Session 3. Linear Regression

I. Introduction

> **Ví dụ**
>
> Một chiếc ô tô có động cơ dung tích x_1 lít, số ghế x_2 và đã đi được x_3 km thì có giá bao nhiêu?

- Giả sử có thống kê từ **1000** chiếc ô tô đã bán trên thị trường, liệu rằng với các thông số trên *ta có thể dự đoán giá* của chiếc ô tô này không?
- Hàm dự đoán: $y = f(x)$ với $x = [x_1, x_2, x_3]$ là vector chứa thông tin input và y là thông tin output.
- Một số mối quan hệ đơn giản có thể nhận thấy:
  1. Dung tích động cơ càng lớn thì giá ô tô thường cao hơn;
  2. Số ghế càng nhiều thì giá ô tô có xu hướng cao hơn;
  3. Số km đã đi càng nhiều thì giá ô tô sẽ giảm.
- Một mô hình đơn giản có thể mô tả mối quan hệ giữa giá ô tô và các thông số đầu vào là:

$$y \approx f(x) = \hat{y}$$
$$f(x) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$$

  trong đó: w_1, w_2, w_3 là các hệ số trọng số và w_0 là giá trị bias.
- Mối quan hệ $y \approx f(x)$ bên trên là một mối quan hệ tuyến tính (linear).
- Bài toán này là một bài toán thuộc loại regression: đi tìm các hệ số tối ưu $\{w_1, w_2, w_3, w_0\}$ chính vì vậy được gọi là bài toán **Linear Regression**.
- **Các chú ý**:
  - y là giá trị thực (dựa trên số liệu thống kê chúng ta có trong tập *training data*), trong khi $\hat{y}$ là giá trị mà mô hình Linear Regression dự đoán được. Nhìn chung, y và $\hat{y}$ là hai giá trị khác nhau do có sai số mô hình $\rightarrow$ mong muốn rằng sự khác nhau này rất nhỏ.
  - *Linear* hay *tuyến tính* hiểu một cách đơn giản là *thẳng, phẳng*.

- Trong không gian hai chiều, một hàm số được gọi là *tuyến tính* nếu đồ thị của nó có dạng một *đường thẳng*.
- Trong không gian ba chiều, một hàm số được goi là *tuyến tính* nếu đồ thị của nó có dạng một *mặt phẳng*.
- Trong không gian nhiều hơn 3 chiều, khái niệm *mặt phẳng* không còn phù hợp nữa, thay vào đó, một khái niệm khác ra đời được gọi là *siêu phẳng* (*hyperplane*).
- Các hàm số tuyến tính là các hàm đơn giản nhất, vì chúng thuận tiện trong việc hình dung và tính toán.

II. Toán học

1. Linear Regression

- Đặt:
  - $w = [w_0, w_1, w_2, w_3]^T$ là vector (cột) hệ số cần phải tối ưu;
  - $\bar{x} = [1, x_1, x_2, x_3]$ là vector (hàng) dữ liệu đầu vào mở rộng.
  - Số 1 ở đầu được thêm vào để phép tính đơn giản hơn và thuận tiện cho việc tính toán.
- Khi đó ta được phương trình:

$$y \approx \hat{y} = \bar{x}w$$

2. Sai số dự đoán

- Mong muốn rằng *sự sai khác* e (error) giữa giá trị thực y và giá trị dự đoán $\hat{y}$ nhỏ nhất. Tương ứng:

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

  hệ số $\dfrac{1}{2}$ để triệt tiêu trong quá trình đạo hàm.
- Ta cần giá trị e^2 nhỏ nhất, thay vì nói e nhỏ nhất do e có thể âm.

3. Hàm mất mát

- Điều tương tự xảy ra với tất cả các cặp (input, outcome) $(\mathrm{x}_i, y_i), i = 1, 2, \ldots, N$ với N là số lượng dữ liệu quan sát được.
- Mong muốn: *tổng sai số là nhỏ nhất*, tương đương với việc tìm w để hàm số sau đạt GTNN:

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

- Hàm số $\mathcal{L}(w)$ ở trên gọi là **hàm mất mát** của bài toán Linear Regression, yêu cầu của ta là sai số này nhỏ nhất $\rightarrow$ tìm vector hệ số w : gọi là *điểm tối ưu* (optimal point)

$$\mathrm{w}^* = \arg\min_{\mathrm{w}} \mathcal{L}(\mathrm{w})$$

- Trước khi đi đến lời giải, ta đơn giản hóa hàm mất mát ở trên với:
  - $\bar{\mathrm{X}} = [\bar{\mathrm{x}}_1, \bar{\mathrm{x}}_2, \ldots, \bar{\mathrm{x}}_N]$: ma trận đầu vào, mỗi dòng là một điểm dữ liệu;
  - $\mathrm{y} = [y_1, y_2, \ldots, y_N]$: vector cột chứa output;
- Ta được:

$$\mathcal{L}(\mathrm{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \bar{\mathrm{x}}_i\mathrm{w})^2 = \frac{1}{2}\|\mathrm{y} - \bar{\mathrm{X}}\mathrm{w}\|_2^2$$

với $\|\mathrm{z}\|_2$ là **Euclidean Norm** (chuẩn Euclid - khoảng cách Euclid), nói cách khác $\|z\|_2^2$ là *tổng bình phương* mỗi phần tử trong vector z.

4. Nghiệm cho bài toán Linear Regression

- Cách tiếp cận đơn giản từ trước là giải phương trình đạo hàm (gradient) bằng 0 - không quá phức tạp và với phương trình tuyến tính thì khả thi.
- Đạo hàm theo w của hàm mất mát:

$$\frac{\partial\mathcal{L}(\mathrm{w})}{\partial\mathrm{w}} = \bar{\mathrm{X}}^T(\bar{\mathrm{X}}\mathrm{w} - y)$$

- Đạo hàm vector: [Source]
- Phương trình đạo hàm trên tương đương với:

$$\bar{\mathrm{X}}^T\bar{\mathrm{X}}\mathrm{w} = \bar{\mathrm{X}}^T y \triangleq \mathrm{b}$$

với $\bar{\mathrm{X}}^T y \triangleq \mathrm{b}$ tức là đặt $\mathrm{b} = \bar{\mathrm{X}}^T y$.

- Nếu ma trận $\mathbf{A} \triangleq \bar{\mathbf{X}}^T\bar{\mathbf{X}}$ khả nghịch (non-singular hay invertible) thì phương trình trên có nghiệm duy nhất:

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{b}$$

  ngược lại, $\mathbf{A}$ không khả nghịch (Det bằng 0) thì phương trình vô nghiệm/ vô số nghiệm.
- Ta sử dụng khái niệm giả nghịch đảo $\mathbf{A}^\dagger$ (A dagger), khi đó, *điểm tối ưu cho bài toán Linear Regression* :

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