Overfitting

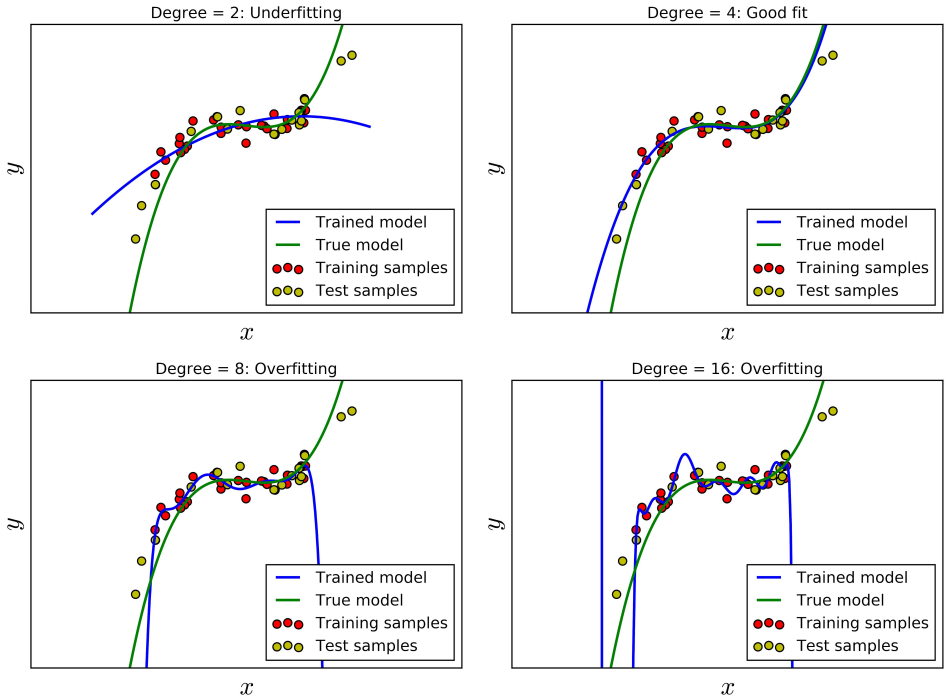
FIC - Ta Dang Khoa

Nội dung

- Underfitting vs overfitting
- Giải pháp
- Cách chia tập data

Underfitting vs overfitting

Bài toán



Underfitting vs overfitting

Giải pháp

- Underfitting:
  - Tăng độ phức tạp của mô hình
  - Tăng số vòng lặp
- Overfitting:
  - Thêm dữ liệu
    - Thu thập thêm dữ liệu
    - Biến đổi, sinh thêm dữ liệu từ những dữ liệu đã có
  - Giảm độ phức tạp của mô hình
    - Giảm số tham số (giảm độ phức tạp của mô hình)
    - Early Stopping
    - Thêm các đại lượng phạt vào hàm loss: **regularized loss function**

Regularization

- Hiệu chỉnh L2:

$$L(x,y) = \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2$$

where $h_\theta x_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3 + \theta_4 x_4^4$

$$L(x,y) \equiv \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_i^2$$

- Thuật toán Gradient descent khi có hiệu chỉnh L2:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \qquad \text{for } j = 0$$

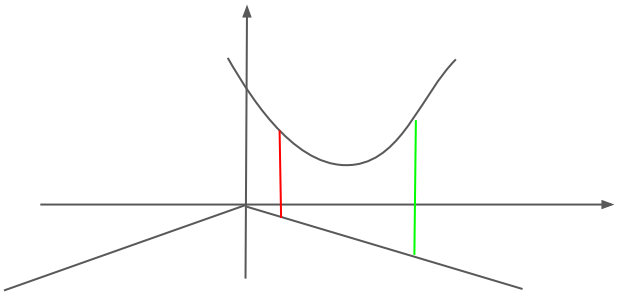
$$\frac{\partial J(\theta)}{\partial \theta_j} = \left(\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \qquad \text{for } j \geq 1$$

Regularization

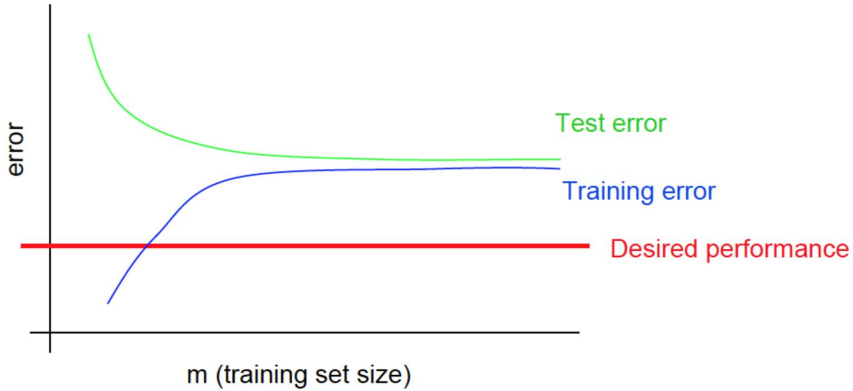
Ý nghĩa

- Việc tối thiểu hàm loss có sử dụng regularization khiến các tham số (w) nhỏ gần với 0. (Có thể giúp làm nhỏ các hệ số của tham số bậc cao, làm giảm overfitting)
- Góc nhìn hình học:

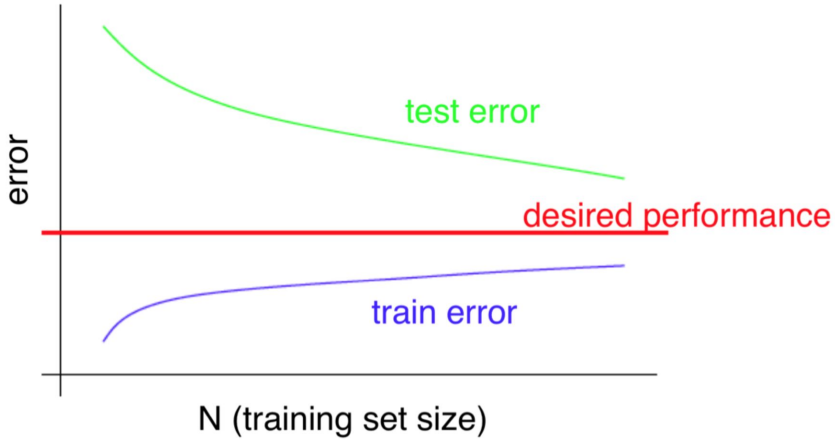
$$J(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 - (-\lambda\|\mathbf{w}\|)$$



Thu thập thêm dữ liệu



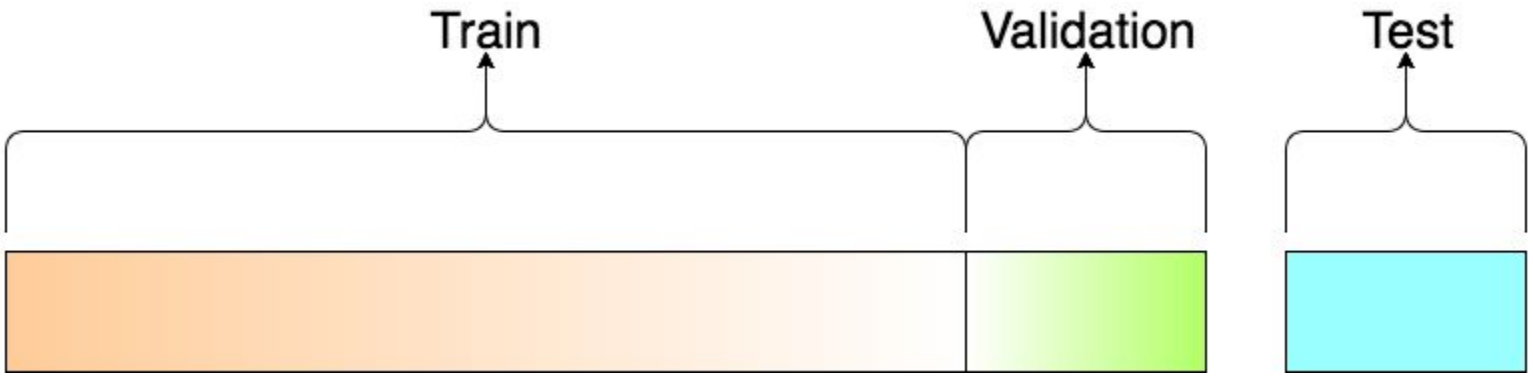
High bias



High variance

Cách chia tập data

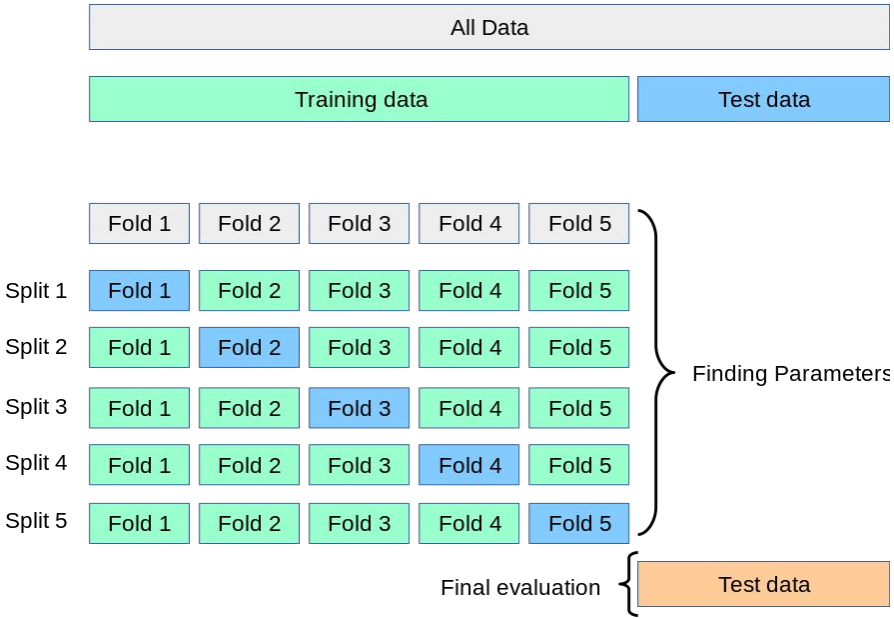
- Thường thì chúng ta sẽ chia tập data thành 3 phần: Train, Validation, Test
- Lý do cần thêm tập validation: tránh overfit trên cả tập test



Cross-Validation

K-Fold CV

Chia tập dữ liệu thành k phần bằng nhau, mỗi lần sử dụng 1 phần làm tập test và k-1 phần làm dữ liệu training (Thường được sử dụng khi tập data ít)



Bài tập

- Thực hiện bài toán sử dụng Regularization và K-Fold
- Tìm hiểu thêm về Feature Engineering (cụ thể là Feature Scaling and Normalization)
- Tìm hiểu về các thuật toán Gradient descent cải tiến, tìm hiểu về batch size