1. Hồi quy

* Hồi quy là một phương pháp phân tích dữ liệu được sử dụng để dự đoán giá trị của một biến phụ thuộc dựa trên một hoặc nhiều biến độc lập. Trong hồi quy, chúng ta cố gắng xây dựng một mô hình toán học mà mô tả mối quan hệ giữa các biến độc lập và biến phụ thuộc, sao cho có thể sử dụng mô hình đó để dự đoán giá trị của biến phụ thuộc khi biết giá trị của các biến độc lập.
* Hồi quy có thể được coi là một dạng học máy có quang Xác

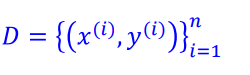
ví dụ:

* Dự đoán giá nhà dựa trên diện tích, số phòng, vị trí, vv.
* Dự đoán doanh số bán hàng dựa trên quảng cáo, mùa vụ, vv.
* Dự đoán điểm số thi dựa trên số giờ học, điểm số quá trình, vv.

Mục tiêu của mô hình học máy hồi quy là ta phải đi tối ưu hàm loss function.

Ta có:

* 1 tập traning D :

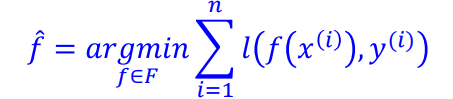


* Hàm loss l: Y x Y ->R



* Một tập giả thiết(Hypothesiss class) F

Thì mục tiêu của là ta sẽ:



Loss function: Hàm mất mát đo lường sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế của dữ liệu.

1. Hồi quy tuyến tính
   1. **Hồi quy tuyến tính đơn giản**

Y= β0\*X + β1 + ε

β0 và β1 là hai hằng số chưa biết đại diện cho độ dốc hồi quy, trong khi ε (epsilon) là thuật ngữ sai số.

Bạn có thể sử dụng hồi quy tuyến tính đơn giản để lập mô hình mối quan hệ giữa hai biến, chẳng hạn như sau:

* Lượng mưa và sản lượng cây trồng
* Tuổi và chiều cao ở trẻ em
* Nhiệt độ và sự giãn nở của thủy ngân kim loại trong nhiệt kế
  1. **Hồi quy tuyến tính bội**

Trong phân tích hồi quy tuyến tính bội, tập dữ liệu chứa một biến phụ thuộc và nhiều biến độc lập. Hàm đường hồi quy tuyến tính thay đổi để bao gồm nhiều yếu tố như sau:

Y= β0\*X0 + β1X1 + β2X2+…… βnXn+ ε

Khi số lượng biến dự báo tăng lên, các hằng số β cũng tăng lên tương ứng.

 Hồi quy tuyến tính bội lập mô hình nhiều biến và tác động của chúng lên một kết quả:

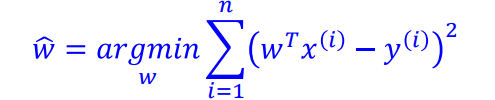
* Lượng mưa, nhiệt độ và mức độ sử dụng phân bón đối với sản lượng cây trồng
* Chế độ ăn uống và tập thể dục đối với bệnh tim
* Tăng trưởng tiền lương và lạm phát trên lãi suất cho vay gia đình

Ta có:

* W là tập hằng số
* X là tập các biến phụ thuộc

Vì vậy  sẽ là y dự đoán

 là y thực tế



Mục tiêu của bài toán là ta phải tìm W sao cho nó làm cho biểu thức trên bé nhất

* 1. Điều chuẩn của hồi quy tuyến tính(Regularizied Linear Regression**)**

Để khắc phục trường hợp overfiting

* Ridge regression ( L2 regularization)
* Lasso regression ( L1 regularization)
* ElasticNet regression ( L2+L1 regularization)

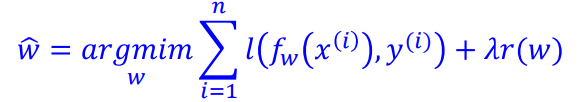
Ngoài phần traning loss thì còn có thêm hàm điều chuẩn phía sau

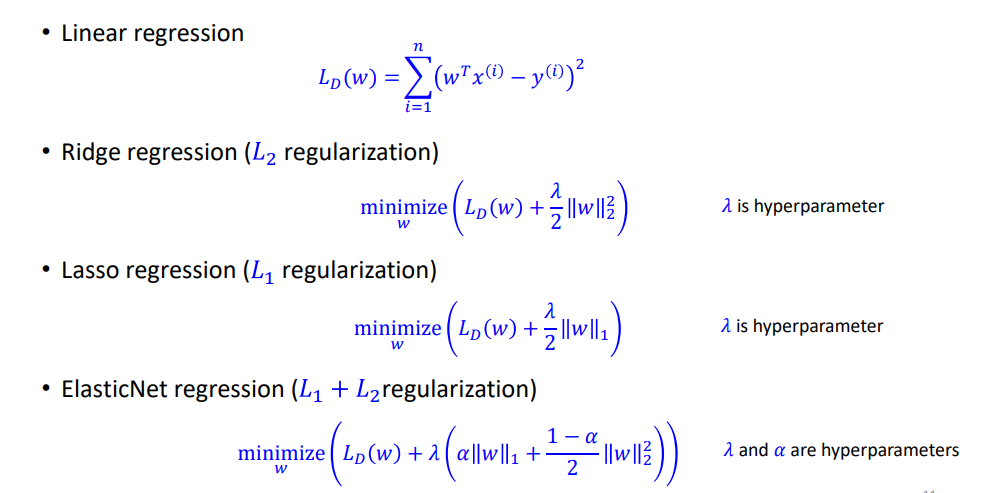
r(w) là hàm đo độ lớn độ phức tạp của mô hình

lambda: là siêu tham sô để cân bằng giữa độ phực tạp của mô hình và traning loss

+ Ưu tiên tranning loss: thì lambda tiến về 0

+ Ưu tiên độ phức tạp của mô hình thì lambda lớn





**Chuẩn L1 (Lasso Regression)**:

* + Trong chuẩn L1, chúng ta thêm một thành phần tuyệt đối của hệ số của các biến vào hàm mất mát (loss function) của mô hình.
  + Thành phần điều chuẩn được thêm vào hàm mất mát là tổng của các giá trị tuyệt đối của hệ số của các biến. Cụ thể, nếu �*θ* là vector hệ số của mô hình, thì thành phần điều chuẩn của chuẩn L1 là ∣∣�∣∣1∣∣*θ*∣∣1​, tức là tổng của các giá trị tuyệt đối của các phần tử trong vector �*θ*.
  + Chuẩn L1 thường dẫn đến một số hệ số của mô hình bằng 0, tạo ra một mô hình thưa thớt (sparse model). Điều này có thể giúp loại bỏ các biến không quan trọng và làm giảm chiều của dữ liệu, giúp tăng tốc độ tính toán và giảm độ phức tạp của mô hình.

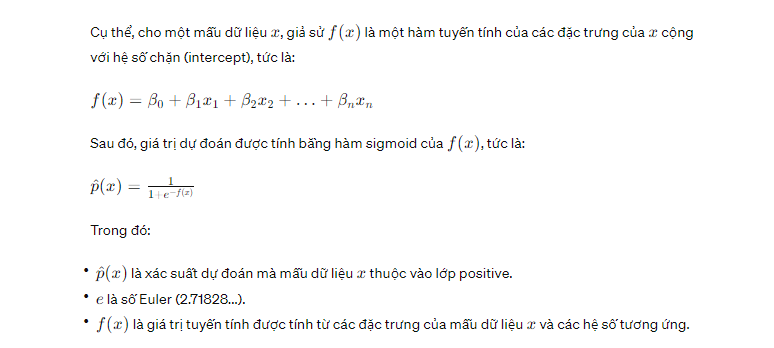
**Chuẩn L2 (Ridge Regression)**:

* + Trong chuẩn L2, chúng ta thêm một thành phần bình phương của hệ số của các biến vào hàm mất mát của mô hình.
  + Thành phần điều chuẩn của chuẩn L2 là bình phương của chuẩn Euclide của vector hệ số, tức là ∣∣�∣∣22∣∣*θ*∣∣22​, nơi ∣∣�∣∣2∣∣*θ*∣∣2​ là độ dài Euclide của vector �*θ*.
  + Chuẩn L2 giúp phân tán giá trị của các hệ số, ngăn chặn sự quá mức hóa và giảm overfitting bằng cách giảm giá trị tuyệt đối của các hệ số mà không loại bỏ chúng hoàn toàn.

1. Logistic regression

Logistic Regression là một thuật toán học máy được sử dụng trong các bài toán phân loại. Mặc dù có từ "regression" trong tên, nhưng Logistic Regression thực sự là một phương pháp phân loại, không phải là một phương pháp hồi quy (regression) trong định nghĩa của nó.

Ý tưởng chính của Logistic Regression là dự đoán xác suất của một mẫu dữ liệu rơi vào một lớp phân loại cụ thể. Điều này thường được thực hiện bằng cách sử dụng một hàm sigmoid (hoặc hàm logistic) để chuyển đổi giá trị đầu ra từ một dãy giá trị liên tục thành một giá trị xác suất nằm trong khoảng từ 0 đến 1.



1. Gradient decent

Phương pháp để tìm giá trị tối ưu của một hàm mất mát (loss function)

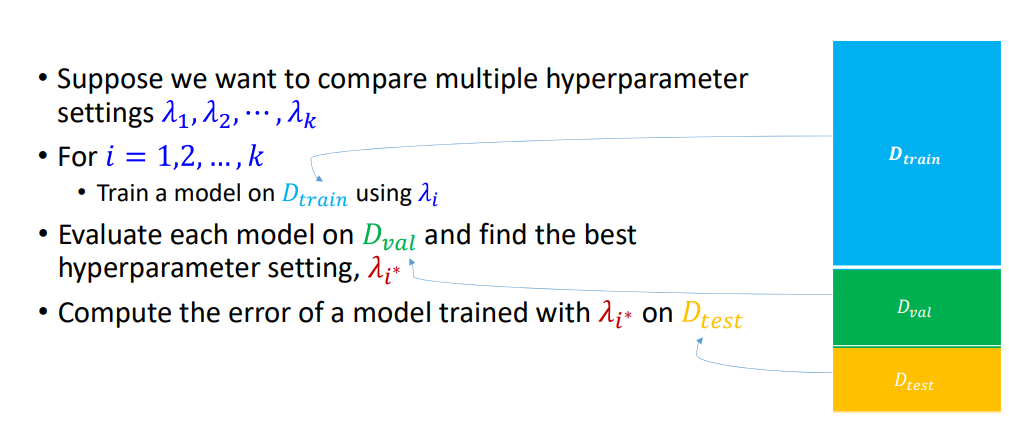
Gradient Descent for Linear Regression:

Thuật toán Gradient Descent được thực hiện qua các bước sau:

* Khởi tạo các tham số của mô hình với các giá trị ngẫu nhiên hoặc giá trị ban đầu.
* Tính gradient của hàm mất mát tại các điểm dữ liệu hiện tại.
* Di chuyển các tham số theo hướng ngược với gradient với một bước kích thước xác định được gọi là learning rate.
* Lặp lại các bước 2 và 3 cho đến khi đạt được một điều kiện dừng được xác định trước hoặc khi đạt đến một số lần lặp cố định.

1. Hyperparameter tuning.

là quá trình tối ưu hóa các siêu tham số (hyperparameters) của một mô hình machine learning để cải thiện hiệu suất của mô hình trên dữ liệu kiểm tra



1. Hiệu suất của mô hình(Hàm loss funtion)

Hiệu suất của mô hình

Một khi bạn xây dựng mô hình, câu hỏi tiếp theo đến trong đầu là để biết liệu mô hình của bạn có đủ để dự đoán trong tương lai hoặc là mối quan hệ mà bạn đã xây dựng giữa các biến phụ thuộc và độc lập là đủ hay không.

Vì mục đích này có nhiều chỉ số mà chúng ta cần tham khảo

R – Square (R^2)

Công thức tính R^2 sẽ bằng :

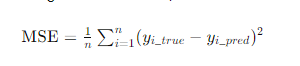


* **Tổng các diện tích (TSS)**: TSS là một phép đo tổng biến thiên trong tỷ lệ đáp ứng / biến phụ thuộc Y và có thể được coi là số lượng biến thiên vốn có trong đáp ứng trước khi hồi quy được thực hiện.
* **Sum of Squares (RSS)**: RSS đo lường lượng biến đổi còn lại không giải thích được sau khi thực hiện hồi quy.
* (TSS - RSS) đo lường mức độ thay đổi trong đáp ứng được giải thích (hoặc loại bỏ) bằng cách thực hiện hồi quy

Trong đó N là số quan sát được sử dụng để phù hợp với mô hình, σx là độ lệch chuẩn của x, và σy là độ lệch chuẩn của y.

* R2 giao động từ 0 đến 1.
* R2 của 0 nghĩa là biến phụ thuộc không thể dự đoán được từ biến độc lập
* R2 của 1 có nghĩa là biến phụ thuộc có thể được dự đoán mà không có sai số từ biến độc lập
* Một R2 giữa 0 và 1 chỉ ra mức độ mà biến phụ thuộc có thể dự đoán được. Một R2 của 0.20 có nghĩa là 20 phần trăm của phương sai trong Y có thể dự đoán được từ X; Một R2 của 0.40 có nghĩa là 40 phần trăm là có thể dự đoán v.v...

**Root Mean Square Error (RMSE)** RMSE cho biết mức độ phân tán các giá trị dự đoán từ các giá trị thực tế. Công thức tính RMSE là



N: Tổng số quan sát

Mặc dù RMSE là một đánh giá tốt cho các sai số nhưng vấn đề với nó là nó rất dễ bị ảnh hưởng bởi phạm vi của biến phụ thuộc của bạn. Nếu biến phụ thuộc của bạn có dải biến thiên hẹp, RMSE của bạn sẽ thấp và nếu biến phụ thuộc có phạm vi rộng RMSE sẽ cao. Do đó, RMSE là một số liệu tốt để so sánh giữa các lần lặp lại khác nhau của mô hình