Overfitting, regularization, và kiểm tra mô hình

Ngô Minh Nhựt

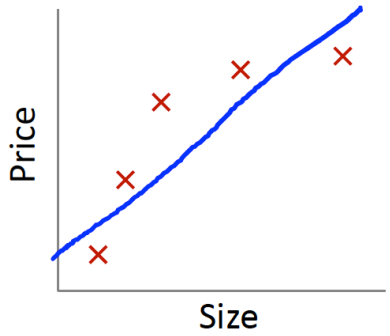
Bộ môn Công nghệ Tri thức

2024

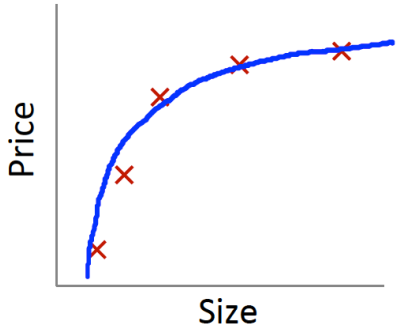
Nội dung

- Mô hình quá khớp dữ liệu (overfitting)
- Bình thường hóa tham số (regularization)
- Kiểm tra mô hình (model validation)

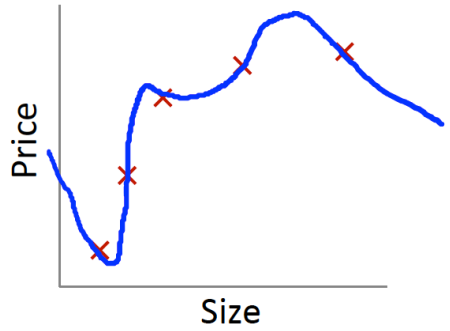
Mô hình quá khớp dữ liệu



$\theta_0 + \theta_1 x$



$\theta_0 + \theta_1 x + \theta_2 x^2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

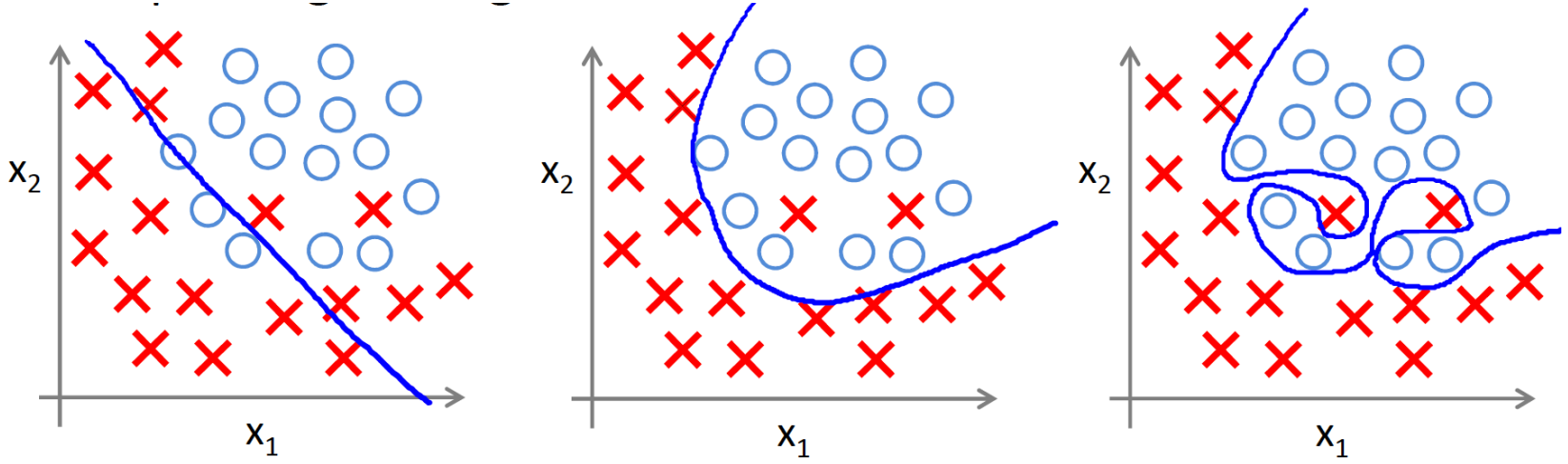
Khi số lượng đặc trưng lớn:

- Mô hình có thể quá khớp với dữ liệu học (chi phí ~ 0)
- Không thể tổng quát hóa cho dữ liệu mới (chi phí >> 0)

Source: Andrew Ng

Mô hình quá khớp dữ liệu

- Hồi qui logistic



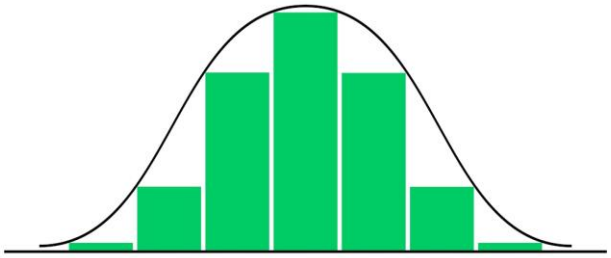
Nguồn: Andrew Ng

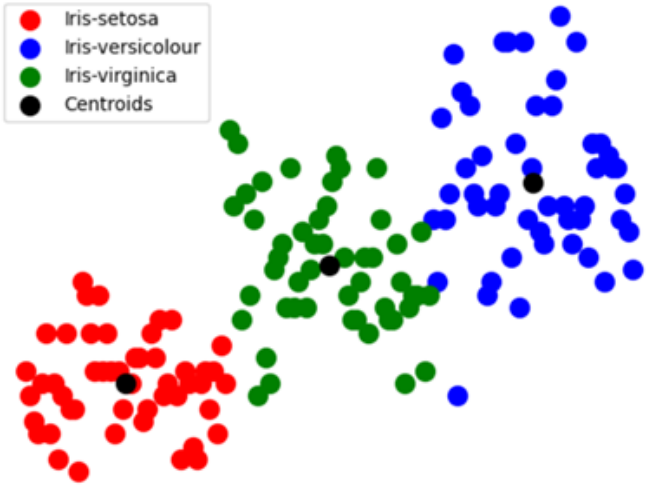
Mô hình quá khớp dữ liệu

- Giải pháp:
  - Giảm số đặc trưng
    - Lựa chọn đặc trưng qua phân tích dữ liệu
    - Lựa chọn đặc trưng qua đánh giá mô hình
  - Bình thường hóa tham số
    - Giảm mức ảnh hưởng của tham số đối với output
    - Tốt trong trường hợp các đặc trưng chi phối nhỏ tới output

Lựa chọn đặc trưng

- Tìm hiểu cấu trúc, phân bố, mối liên hệ giữa input và output
  - Các công cụ như histogram, scatter plot có thể giúp ích





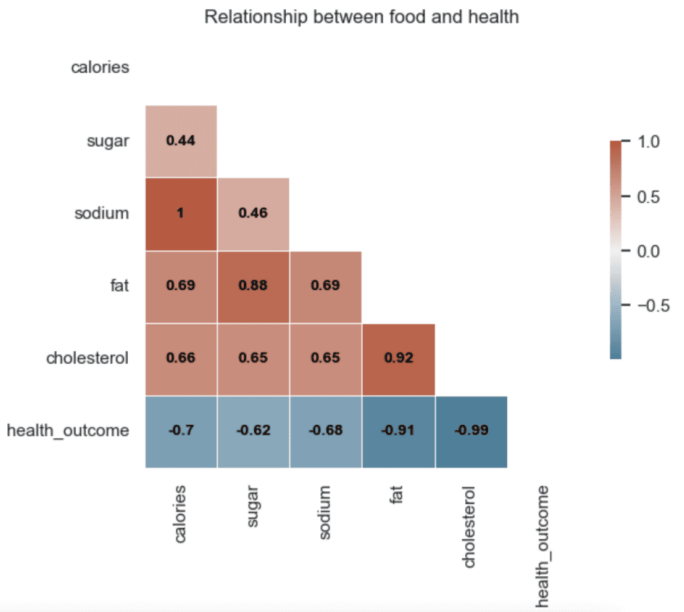
Source: Internet

Lựa chọn đặc trưng

- Đánh giá tầm quan trọng của đặc trưng dùng correlation
  - Áp dụng cho quan hệ tuyến tính
  - Xác định correlation giữa đặc trưng với nhau
  - Xác định correlation giữa đặc trưng và output
- Các đặc trưng có correlation với nhau cao được xem xét loại bỏ
- Đặc trưng có correlation với output cao có tầm quan trọng cao hơn

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

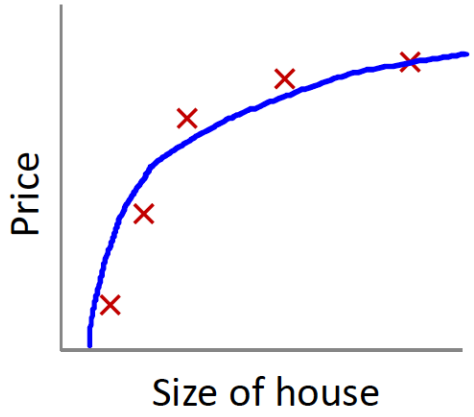


Source: Internet

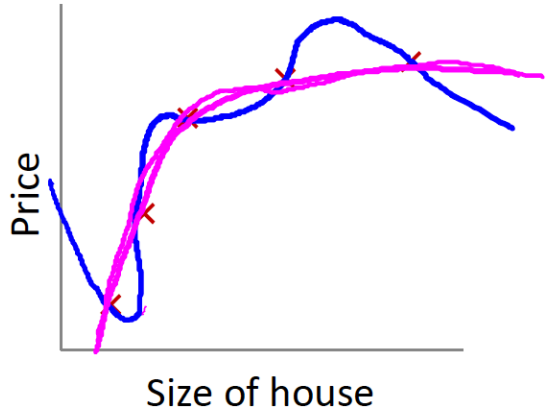
Lựa chọn đặc trưng

- Các bước thực hiện
  - Tìm hiểu cấu trúc, phân bố, mối liên hệ giữa input và output
  - Đánh giá tầm quan trọng của đặc trưng dùng correlation
- Đánh giá mô hình
  - Kiểm tra mô hình với đặc trưng được chọn
  - Lặp lại quá trình lựa chọn cho tới khi mô hình đủ tốt

Bình thường hóa tham số



$$\theta_0 + \theta_1 x + \theta_2 x^2$$



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + 1000\theta_3^2 + 1000\theta_4^2$$

Source: Andrew Ng

Bình thường hóa tham số

- Giá trị tham số nhỏ hơn
  - Tạo ra mô hình đơn giản hơn
  - Giảm overfitting

Bình thường hóa tham số

- Hàm chi phí

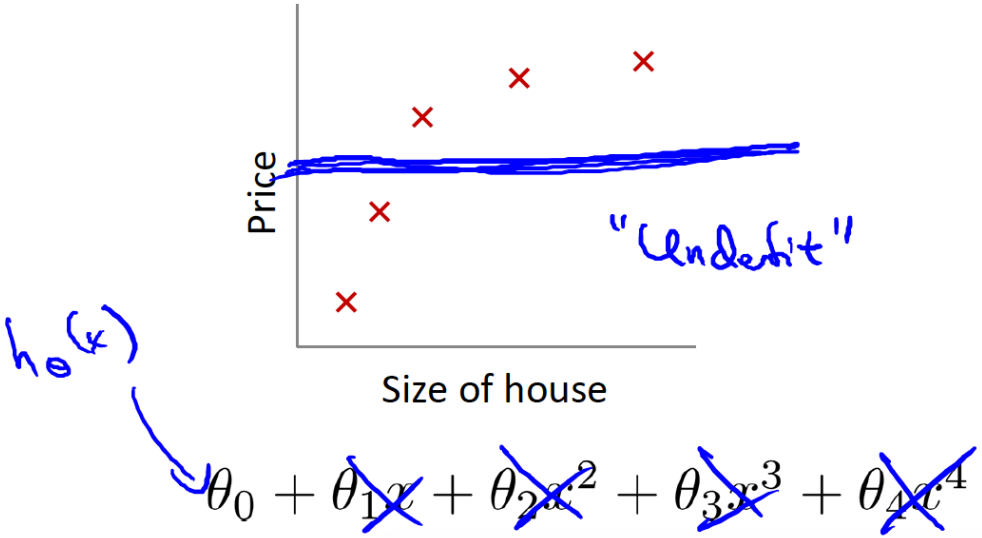
$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

- Lambda: weight decay
- Lambda càng lớn, mô hình càng đơn giản

Bình thường hóa tham số

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

Nếu lambda quá lớn?



Source: Andrew Ng

Bình thường hóa tham số

- Vector gradient

$$\frac{dJ}{d\theta_0} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}$$

$$\frac{dJ}{d\theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)} + \frac{\lambda}{m}\theta_j$$

Thuật toán hạ dốc gradient hoạt động bình thường

Bình thường hóa tham số

- Hồi qui logistic - hàm chi phí

$$J(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

Bình thường hóa tham số

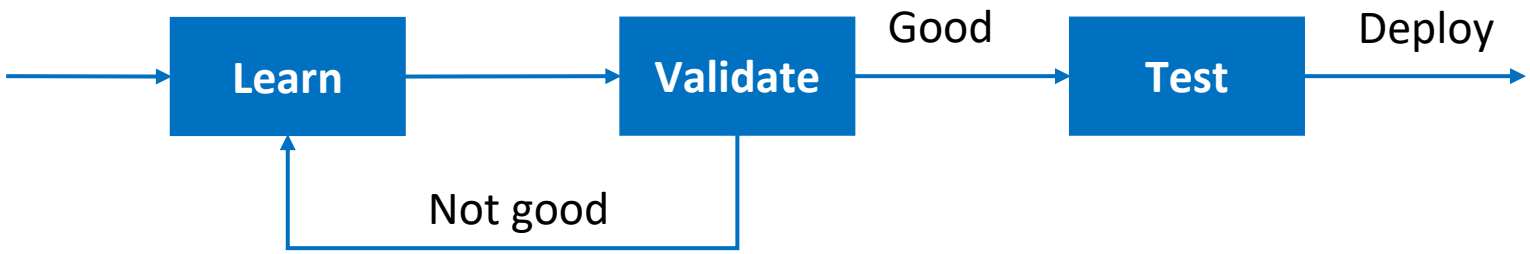
- Hồi qui logistic - vector gradient

$$\frac{dJ}{d\theta_0} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}$$

$$\frac{dJ}{d\theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)} + \frac{\lambda}{m}\theta_j$$

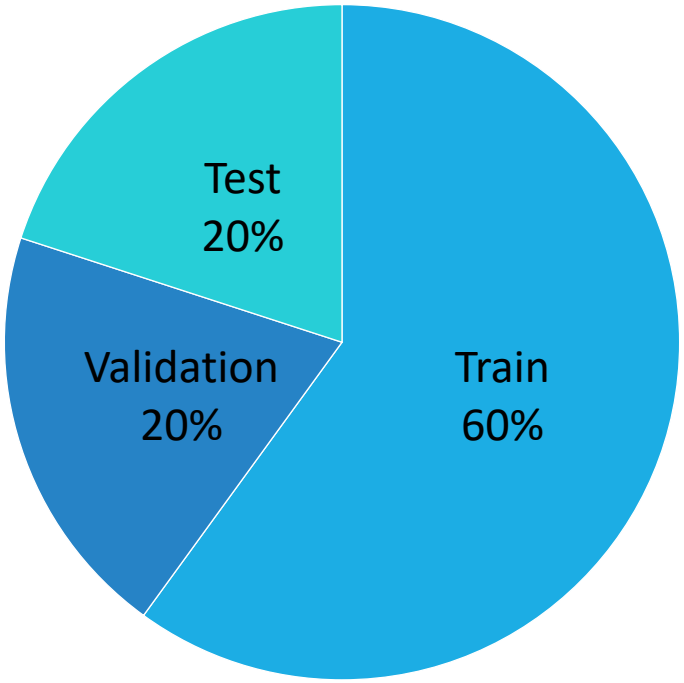
Kiểm tra mô hình

- Quá trình học làm cho mô hình khớp với dữ liệu học
- Mô hình học được có tổng quát hóa cho dữ liệu mới?
  - Kiểm tra mô hình với dữ liệu chưa nhìn thấy
- Giải pháp:
  - Học, kiểm tra, và chọn mô hình tốt nhất
  - Kiểm tra mô hình này có tốt cho dữ liệu mới khác



Tập dữ liệu

- Chia dữ liệu thành 3 tập: train, validation, và test
- Khi tập dữ liệu lớn: 60% : 20% : 20%



Hàm chi phí

Train error

$$J_{\text{train}}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Cross-validation error

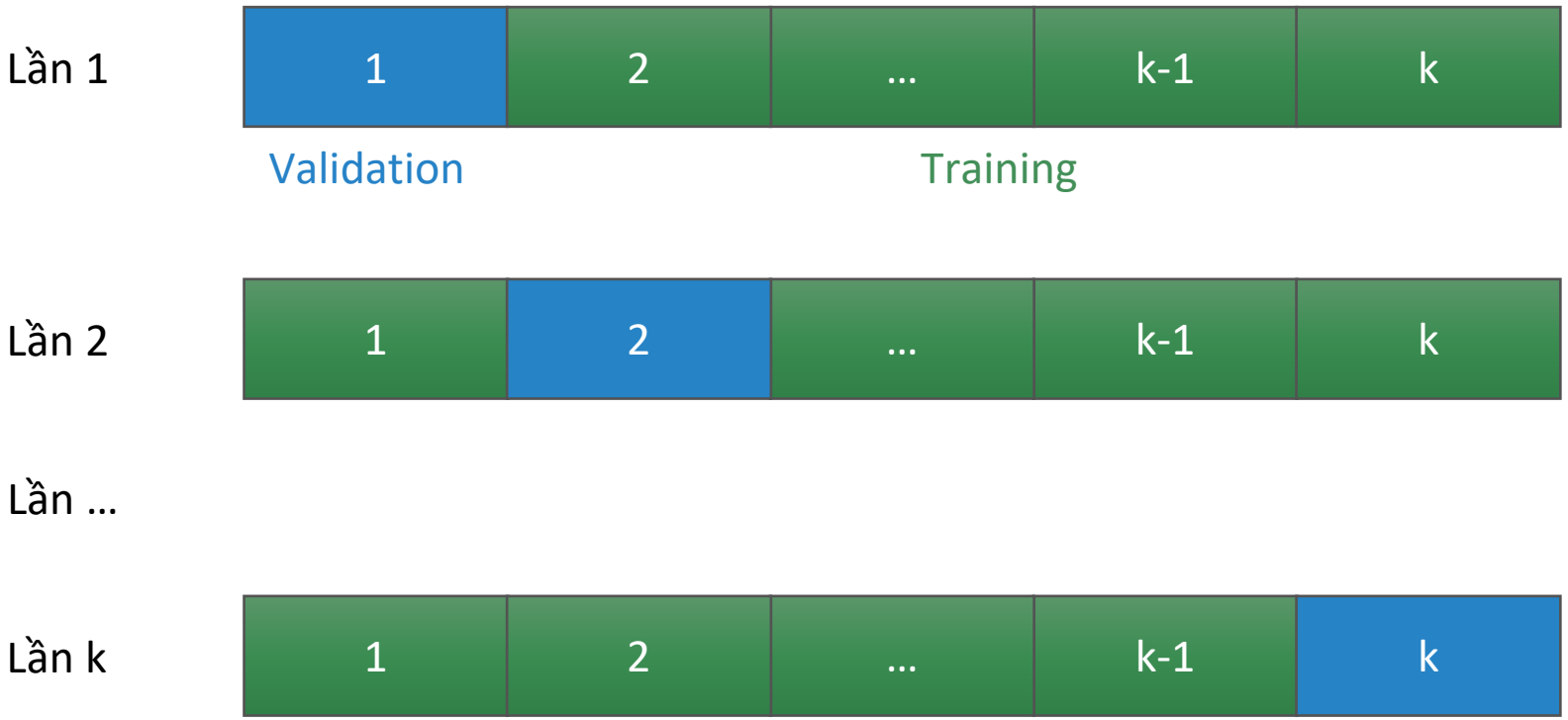
$$J_{\text{cv}}(\theta) = \frac{1}{2m_{\text{cv}}} \sum_{i=1}^{\text{cv}} \left(h_\theta(x_{\text{cv}}^{(i)}) - y_{\text{cv}}^{(i)} \right)^2$$

Test error

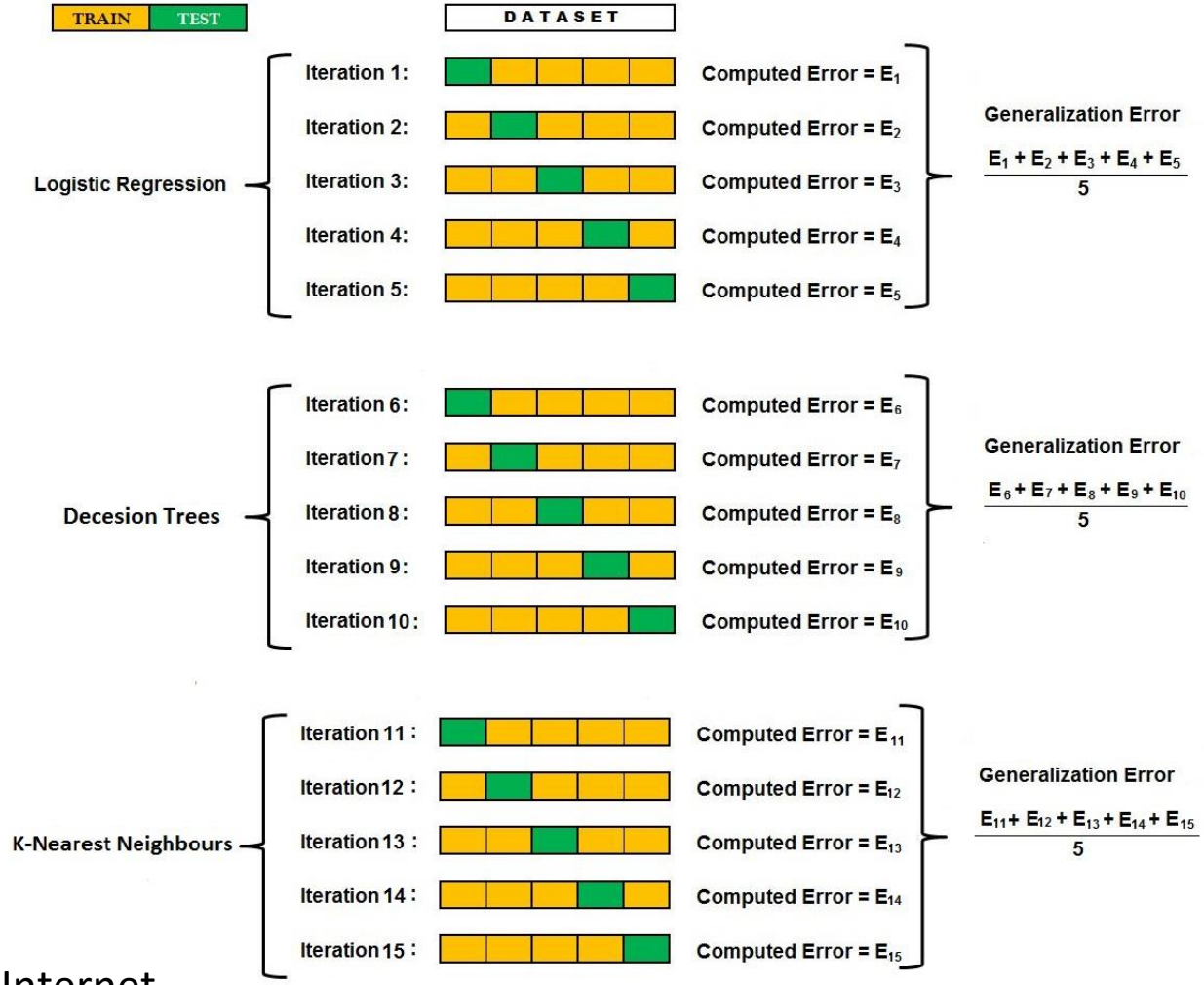
$$J_{\text{test}}(\theta) = \frac{1}{2m_{\text{test}}} \sum_{i=1}^{\text{test}} \left(h_\theta(x_{\text{test}}^{(i)}) - y_{\text{test}}^{(i)} \right)^2$$

k-fold cross validation

- Chia ngẫu nhiên dữ liệu thành k phần
- Học và kiểm tra k lần. k thường là 10



k-fold cross validation

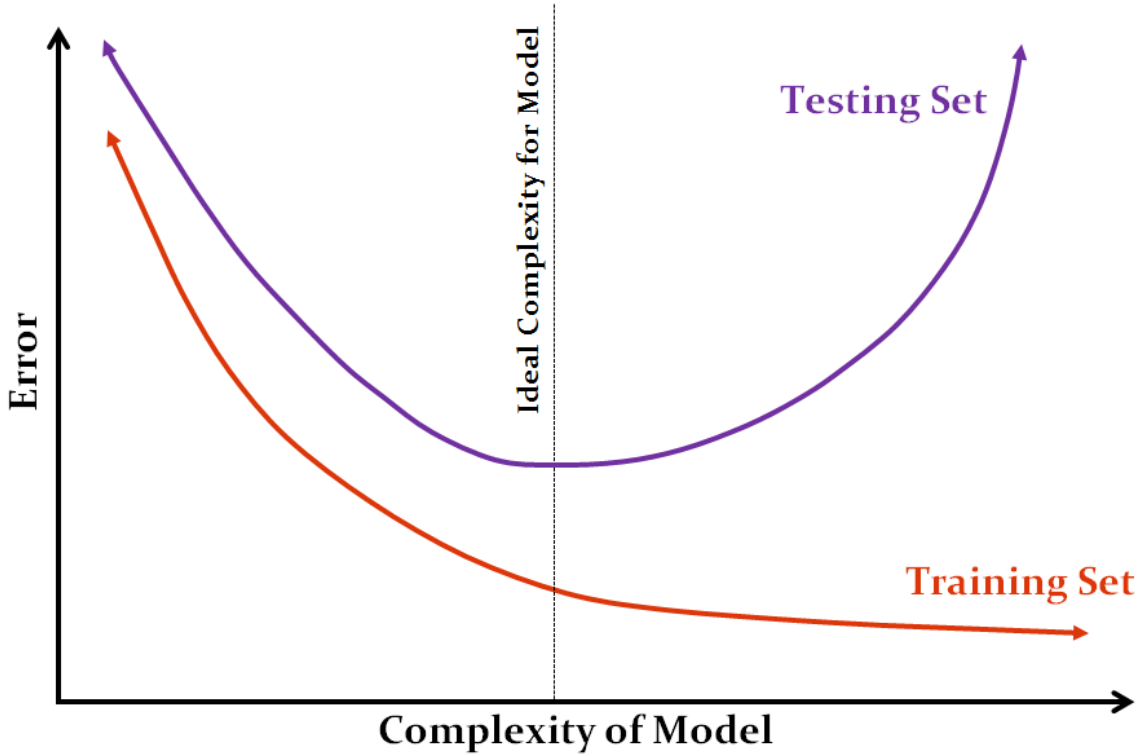


Nguồn: Internet

Hiệu chỉnh mô hình

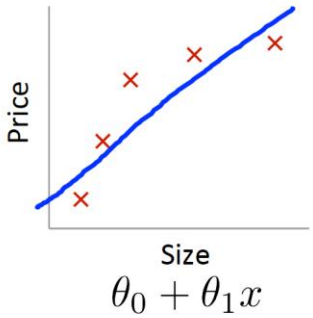
- Mô hình học có nhiều tham số
- Thuật toán học có nhiều siêu tham số (hyper parameter)
- Làm gì khi thuật toán hoạt động không tốt?
  - Thu thập thêm dữ liệu
  - Rút gọn tập đặc trưng
  - Thử với đặc trưng khác
  - Đổi mô hình
  - Giảm weight decay
  - Tăng weight decay

Hiệu chỉnh mô hình



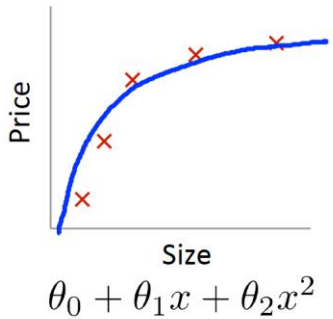
Bias và variance

- Bias: lỗi của mô hình trên tập học
- Variance: độ lệch giữa lỗi của mô hình trên tập đánh giá và lỗi của mô hình trên tập học
- Hiệu chỉnh đến khi bias và variance đều đạt cực tiểu



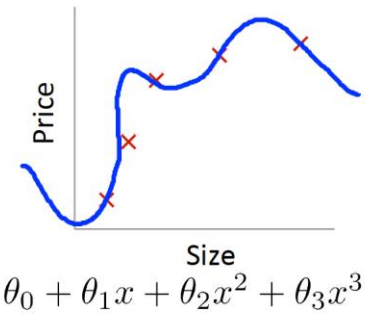
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

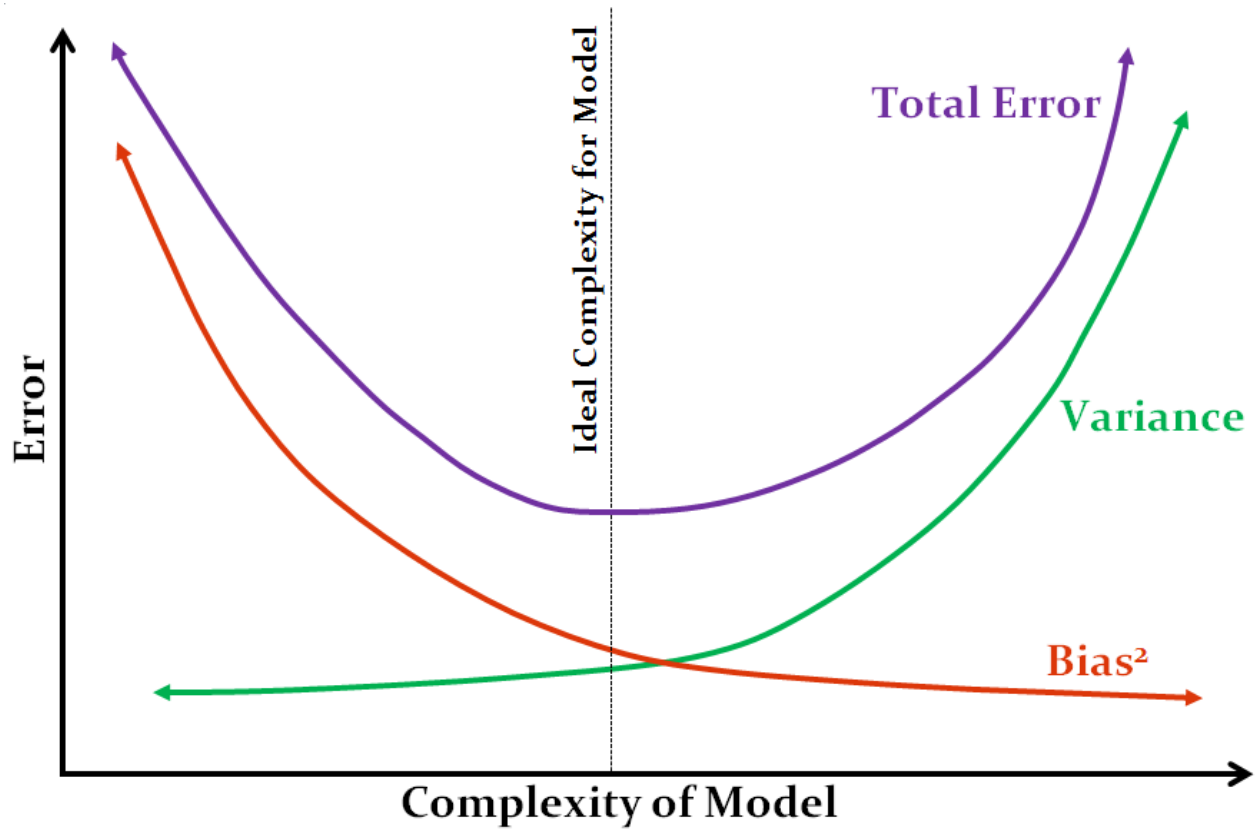
"Just right"



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

Bias và variance

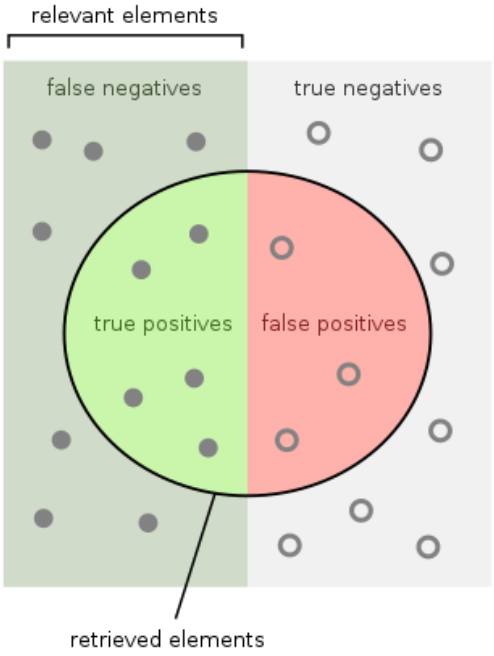


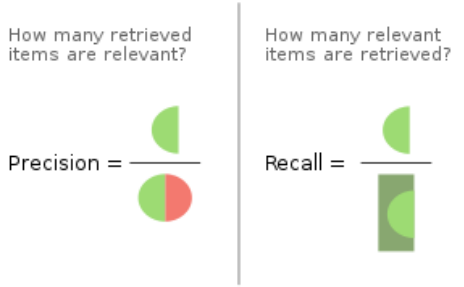
Đánh giá mô hình

- Một số độ đo để đánh giá mô hình
  - Precision, recall
  - Accuracy
  - F1-score
- Tùy vào bài toán, chúng ta cần chọn thước đo phù hợp

Đánh giá mô hình

- $\text{Precision} = \dfrac{\text{True positive}}{\text{Predicted positive}}$
- $\text{Recall} = \dfrac{\text{True positive}}{\text{Positive}}$
- $\text{F1-score} = \dfrac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$
- $\text{Accuracy} = \dfrac{\text{True positive} + \text{True negative}}{\text{Positive} + \text{Negative}}$





Nguồn: Wikipedia

Confusion matrix

		Predicted				
		Cat	Dog	Tiger	Wolf	Total
Actual	Cat	**6**	0	3	1	10
	Dog	2	**4**	0	4	10
	Tiger	3	3	**3**	0	9
	Wolf	1	4	1	**2**	8
	Total	12	11	7	7	

>>> sklearn.metrics.classification_report

>>> y_true = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …]

>>> y_pred = [0, 0, 0, 0, 0, 0, 2, 2, 2, 3, …]

>>> target_names = ['Cat', 'Dog', 'Tiger', 'Wolf']

	precision	recall	f1-score	support
Cat	0.500	0.600	0.545	10
Dog	0.364	0.400	0.381	10
Tiger	0.429	0.333	0.375	9
Wolf	0.286	0.250	0.267	8
accuracy			0.405	37
macro avg	0.394	0.396	0.392	37
weighted avg	0.399	0.405	0.399	37