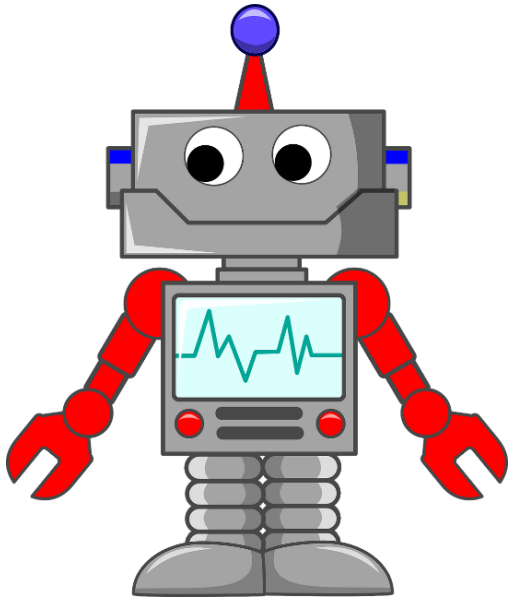
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

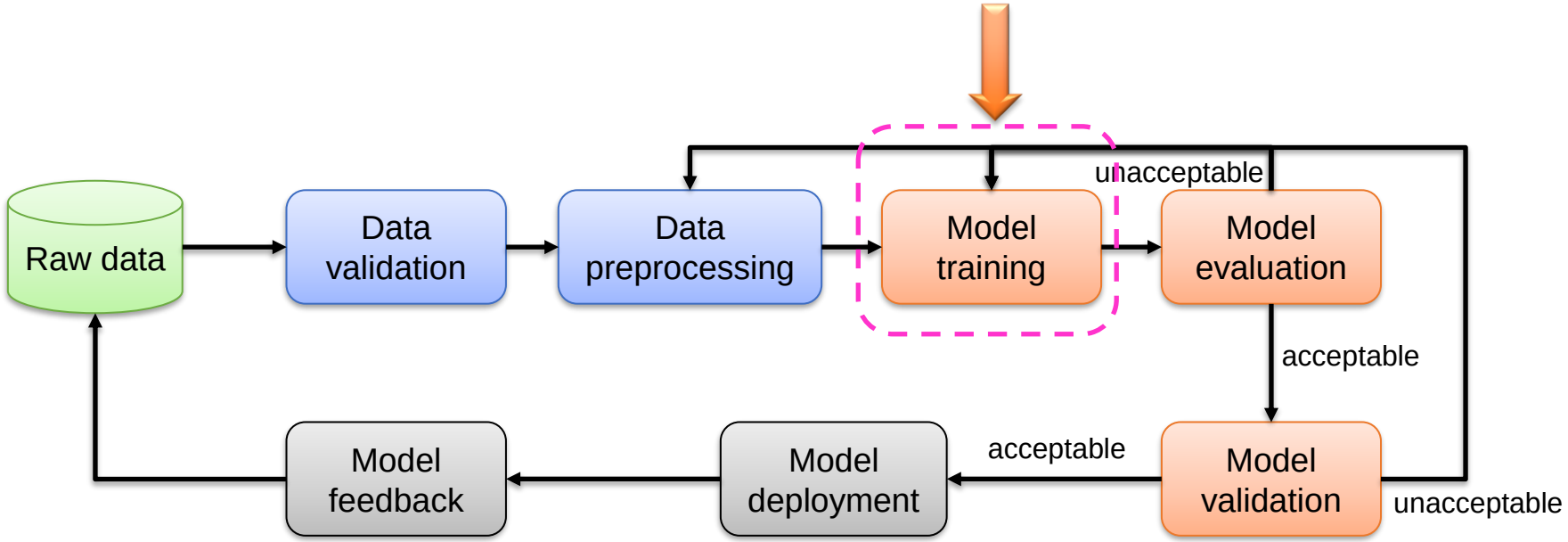
BÀI 07

HỌC CÓ GIÁM SÁT – MÔ HÌNH HỒI QUY (REGRESSION)



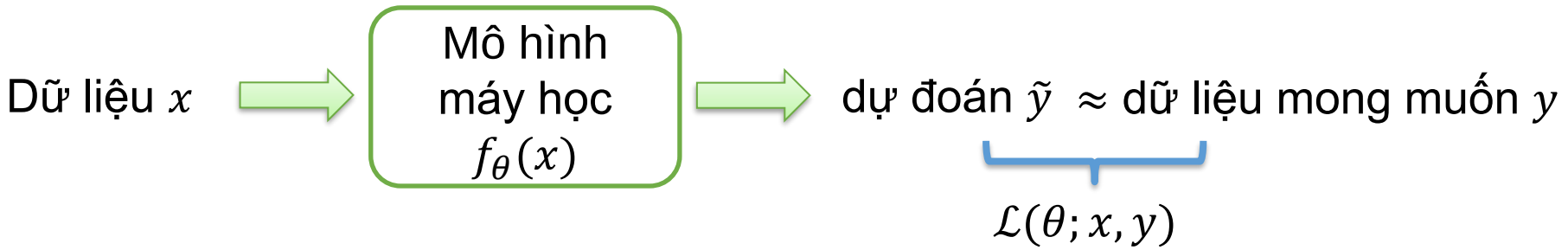
TS. Nguyễn Vinh Tiệp

Vị trí của bài học



Học có giám sát

- **Học có giám sát (supervised learning)**: là một nhánh của máy học, nhằm dự đoán giá trị đầu ra từ một đặc trưng đầu vào dựa trên các dữ liệu huấn luyện trước đó
- Dữ liệu huấn luyện bao gồm cặp đặc trưng đầu vào và giá trị đầu ra mong muốn (x, y)

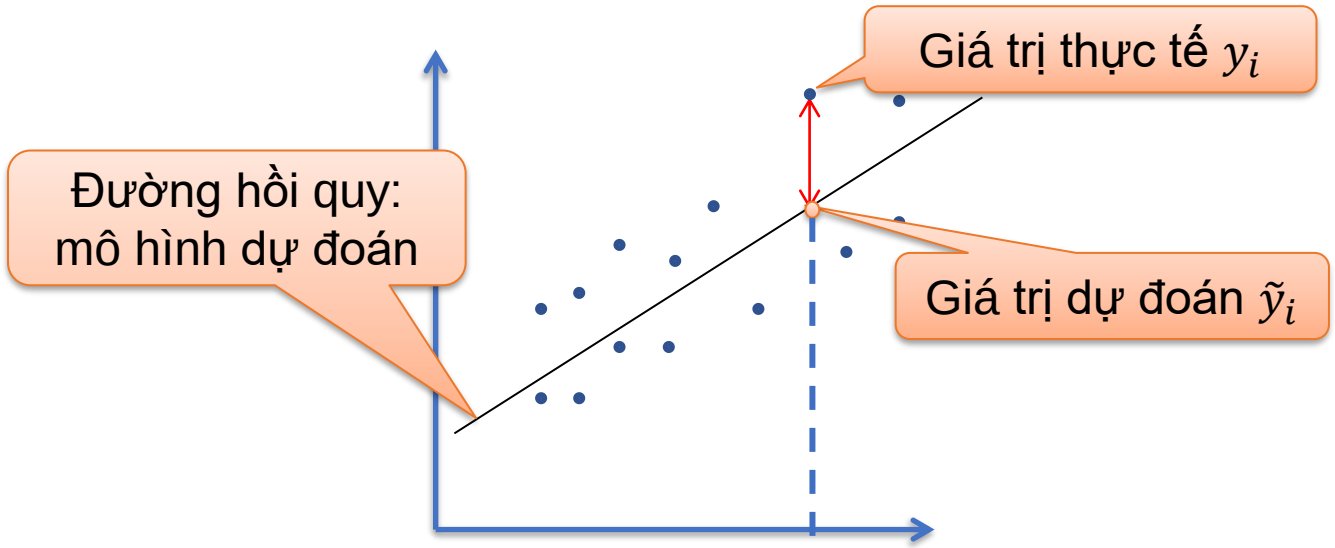


- Có hai loại bài toán chính: hồi quy và phân lớp

NỘI DUNG

Mô hình hồi quy tuyến tính – Linear Regression

- Mô hình thực tế: $y = \varepsilon + \beta x$
- Hàm mô hình dự đoán: $\hat{y} = \hat{\beta} x$
- Hàm độ lỗi: $\mathrm{MSE} = \mathrm{L}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta} x_i)^2 = \|Y - \hat{\beta} X\|^2$



Mô hình hồi quy tuyến tính – Linear Regression

- Huấn luyện mô hình:
  - Bằng normal equation: tham số ước lượng: $\hat{\beta} = (X^T X)^{-1}(X^T Y)$
  - Bằng thuật toán gradient descent với công thức cập nhật:

$$\hat{\beta} = \hat{\beta} - \alpha \nabla_{\hat{\beta}} L$$

- Trong sklearn đã cài đặt module: **LinearRegression**

Mô hình hồi quy tuyến tính – Linear Regression

- **Ưu điểm:**
  - Đơn giản, dễ hiểu
  - Phù hợp với dữ liệu có mối quan hệ tuyến tính (đồng biến, nghịch biến)

- **Khuyết điểm:**
  - Không hiệu quả khi dữ liệu có mối quan hệ phức tạp
  - Dễ bị ảnh hưởng khi có dữ liệu nhiễu

NỘI DUNG

1. MÔ HÌNH HỒI QUY TUYẾN TÍNH

2. BIAS VÀ VARIANCE

3. MÔ HÌNH LASSO, RIDGE VÀ ELASTIC NET

4. MỘT SỐ MÔ HÌNH PHI TUYẾN

Khái niệm Bias

- **Bias**: là sai số giữa trung bình (kỳ vọng) mô hình dự đoán với mô hình thực tế:

$$\text{bias}(\hat{\beta}) = E(\hat{\beta}) - \beta$$

- **Bias thấp**: thể hiện mô hình học được mối quan hệ của dữ liệu huấn luyện
- **Bias cao:** thể hiện mô hình không học được quan hệ của dữ liệu huấn luyện

Khái niệm Variance

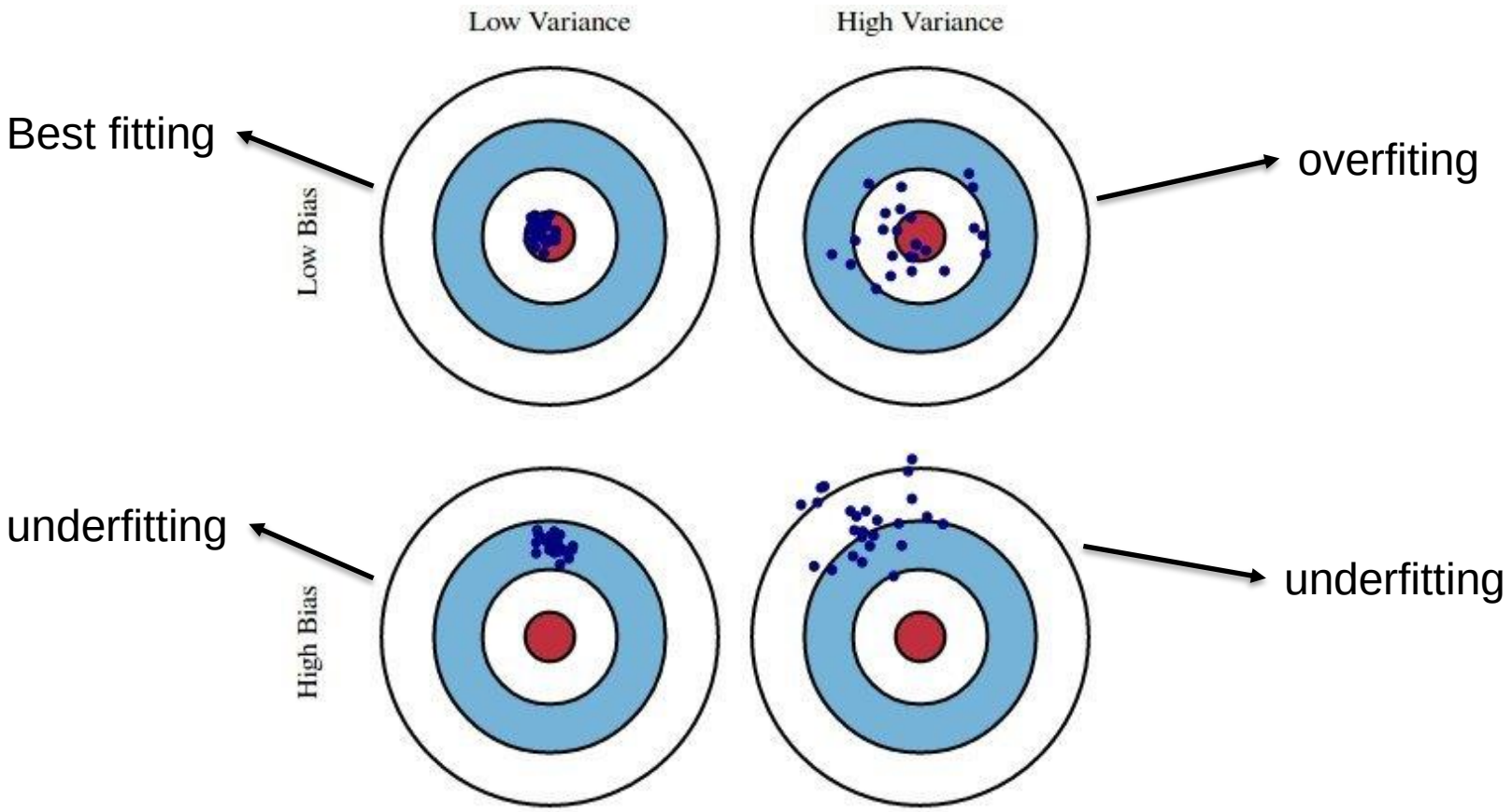
- **Variance**: là sai số trung bình của mô hình dự đoán so với kỳ vọng (trung bình) mô hình được huấn luyện trên toàn bộ dữ liệu thực

$$\text{variance}(\hat{\beta}) = E\left[\left(E[\hat{\beta}] - \hat{\beta}\right)^2\right]$$

- **Variance thấp**: thể hiện tính tổng quát cao của mô hình, dù huấn luyện trên một tập con vẫn đoán đúng trên dữ liệu chưa thấy (tập test)
- **Variance cao**: thể hiện mô hình không đoán tốt trên dữ liệu chưa gặp

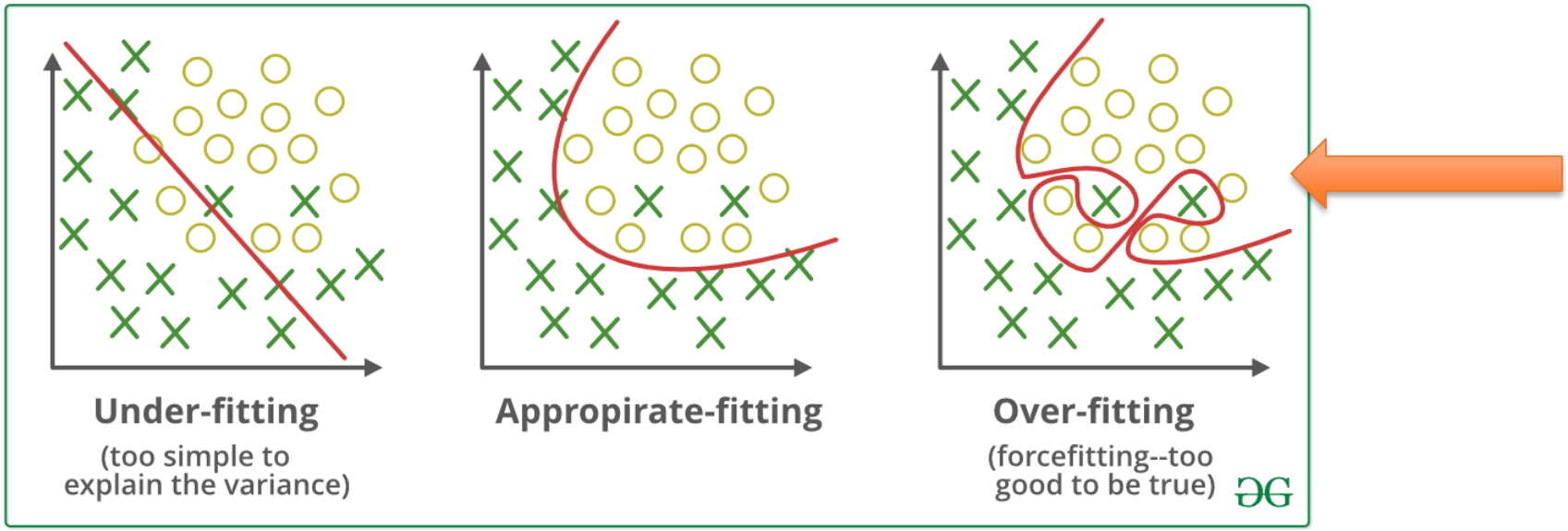
Cân bằng bias-variance

- Khi xây dựng mô hình và huấn luyện trên tập dữ liệu cần chú ý đảm bảo cả bias và variance đều thấp



Hiện tượng overfiting

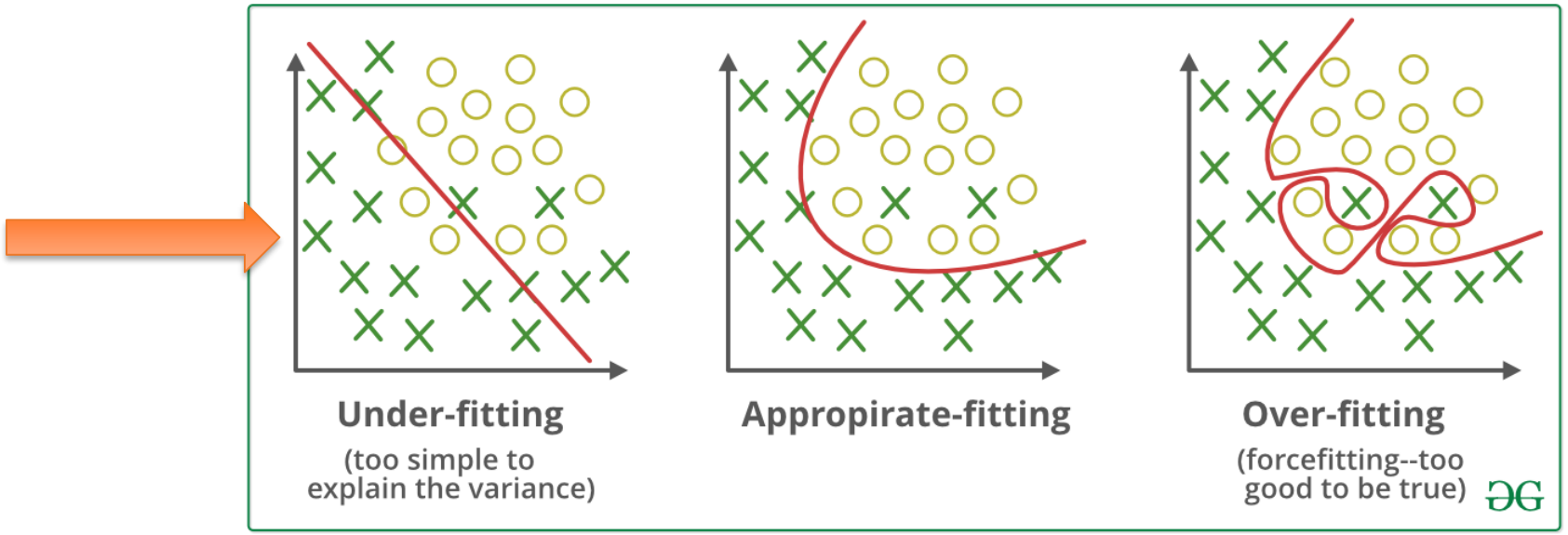
- **Nguyên nhân**: Xảy ra khi mô hình quá phức tạp hoặc dữ liệu không đủ khái quát



- **Để tránh overfiting**:
  - ➔ Giảm bớt sự phức tạp của mô hình (**giảm tham số**)
  - ➔ Lấy mẫu thêm dữ liệu để bao gồm các tình huống thực tế (**tăng dữ liệu**)

Hiện tượng underfitting

- **Nguyên nhân**: Xảy ra khi mô hình quá đơn giản so với tính chất phức tạp của dữ liệu, hoặc dữ liệu không đủ khái quát



- **Để tránh underfiting**:
  - ➔ Chuyển sang mô hình có khả năng biểu diễn phức tạp hơn (**thay hàm mô hình**)
  - ➔ Lấy mẫu thêm dữ liệu để bao gồm các tình huống thực tế (**tăng dữ liệu**)

NỘI DUNG

1. MÔ HÌNH HỒI QUY TUYẾN TÍNH

2. BIAS VÀ VARIANCE

3. MÔ HÌNH LASSO, RIDGE VÀ ELASTIC NET

4. MỘT SỐ MÔ HÌNH PHI TUYẾN

LASSO Regression

- Hồi quy tuyến tính + chính quy hóa L_1

- Có thể dùng để lựa chọn đặc trưng (Feature selection): mô hình cố gắng đưa các hệ số về 0 đối với những đặc trưng không quan trọng

$$\underset{\hat{\beta}}{\operatorname{argmin}} \left\| Y - \hat{\beta} X \right\|^2 + \lambda \left\| \hat{\beta} \right\|_1$$

➔hướng đến chọn lựa các đặc trưng quan trọng

Ridge Regression

- Hồi quy tuyến tính + chính quy hóa L_2

$$\underset{\hat{\beta}}{\operatorname{argmin}} \left\| Y - \hat{\beta} X \right\|^2 + \lambda \left\| \hat{\beta} \right\|_2^2$$

➔hướng đến khai thác hết các đặc trưng

Elastic Net

- Hồi quy tuyến tính + chính quy hóa $L_1 + L_2$

$$\underset{\hat{\beta}}{\operatorname{argmin}} \left\| Y - \hat{\beta} X \right\|^2 + \lambda \left(\frac{1-\alpha}{2} \left\| \hat{\beta} \right\|_2^2 + \alpha \left\| \hat{\beta} \right\|_1 \right)$$

➔ Cân bằng các tiêu chí của LASSO và Ridge Regression

NỘI DUNG

1. MÔ HÌNH HỒI QUY TUYẾN TÍNH

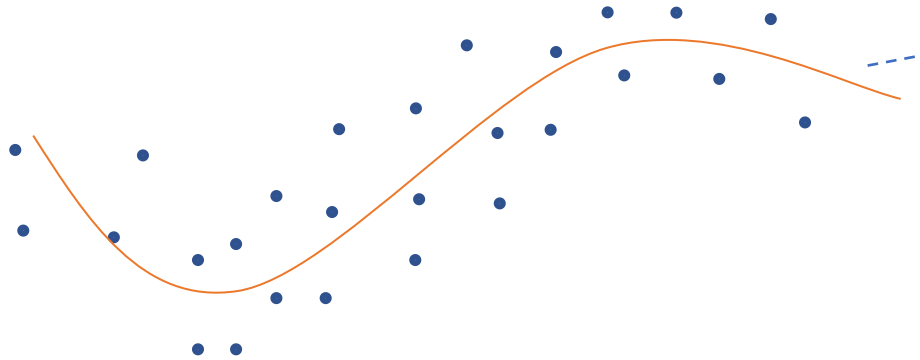
2. BIAS VÀ VARIANCE

3. MÔ HÌNH LASSO, RIDGE VÀ ELASTIC NET

4. MÔ HÌNH VỚI DỮ LIỆU CÓ QUAN HỆ PHI TUYẾN

Một số mô hình Regression khác

- Vẫn dùng LinearRegression, nhưng với feature engineer:

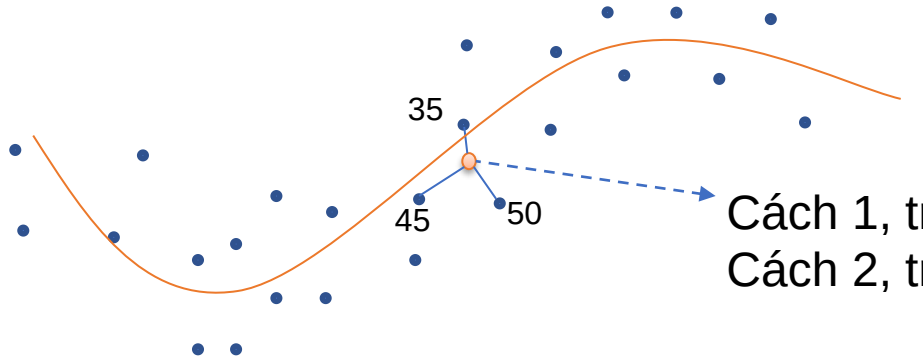


Dự đoán dạng của hàm, ta tạo các đặc trưng tương ứng.
Ví dụ: mô hình hàm số bậc 3

- Trước đây $X = \begin{bmatrix} 1 \\ x \end{bmatrix}$, ta tạo thêm các đặc trưng $X_{\text{new}} = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}$

Một số mô hình Regression

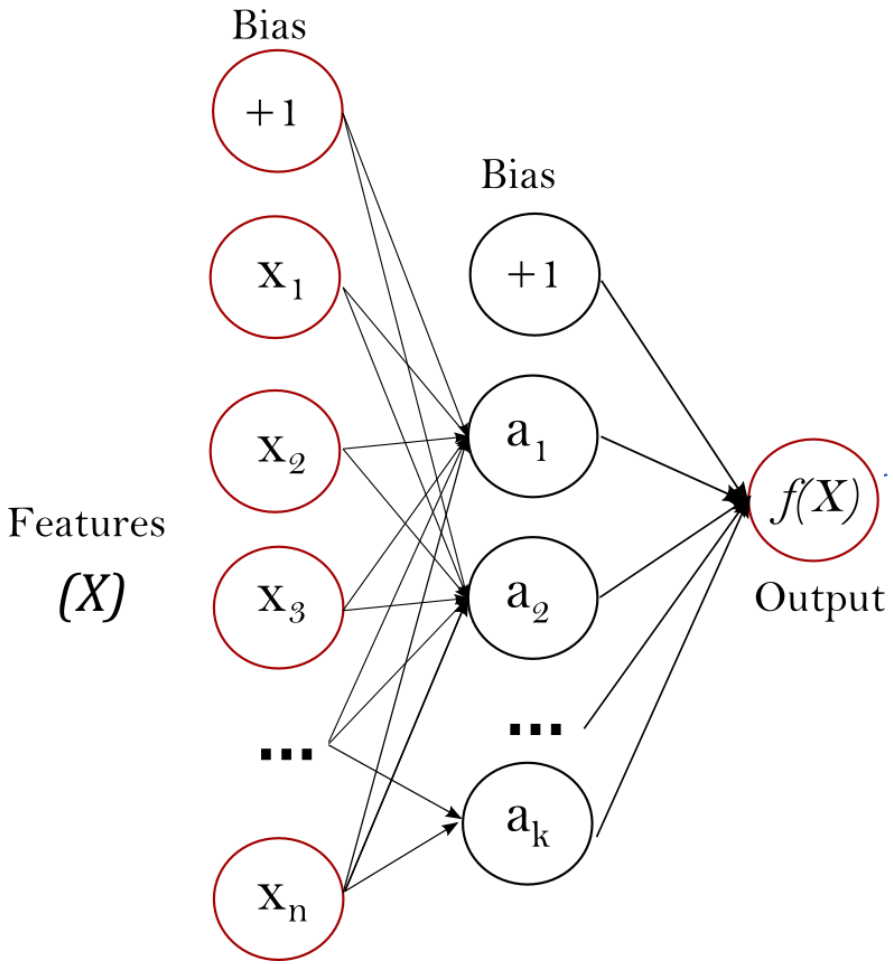
- **KNNRegressor**



Cách 1, trung bình: (35+45+50)/3 = 43.3
Cách 2, trung bình trọng số theo khoảng cách

Một số mô hình Regression

- **Mạng Neural Network** (hay Multi-Layer Perceptron)



Không có hàm kích hoạt cuối

Một số mô hình Regression

- Các mô hình có nguồn gốc từ mô hình phân lớp:
  - Support Vector Regressor
  - Decision Tree Regressor
  - Random Forest Regressor
  - Gradient Boost, XGBoost, LightGBM, CatBoost Regressor

BÀI QUIZ VÀ HỎI ĐÁP