# **Lesson 3 : Regression (cont)**

Review knowledge : Regression

* Linear Regression
* Multivariate Regression
* Non – Linear Regression
* Feature Scaling
* Correlation , Causation , Outlier
* Loss Function
* Gradient Descent
* Underfit , Overfit

## Logistic Regression

## Khái niệm :

Logistic Regression là một thuật toán Supervised dùng để phân loại ( Classification).

Là một mô hình Regression nhằm dự đoán giá trị đầu ra rời rạc (discrete target variable) y với một giá trị đầu vào x. Việc này tương đương với chuyện phân loại các đầu vào x vào các nhóm y tương ứng.

Gọi là rời rạc vì đầu ra chỉ là yes / no ( tức 0 , 1 ), hoặc (1,2,3...,n) , không thể nào liên tục từ -∞ đến +∞.

Chart, scatter chart

Description automatically generated

Ví dụ phân loại với 2 biến x1 x2

Diagram

Description automatically generated

Ví dụ :

Bài toán Ngân hàng bạn đang làm có chương trình cho vay ưu đãi cho các đối tượng mua chung cư. Tuy nhiên gần đây có một vài chung cư rất hấp dẫn (giá tốt, vị trí đẹp,...) nên lượng hồ sơ người nộp cho chương trình ưu đãi tăng đáng kể. Bình thường bạn có thể duyệt 10-20 hồ sơ một ngày để quyết định hồ sơ có được cho vay hay không, tuy nhiên gần đây bạn nhận được 1000-2000 hồ sơ mỗi ngày. Bạn không thể xử lý hết hồ sơ và bạn cần có một giải pháp để có thể dự đoán hồ sơ mới là có nên cho vay hay không. Sau khi phân tích thì bạn nhận thấy là hai yếu tố chính quyết định đến việc được vay tiền đó là mức lương và thời gian công tác. Đây là dữ liệu bạn có từ trước đến nay:

Chart, scatter chart

Description automatically generated

Về mặt logic, chúng ta có thể nghĩ ngay đến việc vẽ 1 đường thẳng phân chia các điểm xanh và đó, rồi đưa ra quyết định cho 1 điểm mới dựa vào đường thẳng đó. Ví dụ thế này:

Chart, scatter chart

Description automatically generated

* Ví dụ đường xanh là đường phân chia. Dự đoán cho hồ sơ của người có mức lương 6 triệu và 1 năm kinh nghiệm là không chấp nhận
* Tuy nhiên, do ngân hàng đang gặp khó khăn nên hạn chế cho vay, ngân hàng yêu cầu xác suất hồ sơ đạt trên 80% mới cho vay. Bây giờ không chỉ dừng lại ở việc quyết định cho vay hay không, mà phải tìm xác suất hồ sơ đó cho vay là bao nhiêu.

Ôn lại 1 tí xác suất :

* Xác suất là các số trong khoảng [0,1] ( 0% - 100% ) , được gán vào 2 biến xảy ra – không xảy ra.
* Có những vấn đề liên quan tới xác suất : Tung đồng xu 2 mặt , thì tỉ lệ ngửa – sấp là 50 – 50. Vì có 2 mặt , tung 1 lần thì xác suất đc 1 trong 2 mặt là ½ là 50%.
* Tuy nhiên có những vấn đề mà xác suất sẽ phức tạp hơn 1 tí

Ví dụ như xác suất để trời mưa , trời không mưa là 50 – 50 . Tuy nhiên:

+ Xác suất để trời mưa nếu đó là mùa mưa sẽ lớn hơn xác suất để trời mưa nếu đó không phải là mùa mưa.

+ Xác suất để trời mưa nếu đó là mùa mưa và 7 ngày rồi chưa mưa sẽ lớn hơn xác suất để trời mưa tiếp vào ngày mai dù đã mưa 2 ngày liên tục.

* + Đó gọi là xác suất có điều kiện.

Như vậy , xác suất càng cao thì sự việc xảy ra càng chắc chắn. Ví dụ , bạn lương cao và công việc ổn định lâu năm thì xác suất bạn được ngân hàng cho vay là cao vì chắc chắn bạn có khả năng trả nợ.

Bạn không siêng năng học bài và kiến thức bạn kém thì xác suất bạn thi rớt là cao vì chắc chắn bạn không có khả năng làm tốt được bài.

Tuy nhiên, mọi thứ đều mang tính tương đối, rất khó để có chuyện xác suất là 0% hay 100% , mà chỉ dao động trong khoảng đó. Vì vậy có nhiều chuyện sẽ không thể chắc chắn thành công hay thất bại mà chỉ có thể nói xác suất thành công là bao nhiêu.

Công thức :

Nhìn vào bài toán ở trên , ta có :

P(x = 1 ) = y , xác suất mà hồ sơ được cho vay

P(x = 0 ) = 1 – y , xác suất mà hồ sơ không được cho vay

* + P(x = 1 ) + p(x = 0 ) = 1

Dựa vào công thức xác suất hậu nghiệm ( là xác suất có điều kiện của 1 biến cố ngẫu nhiên khi có 1 bằng chứng liên quan được xét đến ).

P(y/x) = =

Đặt a = ln => p(y0/x) =

Đây gọi là hàm sigmoid , một hàm đi từ 0 – 1 , hợp lý cho bài toán classification ( 0 , 1 )

A screenshot of a computer

Description automatically generated with medium confidence

công thức của linear regression là: y = w0 +w1 ∗ x1

Công thức của logistic regression là : y = = σ(w0 +w1 ∗ + w2 ∗ ) =

Loss Function : Tương tự như linear / non linear regression, logistic regression cũng có hàm loss function để đánh giá độ tốt của model, hoạt động tương tự , y dự đoán càng gần y thực tế càng tốt.

L =

Thì Loss function của Logistic Regression là :

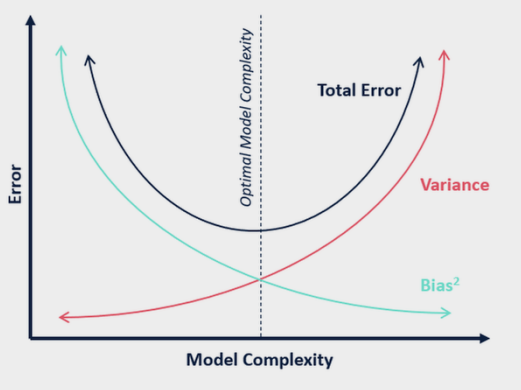
L = -y \* log(y^) – (1-y)log(1-y^) ( log ở đây là ln )

Gradient Descent của Logistic Regression cũng tương tự, chúng ta sẽ thực hành bằng google colab.

## Những vấn đề hay gặp của logistic regression.

Bias : Là mô hình có xu hướng thiên vị về mức 0 , hoặc 1.

Mức bias quá cao sẽ khiến model dự đoán sai.

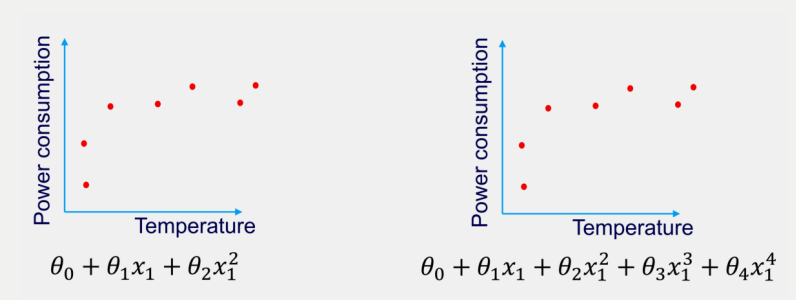


Variance ( phương sai ) : Phương sai tức là sự chênh lệch giữa dự đoán và thực tế, phương sai quá cao thì model cũng hoạt động không hiệu quả.

Shape, circle

Description automatically generated

Ngoài ra còn có một vài yếu tố khác như : regularization , quá nhiều features trong mô hình khiến mô hình bị overfit, hoặc bậc của đa thức quá cao cũng khiến mô hình bị overfit .



Có thể thấy trong model 1 chỉ có 2 features, nhưng model 1 chỉ là poly bậc 2 nhưng model 2 tới bậc 4.

Ngoài ra còn một số yếu tố về structure data, input data cũng khiến mô hình hoạt động không chuẩn.

Và data được dùng để huấn luyện đều là dạng số, vì vậy chúng ta cũng phải biết cách convert từ chữ sang số. Hoặc convert nominal data về categorical data. ( những thứ này chúng ta sẽ practice trong lab ).

* One hot encoding.
* Label encoder
* Preprocessing techniques : [Link](https://scikit-learn.org/stable/modules/preprocessing.html)



Ví dụ cho bài toán phân loại :

Before

A picture containing application

Description automatically generated

After

Table

Description automatically generated with medium confidence

Before

Table

Description automatically generated

After

Table

Description automatically generated

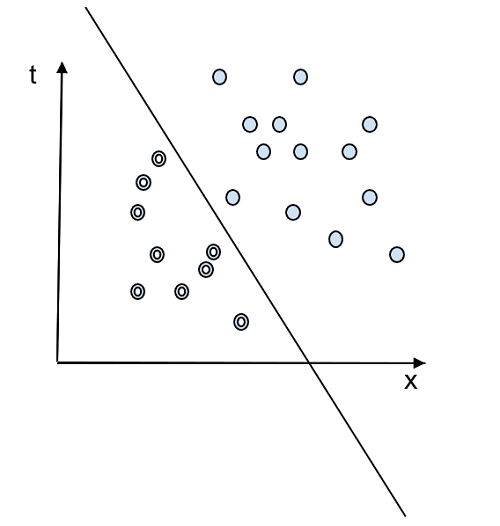
PCA để tìm ra những features chính.

## Ứng dụng

* Spam detection: Dự đoán mail gửi đến hòm thư của bạn có phải spam hay không.
* Credit card fraud: Dự đoán giao dịch ngân hàng có phải gian lận không.
* Health: Dự đoán 1 u là u lành hay u ác tính.
* Banking: Dự đoán khoản vay có trả được hay không.
* Investment: Dự đoán khoản đầu tư vào start-up có sinh lợi hay không.

## Extras :

Ngoài ra, có một thuật toán cũng gần giống với Logistic Regression , đó là thuật toán PLA ( Perceptron Learning Algorithm)



Thuật toán này cũng dùng để phân lớp 2 class . Tuy nhiên, chúng ta sẽ tìm hiểu nó kỹ hơn vào buổi học về Support Vector Machine. Còn ở buổi này, chúng ta chỉ demo nó trên colab và tìm hiểu sơ qua xem nó khác Logistic Regression chỗ nào.

Google Colab : [Link](https://colab.research.google.com/drive/1E92V6uxl3ER2nRlRhHF0vM6QIbL_vQJq?usp=sharing)