Thuật toán Linear Regression(Hồi quy tuyến tính)

1.Giới thiệu

Ví dụ:

Một căn nhà rộng , có phòng ngủ và cách trung tâm thành phố km có giá bao nhiêu. Giả sử chúng ta đã có số liệu thống kê từ 1000 căn nhà trong thành phố đó, liệu rằng khi có một căn nhà mới với các thông số về diện tích, số phòng ngủ và khoảng cách tới trung tâm, liệu chúng ta có dự đoán được giá của căn nhà không? Nếu có thì hàm dự đoán y = f(x) có dạng như thế nào? Ở đây x = [] là một vector hang chứa thông tin input, y là một đại lượng vô hướng biểu diễn output(tức là giá nhà trong bài toán này).

Chúng ta có thể thấy rằng:

1.Diện tích nhà càng lớn thì giá nhà càng cao.

2.Số lượng phòng ngủ càng lớn thì giá nhà càng cao.

3.Càng xa trung tâm thì giá nhà càng giảm.

Một hàm số đơn giản nhất có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào trên là:

F(x) = + (1)

Trong đó Mối quan hệ y≈f(x) trên là mối quan hệ tuyến tính(linear). Bài toán đi tìm các hệ số tối ưu { } là bài toán Linear Regression.

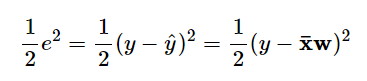
2.Phân tích toán học

2.1 Dạng của Linear Regression

Trong phương trình (1) trên, nếu đặt w = là vector cột hệ số cần phải tối ưu và x=[1,x1,x2,x3] là vector hang dữ liệu đầu vào mở rộng.

2.2 Sai số dự đoán

Chúng ta mong muốn rằng sự sai khác e giữa giá trị thực y và giá trị dự đoán  (đọc là y hat trong tiếng Anh) là nhỏ nhất. Nói cách khác, chúng ta muốn giá trị sau đây càng nhỏ càng tốt:



2.3 Hàm mất mát

Điều tương tự xảy ra với tất cả các cặp (input, outcome) (xi,yi),i=1,2,…,N,với N là số lượng dữ liệu quan sát được. Điều chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm w để hàm số sau đạt giá trị nhỏ nhất

Text

Description automatically generated with medium confidence

Hàm số L(w) được gọi là **hàm mất mát** (loss function) của bài toán Linear Regression. Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số w sao cho giá trị của hàm mất mát này càng nhỏ càng tốt. Giá trị của w làm cho hàm mất mát đạt giá trị nhỏ nhất được gọi là điểm tối ưu (optimal point), ký hiệu



Trước khi đi tìm lời giải, chúng ta đơn giản hóa phép toán trong phương trình hàm mất mát (2). Đặt y=[y1;y2;…;yN] là một vector cột chứa tất cả các output của training data; ¯X=[¯x1;¯x2;…;¯xN]¯ là ma trận dữ liệu đầu vào (mở rộng) mà mỗi hàng của nó là một điểm dữ liệu. Khi đó hàm số mất mát L(w) được viết dưới dạng ma trận đơn giản hơn:

Diagram, schematic

Description automatically generated

với  là Euclidean norm (chuẩn Euclid, hay khoảng cách Euclid), nói cách khác là tổng của bình phương mỗi phần tử của vector z. Tới đây, ta đã có một dạng đơn giản của hàm mất mát được viết như phương trình (3)

2.4. Nghiệm cho bài toán Linear Regression

**Cách phổ biến nhất để tìm nghiệm cho một bài toán tối ưu là giải phương trình đạo hàm (gradient) bằng 0!**

Đạo hàm theo w của hàm mất mát là:

Text

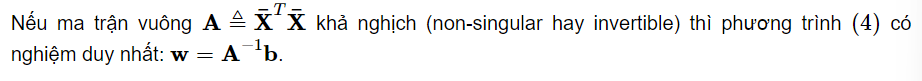
Description automatically generated

Phương trình đạo hàm bằng 0 tương đương với:



Logo

Description automatically generated with low confidence



Nếu mà trận A không khả nghịch thì phương trình vô nghiệm hoặc phương trình có vô số nghiệm.

Với khái niệm giả nghịch đảo, điểm tối ưu của bài toán Linear Regression có dạng:

A picture containing text

Description automatically generated

3.Ví dụ trên Python

Chúng ta có 1 bảng dữ liệu về chiều cao và cân nặng 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 |  |  |

Bài toán: Liệu ta có thể dự đoán cân nặng của một người dựa vào chiều cao của họ không?

Ta có thể thấy cân nặng sẽ tỉ lệ thuận theo chiều cao, nên ta sẽ sử dụng thuật toán Linear Regression để dự đoán.

Ta sẽ chia dữ liệu làm 2 phần:

Test: gồm 2 cột 155cm và 160cm

Train: Các cột còn lại

3.2 Hiển thị dữ liệu lên đồ thị

Chart, scatter chart

Description automatically generated

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 năng sẽ cho kết quả tốt:

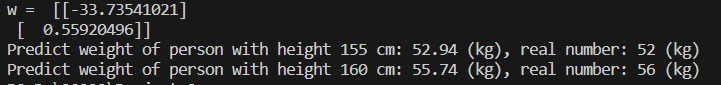
(cân nặng) = w\_1\*(chiều cao) + w\_0

3.3 Tìm nghiệm

Chart, scatter chart

Description automatically generated

Từ đồ thị bên trên ta thấy rằng các điểm dữ liệu màu đỏ nằm khá gần đường thẳng dự đoán màu xanh. Vậy mô hình Linear Regression hoạt động tốt với tập dữ liệu training.



Chúng ta thấy rằng kết quả dự đoán khá gần với số liệu thực tế.