ITF: **Information Technology Faculty**
Thuyloi University

Hồi quy tuyến tính

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Bài giảng được dựa trên giáo trình machine learning cơ bản

Giới thiệu

Xét ví dụ:

- G/S chúng ta đã có số liệu thống kê từ 1000 căn nhà trong thành phố một căn nhà rộng x_1 m2, có x_2 phòng ngủ và cách trung tâm thành phố x_3 km có giá là bao nhiêu?
- Hàm dự đoán $\mathrm{y} = f(\mathrm{x})$ có dạng thế nào?
- Ở đây $\mathrm{x} = [x_1, x_2, x_3]$ là một vector hàng chứa thông tin *input*, y là một số vô hướng (scalar) biểu diễn *output*.

Giới thiệu

- Chúng ta thấy:
  - diện tích nhà càng lớn thì giá nhà càng cao
  - số lượng phòng ngủ càng lớn thì giá nhà càng cao
  - càng xa trung tâm thì giá nhà càng giảm
- Một hàm số có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào:

$$\mathrm{y} \approx f(\mathrm{x}) = \hat{\mathrm{y}}$$

$$f(\mathrm{x}) = w_1\mathrm{x}_1 + w_2\mathrm{x}_2 + w_3\mathrm{x}_3 + w_0$$

  - Mối quan hệ y≈f(x) là một mối quan hệ tuyến tính (linear).
  - Bài toán trên là bài toán thuộc loại regression. Do đó, bài toán đi tìm các hệ số tối ưu $\{w_1, w_2, w_3, w_0\}$ được gọi là bài toán Linear Regression.

Giới thiệu

- y là giá trị thực của *outcome* (dựa trên số liệu thống kê chúng ta có trong tập *training data*),
- $\hat{y}$ là giá trị mà mô hình Linear Regression dự đoán được.
- y và $\hat{y}$ là hai giá trị khác nhau do có sai số mô hình, tuy nhiên, chúng ta mong muốn rằng sự khác nhau này rất nhỏ.
- *Linear* 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, là *siêu mặt phẳng* (*hyperplane*).

Dạng của Linear Regression

- Trong phương trình: $f(\mathrm{x}) = w_1\mathrm{x}_1 + w_2\mathrm{x}_2 + w_3\mathrm{x}_3 + w_0$, nếu chúng ta đặt
  - $\mathrm{w} = [w_1, w_2, w_3, w_0]^T$ là vector hệ số cần phải tối ưu
  - $\mathrm{x} = [1, \mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3]$ là vector (hàng) dữ liệu đầu vào
- Phương trình trên có thể được viết lại dưới dạng:

$$\mathrm{y} \approx \mathrm{xw} = \hat{\mathrm{y}}$$

Sai số dự đoán

- Với một cặp (x_i, y_i) $i = 1,2 \dots N$, chúng ta muốn sự sai khác e_i giữa giá trị thực y_i và giá trị dự đoán $\hat{y}_i$ là nhỏ nhất:

$$\frac{1}{2}{e_i}^2 = \frac{1}{2}(y_i - \hat{y}_i)^2 = \frac{1}{2}(y_i - x_i w)^2$$

- Hệ số $\frac{1}{2}$ là để thuận tiện cho việc tính toán (khi tính đạo hàm thì số $\frac{1}{2}$ sẽ bị triệt tiêu).

- Chúng ta cần${e_i}^2$vì $e_i = y_i - \hat{y}_i$ có thể là một số âm, việc nói e_i nhỏ nhất sẽ không đúng vì khi $e_i = -\infty$ là rất nhỏ nhưng sự sai lệch là rất lớn.ss

Hàm mất mát

- Chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm w để hàm số sau đạt giá trị nhỏ nhất:

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

- Hàm số $\mathcal{L}(w)$ được gọi là **hàm mất mát** (loss function) của bài toán Linear Regression.

- Chúng ta cần tìm vector hệ số w sao cho giá trị của hàm mất mát này càng nhỏ càng tốt

$$\mathrm{w}^* = \underset{w}{\operatorname{argmin}}\,\mathcal{L}(w)$$

Hàm mất mát

- Trước khi đi tìm lời giải, chúng ta đơn giản hóa phép toán trong phương trình hàm mất mát:

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

- Đặt
  - $\mathrm{y} = [y_1; y_2; \ldots y_N]$ là một véc tơ cột chứa tất cả các output của dữ liệu huấn luyện
  - $\mathrm{X} = [\mathrm{x}_1; \mathrm{x}_2; \ldots \mathrm{x}_N]$ là ma trận dữ liệu đầu vào mà mỗi hàng là một điểm dữ liệu
- Khi đó, hàm $\mathcal{L}(w)$ được viết thành:

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

$$= \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\|_2^2$$

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! Nhưng chỉ trường hợp
  - tính đạo hàm và
  - việc giải phương trình đạo hàm bằng 0

  không quá phức tạp.
- Thật may, với các mô hình tuyến tính, hai việc này là khả thi
- Đạo hàm theo w của hàm mất mát là:

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

Nghiệm cho bài toán Linear Regression

- Phương trình đạo hàm bằng 0 tương đương với:
- $$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y} \quad (*)$$
- Đặt $\mathbf{A} \triangleq \mathbf{X}^T\mathbf{X}$ và $\mathbf{X}^T\mathbf{y} \triangleq \mathbf{b}$
- Nếu ma trận vuông A khả nghịch thì phương trình (*) có nghiệm duy nhất

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