Khái niệm Linear Regression

Bài toán Linear Regression là gì?

Bài toán Linear Regression là bài toán đi tìm các hệ số tối ưu $w_1, w_2, ..w_0$ cho phương trình:

$$f(x) = w_0 + w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n$$

Ta đặt:

$$\bar{\mathbf{x}} = \begin{bmatrix} 1 & x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \dots \\ w_n \end{bmatrix}$$

Suy ra:

$$y = f(x) = w_0 + w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n = \begin{bmatrix} 1 & x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} * \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \dots \\ w_n \end{bmatrix} = \bar{\mathbf{x}} * \mathbf{w}$$

Từ đó dùng hệ số để dự đoán giá trị $\hat{y}$ với mỗi bộ dữ liệu $x'_1, x'_2, ..x'_n$

Cách giải bài toán

Hệ số $w_1, w_2, ..w_0$ là tối ưu khi sự sai khác e của giá trị thực y và giá trị dự đoán $\hat{y}$ là nhỏ nhất (sai số dự đoán). Sai khác e được tính bằng công thức:

$$\frac{1}{2} * e^2 = \frac{1}{2} * (y - \hat{y})^2 = \frac{1}{2} * (y - \bar{\mathbf{x}} * \mathbf{w})^2$$

Như vậy với mỗi bộ dữ liệu $x_1, x_2, ..x_n$ và giá trị thực y ta lại có một sai khác e khác nhau, vậy để tối ưu ta cần tìm hệ số $w_1, w_2, ..w_0$ sao cho tổng sai khác giữa giá trị thực y và giá trị $\hat{y}$ của các bộ dữ liệu là nhỏ nhất, từ đấy ta đi đến khái niệm hàm mất mát:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} * \sum_{i=1}^{N} (y_i - \bar{\mathbf{x}}_i * \mathbf{w})^2$$

Nếu ta đặt:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}$$

và

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \dots \\ \bar{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ 1 & x_{31} & x_{32} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix}$$

thì tích : $\bar{\mathbf{X}} * \mathbf{w}$ sẽ có dạng:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ 1 & x_{31} & x_{32} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix} * \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \dots \\ w_n \end{bmatrix} = \begin{bmatrix} w_0 + w_1 * x_{11} + w_2 * x_{12} + \dots + w_n * x_{1n} \\ w_0 + w_1 * x_{21} + w_2 * x_{22} + \dots + w_n * x_{2n} \\ w_0 + w_1 * x_{31} + w_2 * x_{32} + \dots + w_n * x_{3n} \\ \dots \\ w_0 + w_1 * x_{n1} + w_2 * x_{n2} + \dots + w_n * x_{nn} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{x}}_1 * \mathbf{w} \\ \bar{\mathbf{x}}_2 * \mathbf{w} \\ \bar{\mathbf{x}}_3 * \mathbf{w} \\ \dots \\ \bar{\mathbf{x}}_n * \mathbf{w} \end{bmatrix}$$

và hiệu $\mathbf{y} - \bar{\mathbf{X}} * \mathbf{w}$ sẽ bằng:

$$\begin{bmatrix} y_1 - \bar{\mathbf{x}}_1 * \mathbf{w} \\ y_2 - \bar{\mathbf{x}}_2 * \mathbf{w} \\ y_3 - \bar{\mathbf{x}}_3 * \mathbf{w} \\ \dots \\ y_n - \bar{\mathbf{x}}_n * \mathbf{w} \end{bmatrix}$$

Như vậy có thể viết gọn $\mathcal{L}(\mathbf{w})$ như sau:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} * \|\mathbf{y} - \bar{\mathbf{X}} * \mathbf{w}\|_2^2$$

Để giải bài toán tối ưu, ta cần tìm nghiệm bằng 0 của đạo hàm của hàm mất mát:

$$\mathcal{L}'(\mathbf{w}) = \sum_{i=1}^{N}(\frac{1}{2}(y_i - \bar{\mathbf{x}}_i * \mathbf{w})^2)'$$

$$= \sum_{i=1}^{N}\frac{1}{2} * 2(y_i - \bar{\mathbf{x}}_i * \mathbf{w}) * (y_i - \bar{\mathbf{x}}_i * \mathbf{w})'$$

$$= \sum_{i=1}^{N}(y_i - \bar{\mathbf{x}}_i * \mathbf{w}) * -(\bar{\mathbf{x}}_i * \mathbf{w})'$$

$$= \sum_{i=1}^{N}(y_i - \bar{\mathbf{x}}_i * \mathbf{w}) * -(\bar{\mathbf{x}_i}^T)$$

$$= \sum_{i=1}^{N}(y_i - \bar{\mathbf{x}}_i * \mathbf{w}) * -(\bar{\mathbf{x}}_i^T)$$

$$= \sum_{i=1}^{N}(\bar{\mathbf{x}}_i * \mathbf{w} - y_i) * \bar{\mathbf{x}_i^T}$$

$$= \bar{\mathbf{X}}^T * (\bar{\mathbf{X}}\mathbf{w} - \mathbf{y})$$

$$= \bar{\mathbf{X}}^T * \bar{\mathbf{X}}\mathbf{w} - \bar{\mathbf{X}}^T * \mathbf{y}$$

như vậy phương trình bằng 0 tương đương với

$$\bar{\mathbf{X}}^T * \bar{\mathbf{X}}\mathbf{w} = \bar{\mathbf{X}}^T * \mathbf{y}$$

TH1: ma trận $\bar{\mathbf{X}}^T * \bar{\mathbf{X}}$ là khả nghịch, suy ra:

$$(\bar{\mathbf{X}}^T * \bar{\mathbf{X}})^{-1} * \bar{\mathbf{X}}^T * \bar{\mathbf{X}} * \mathbf{w} = (\bar{\mathbf{X}}^T * \bar{\mathbf{X}})^{-1} * \bar{\mathbf{X}}^T * \mathbf{y}$$

suy ra:

$$\mathbf{w} = \bar{\mathbf{X}}^T * \bar{\mathbf{X}})^{-1} * \bar{\mathbf{X}}^T * \mathbf{y}$$

TH2: ma trận $\bar{\mathbf{X}}^T * \bar{\mathbf{X}}$ là không khả nghịch thì:

$$\mathbf{w} = (\bar{\mathbf{X}}^T * \bar{\mathbf{X}})^{\dagger} * \bar{\mathbf{X}}^T * \mathbf{y}$$

Code trong python

1. **Đặt bài toán thực tế**
   Anh X đang chuẩn bị ứng tuyển vào công ty A và muốn xem liệu anh nên yêu cầu mức lương bao nhiêu. Là một người trẻ mới có kinh nghiệm x năm trong nghề và ra trường với bằng giỏi, anh X không muốn yêu cầu mức lương cao quá để bị từ chối cũng như không muốn một mức lương quá thấp so với năng lực của mình.
   Nhờ quen biết anh đã xin được bảng lương của mọi người trong công ty cùng số năm kinh nghiệm, trình độ đại học của mỗi người trong nghề.
   Dữ liệu có thể xem tại đây
   Ta có thể giả định số năm trong nghề và tấm bằng sau khi ra trường sẽ tỷ lệ thuận với mức lương nhận được, từ đó ta có thể tạo bài toán

   $$f(x) = w_0 + w_1 * a + w_2 * b$$

   trong đó :

   (a) f(x) là mức lương nhận được

   (b) a là số năm kinh nghiệm

   (c) b là mức bằng đại học, ở đây ta biểu diễn (bằng trung bình, bằng khá, bằng giỏi) bằng (1,2,3)

2. **Python code**
   Xử lý dữ liệu:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 import pandas as pd
4
5 # import the data
6
7 df = pd.read_csv("Salary_Data.csv")
8
9 X = ["YearsExperience","CollegeDegree"]
10
11 Y = ["Salary"]
12 X = df[X]
13 Y = df[Y]
14
```

   Áp dụng công thức để tính toán tìm hệ số w tối ưu:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 import pandas as pd
4
5 # import the data
6
7 df = pd.read_csv("Salary_Data.csv")
8
9 X = ["YearsExperience","CollegeDegree"]
10
11 Y = ["Salary"]
12 X = df[X]
13 Y = df[Y]
14
```

Từ đó ta tính được bộ dữ liệu w bằng [**15413,10050.8,4491.8**]

3. **Kết quả**
   Vậy từ đó anh X có thể đánh giá mức lương của mình theo công thức để đề nghị mức lương hợp lý:

$$f(x) = 15413 + 10050.8 * \textit{số năm kinh nghiệm} + 4491.8 * \textit{mức bằng đại học}$$

   ví dụ anh T có 3 năm kinh nghiệm và bằng đại học mức khá, khi được tuyển vào công ty A anh T sẽ có thể dự đoán mức lương xấp xỉ bằng:

$$f(x) = 15413 + 10050.8 * 3 + 4491.8 * 2 = 54549$$

Nhận xét thuật toán

Ta có thể thấy Linear Regression là một thuật toán ngắn gọn và dễ dàng cài đặt, mặc dù độ chính xác có thể dễ sai lệch khi có bộ dữ liệu chưa tốt: có một số dữ liệu sai khác quá lớn với các dữ liệu khác (ví dụ trong bộ dữ liệu phía trên nếu có nhân viên lương quá cao trong khi không có kinh nghiệm gì và bằng thấp)