giá rất khác biệt nhau.

**IV. ĐỀ XUẤT GIẢI PHÁP**

- Sự không phù hợp của hồi quy tuyến tính có thể được khắc phục bằng một số đề xuất như sau:

+ Hồi quy đa thức (Polynomial Regression). Tuy nhiên có thể giải pháp này có thể cũng chưa tốt hơn nhiều so với hồi quy tuyến tính vì sự phân bố của dữ liệu có độ phân tán cực kỳ lớn, không có hình dạng của một hàm nhất định nào.

+ Random Forest với những cây con là Regressor Tree.

+ Sử dụng một mạng neural nông (Shallow Neural Network) để phân tích hồi quy.

- Và trong bài làm của nhóm, nhóm sẽ thực hiện giải pháp Shallow Neural Network cho việc cải tiến quy trình phân tích hồi quy.

**V. SHALLOW NEURAL NETWORK**

- Kiến trúc mạng:

Ảnh có chứa vòng tròn, hàng, biểu đồ, ảnh chụp màn hình

Mô tả được tạo tự động

- Shallow Neural Network là một mạng neural nông gồm một lớp đầu vào, một lớp ẩn và một lớp đầu ra như hình trên. Lớp ẩn nhóm chúng em sẽ sử dụng 10 units trong lớp này và activation function được sử dụng là **hàm ReLU**.

- **Phần** **cài đặt được trình bày chi tiết trong file notebook source code mà nhóm nộp kèm theo.**

- R2 score dao động trong khoảng 0.3 đến 0.31, một kết quả tốt hơn khá nhiều với mô hình hồi quy truyền thống. Với một bộ dữ liệu phân tán cực lớn nhưng R2 score vẫn dương và đạt tới trên 0.3 là một thành công cho mô hình mạng học nông.