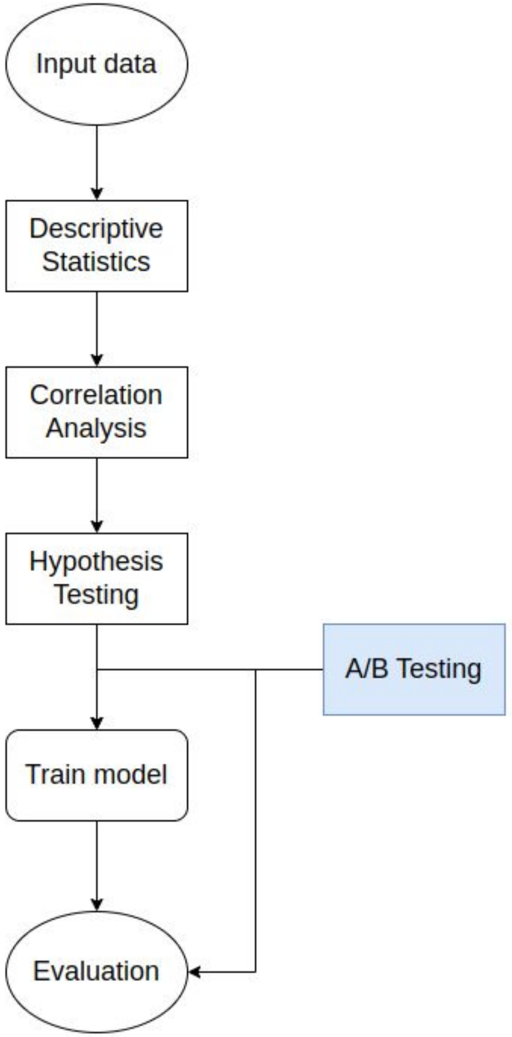
AI VIET NAM
@aivietnam.edu.vn

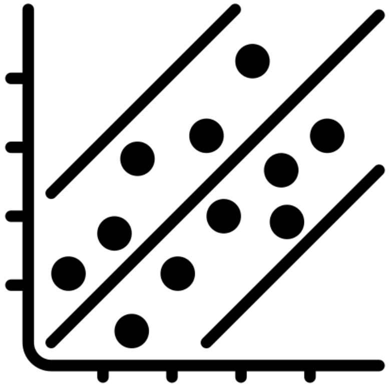
Regression and Variants

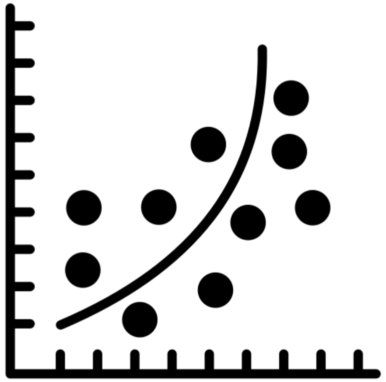
Minh-Hung An - TA
Minh-Loi Nguyen - STA

Objectives

- ✔ Understand the concept of Regression
- ✔ Understand the variants of Linear Regression
- ✔ Linear Regression & Prompt Engineering







Nguyên lý hoạt động của A/B testing với vai trò feature như sau:
Mỗi dòng dữ liệu được gán ngẫu nhiên vào một trong hai nhóm, nhóm A hoặc nhóm B. Việc ngẫu nhiên hóa này giúp đảm bảo rằng không có sự thiên vị hoặc thành kiến nào trong việc phân nhóm, làm cho kết quả thử nghiệm đáng tin cậy và có thể tổng quát hóa.

Outline

1 Regression

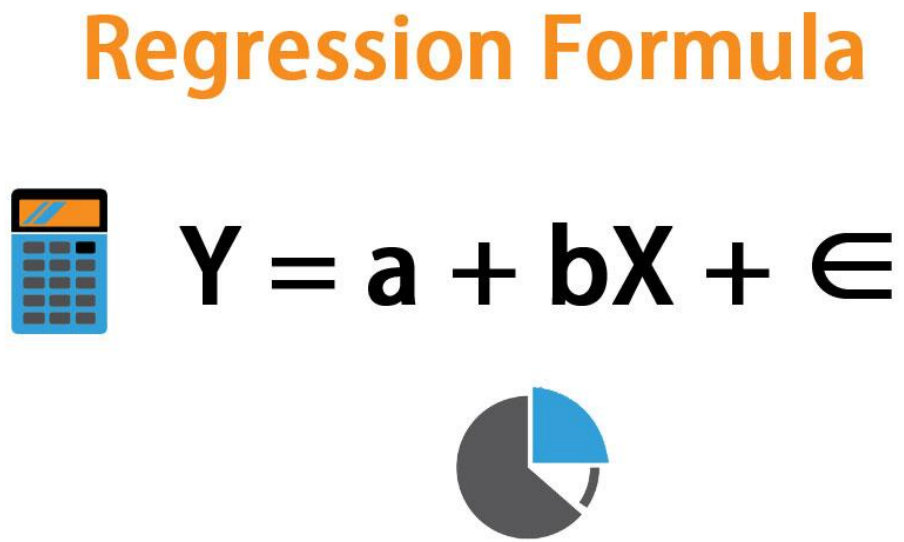
2 Linear Regression

3 Regularization

4 Other

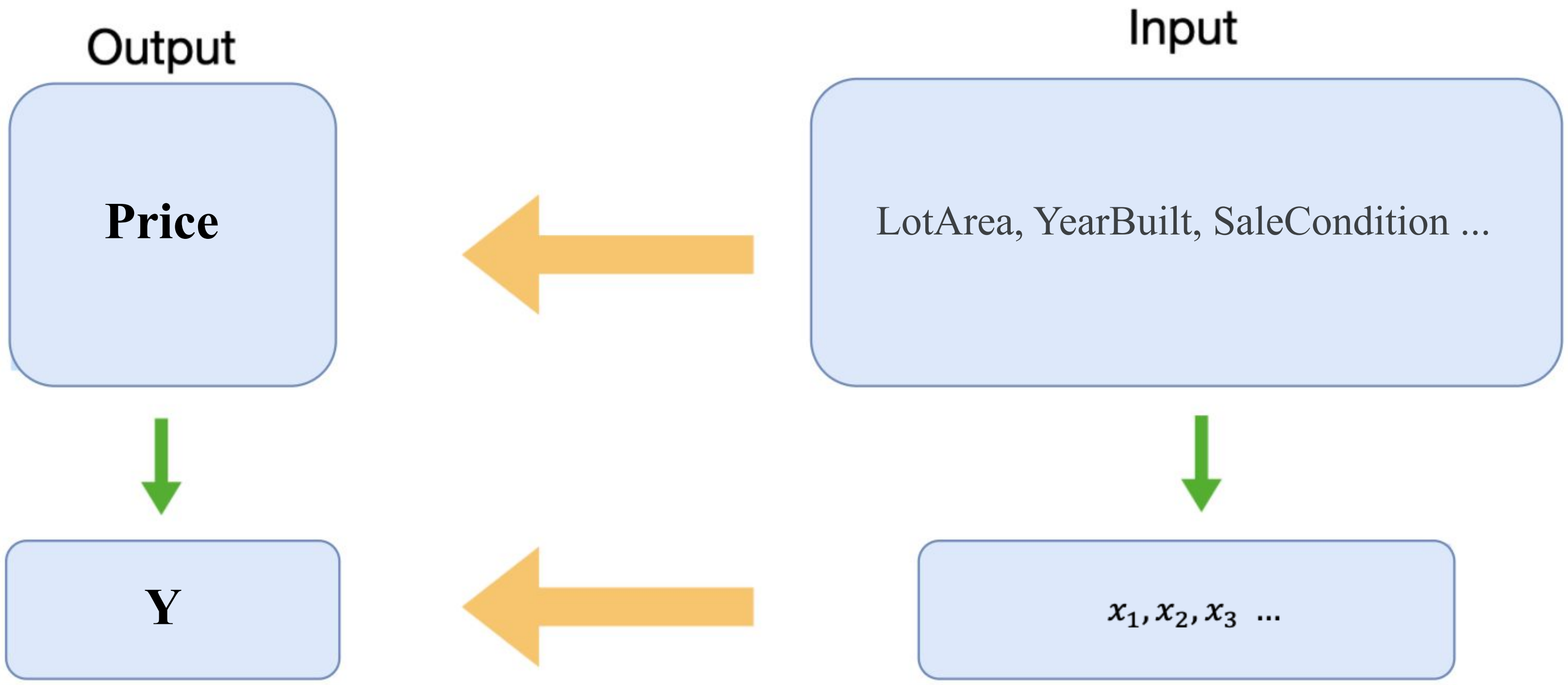
Outline

1. **Regression**
2. **Linear Regression**
3. **Regularization**
4. **Other**



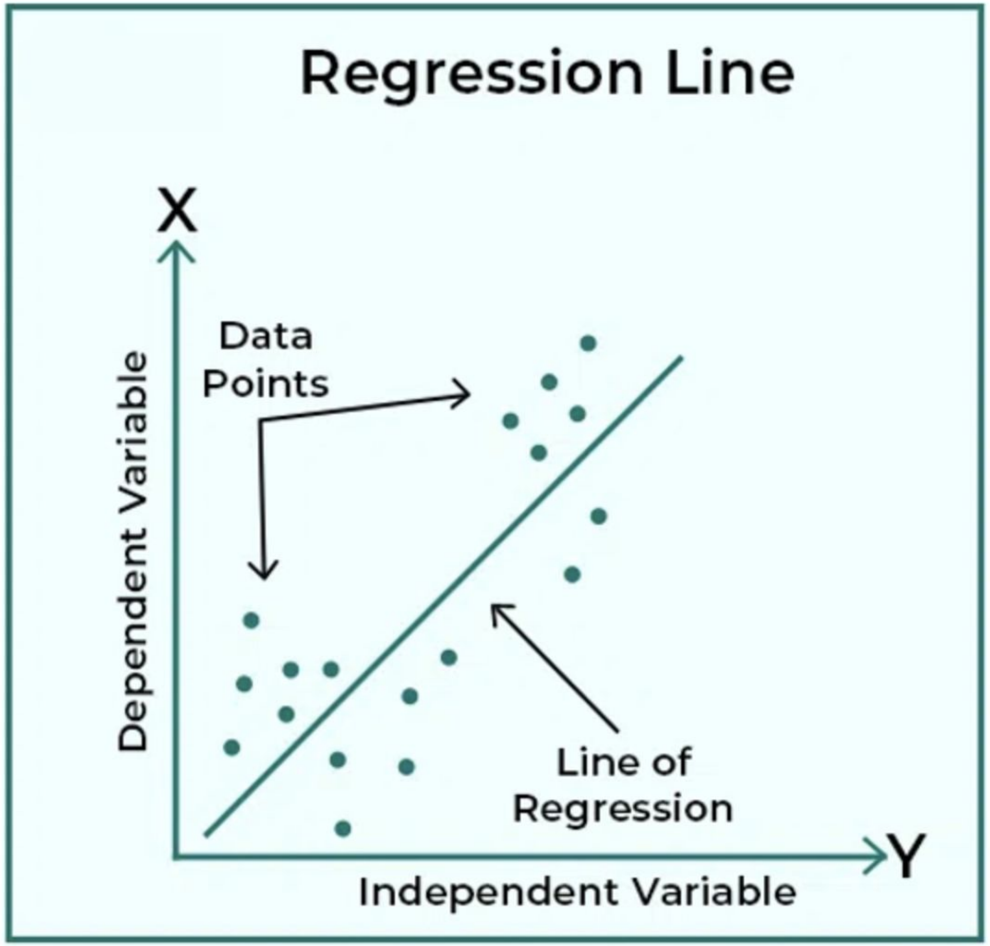
Regression

❖ Regression là gì?



Regression

❖ Regression là gì?



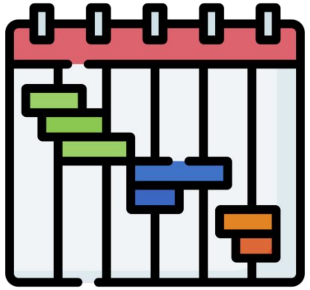
Supervised Learning

Predict continous output

Regression

❖ Ứng dụng

HealthCare



Projection

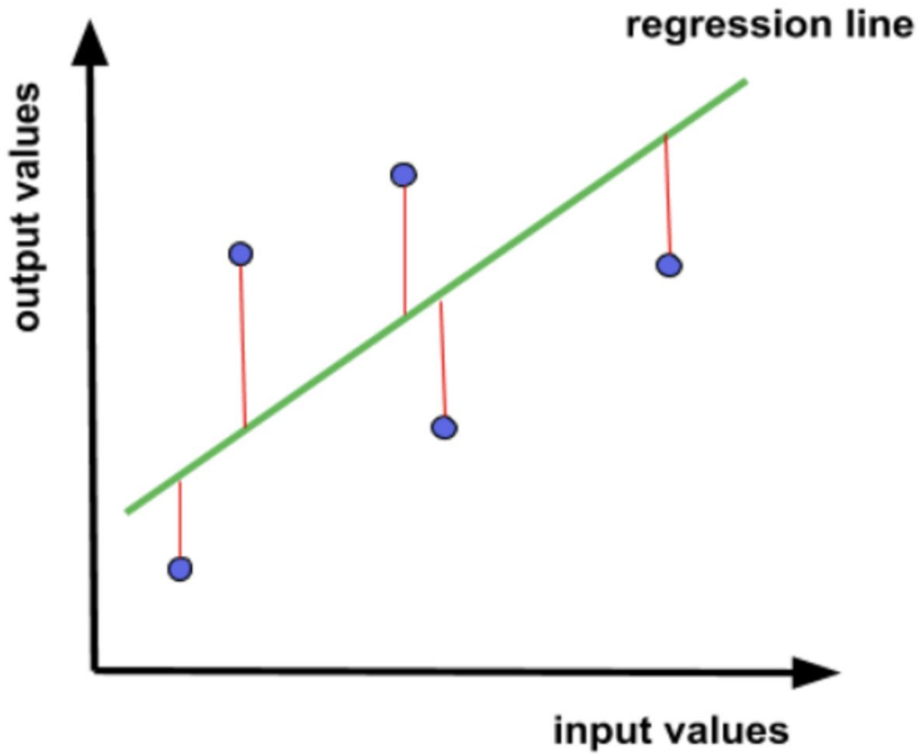
Regression

Marketing

Finance

Regression

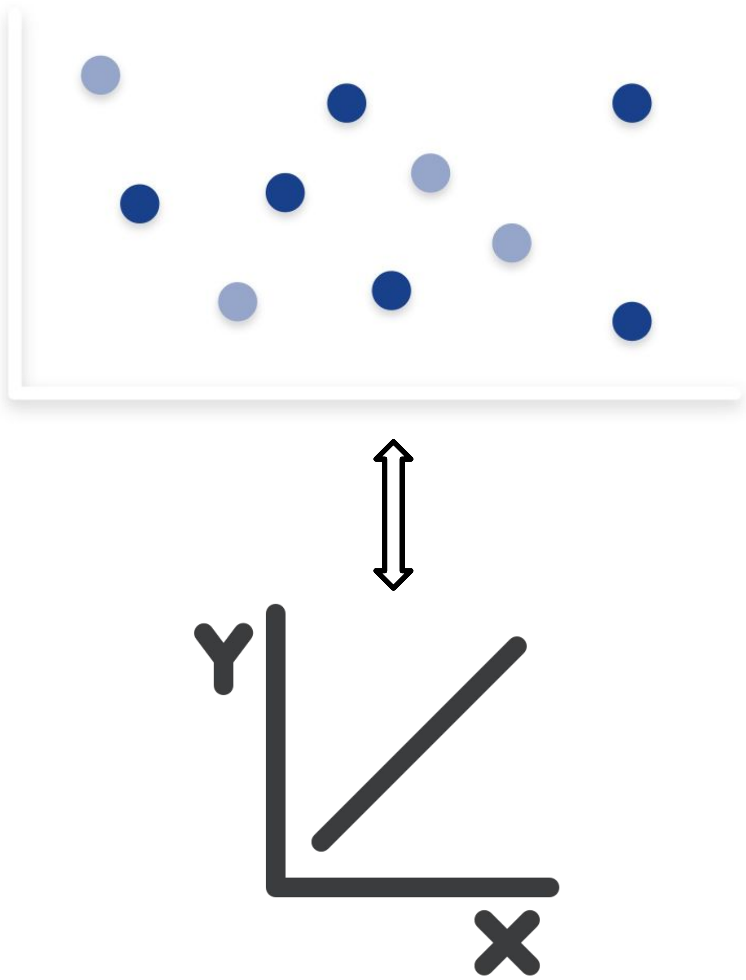
❖ Common Regression Metrics



1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. R-squared(R^2)

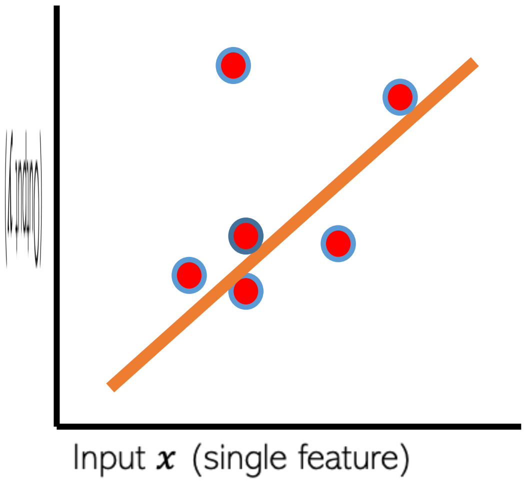
Outline

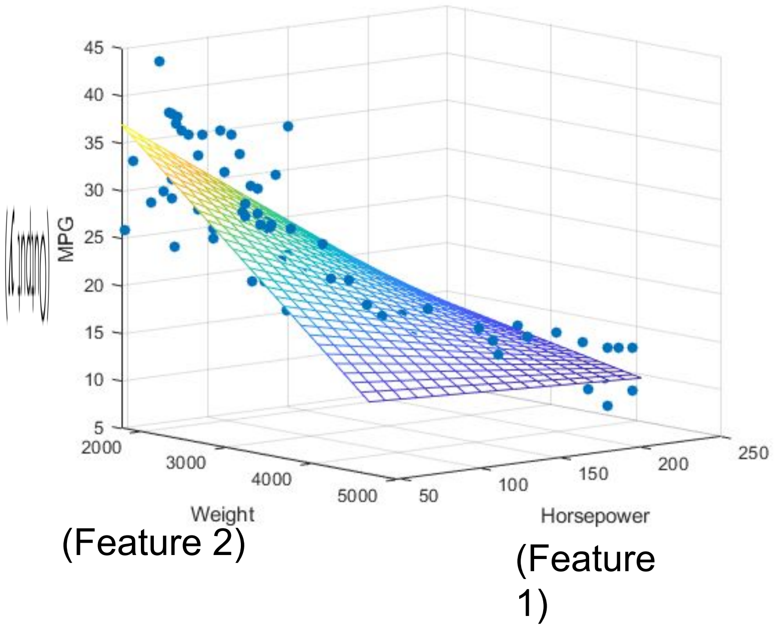
1. Regression
2. Linear Regression
3. Regularization
4. Other



Linear Regression

❖ Linear Regression là gì?





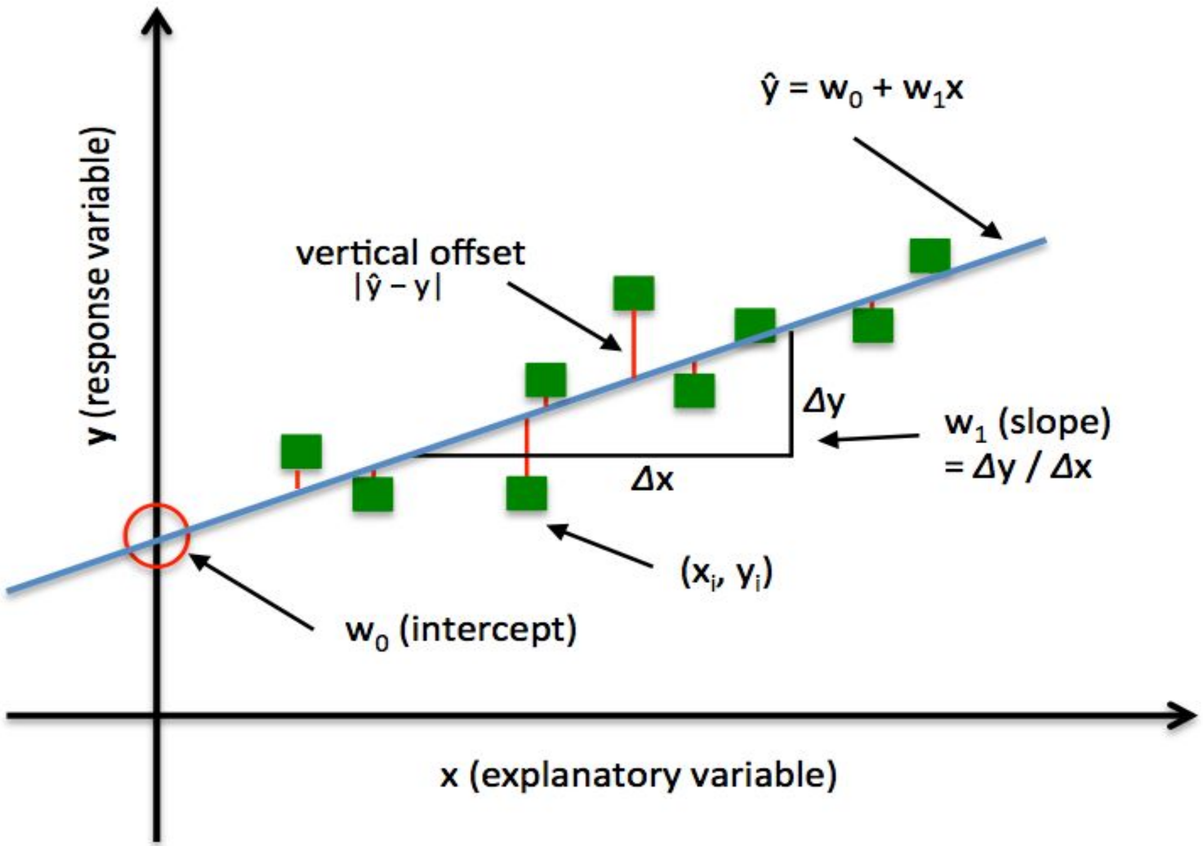
Là việc khớp một **đường thẳng** hoặc **mặt phẳng** (siêu) với một tập hợp các điểm.

Linear Regression

❖ Linear Regression

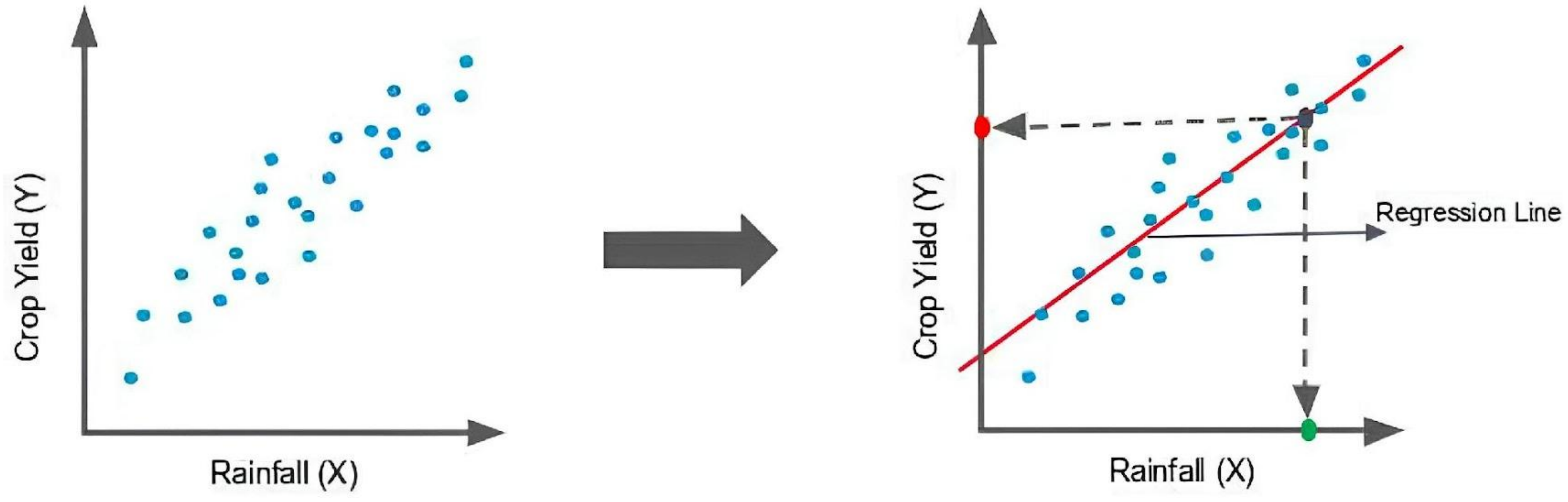
$$\hat{y} = w_0 + w_1 x + \varepsilon$$

- ε : thành phần ngẫu nhiên hoặc thành phần lỗi
- x : feature của biến dữ liệu
- $\hat{y}$: giá trị mong muốn dự đoán
- w_1 : Hệ số góc của đường thẳng
- w_0 : Hệ số tự do của đường thẳng



Linear Regression

❖ Linear Regression

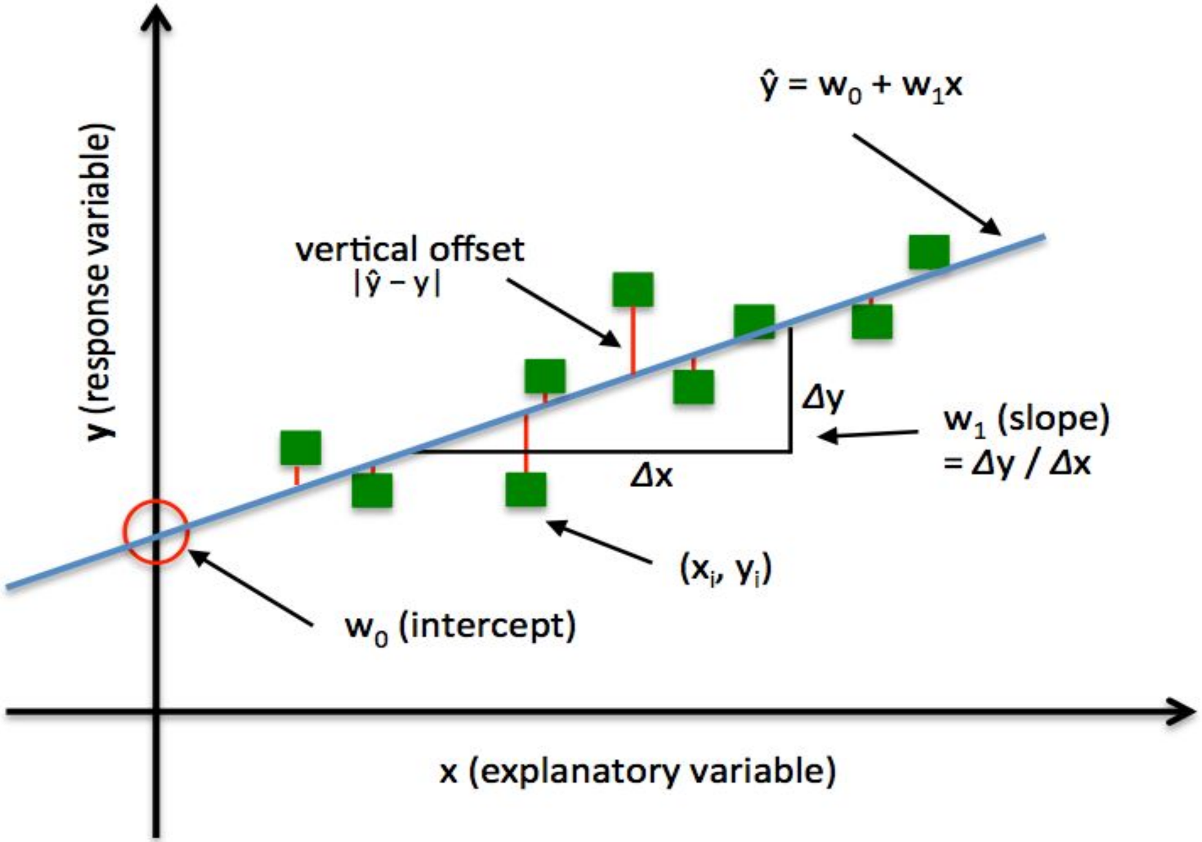


Linear Regression

❖ Giải bài toán Regression

Tìm bộ $\mathbf{w_1}$, $\mathbf{w_0}$ khớp với bộ dữ liệu

$$\arg\min_{w_0, w_1} \sum_{i}^{n} (y_i - w_0 - w_1 x_i)^2$$



Linear Regression

❖ Cost Function

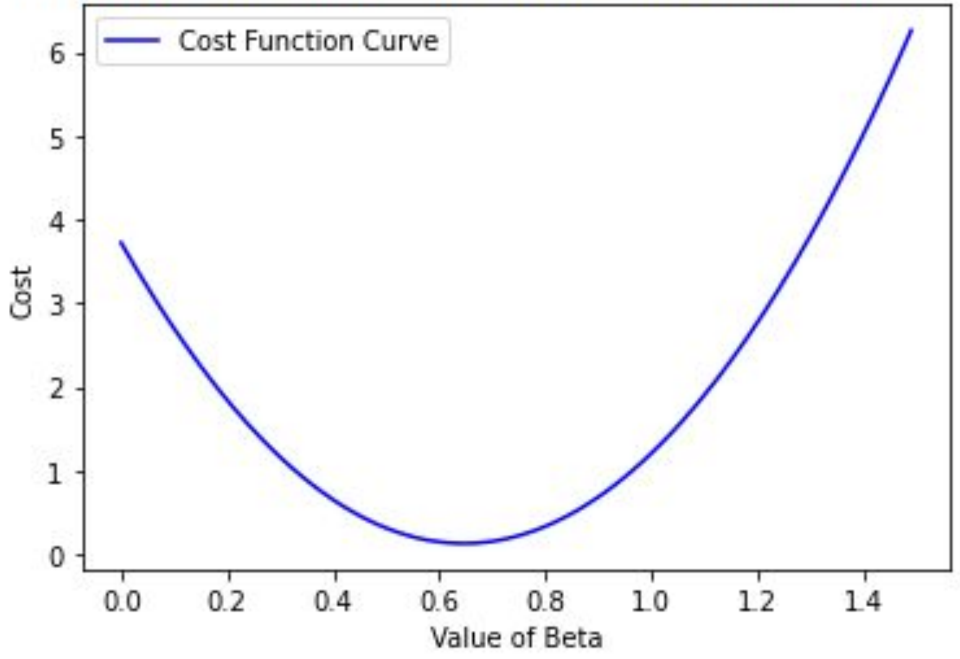
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Tổng số điểm dữ liệu

Giá trị thực

Giá trị dự đoán

<matplotlib.legend.Legend at 0x7f6fd81050d0>



Minimize Cost Function

Linear Regression

Cost Function

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Linear Regression

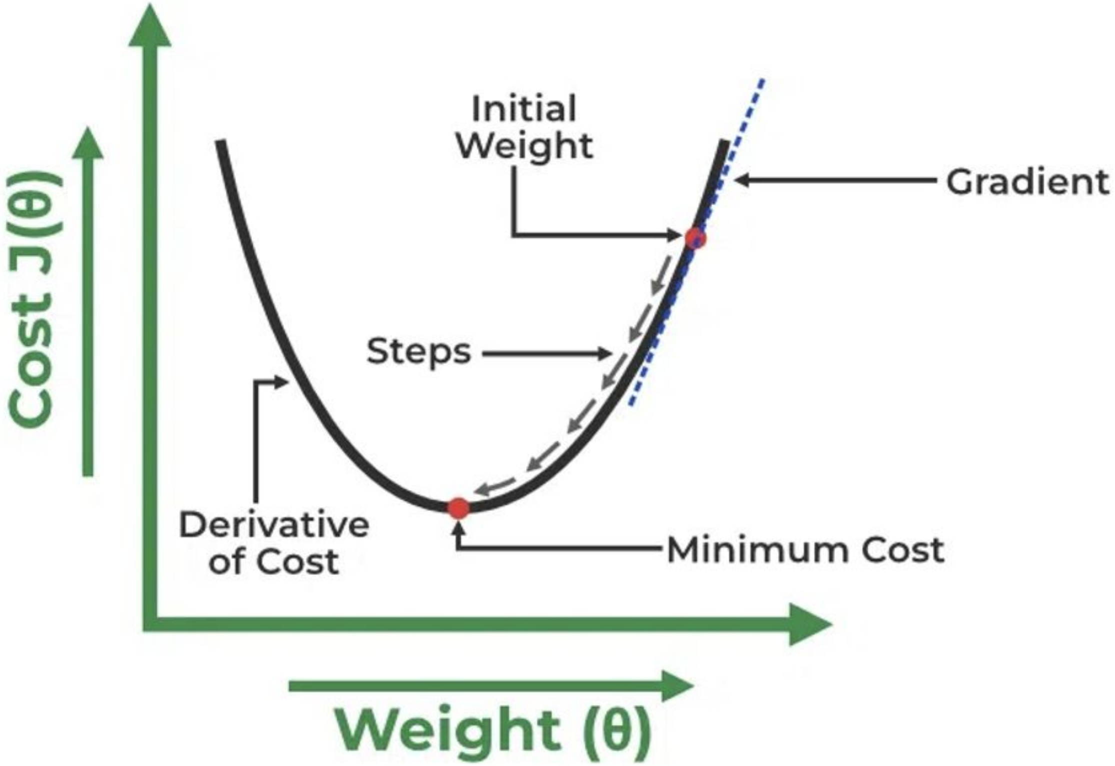
❖ Calculus

$$\frac{\partial \epsilon^2}{\partial w_0} = \sum_i^n -2(y_i - w_0 - w_1 x_i) = 0$$

$$\frac{\partial \epsilon^2}{\partial w_1} = \sum_i^n -2x_i(y_i - w_0 - w_1 x_i) = 0$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$w_1 = \frac{n\sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{n\sum_i^n x_i x_i - \sum_i^n x_i \sum_i^n x_i}$$

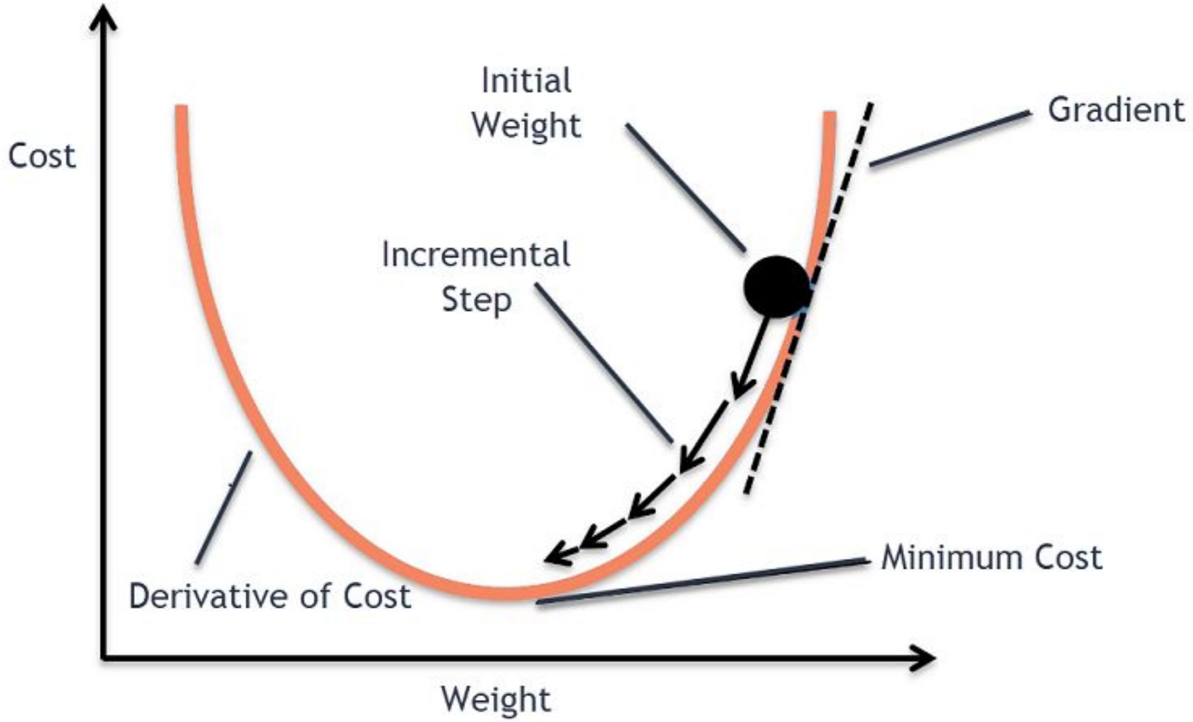


Linear Regression

Gradient Descent

Gradient descent algorithm

$$
\begin{aligned}
&\text{repeat until convergence } \{ \\
&\qquad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \\
&\qquad\qquad (\text{for } j = 1 \text{ and } j = 0) \\
&\}
\end{aligned}
$$



❖ Đánh giá mô hình Linear Regression

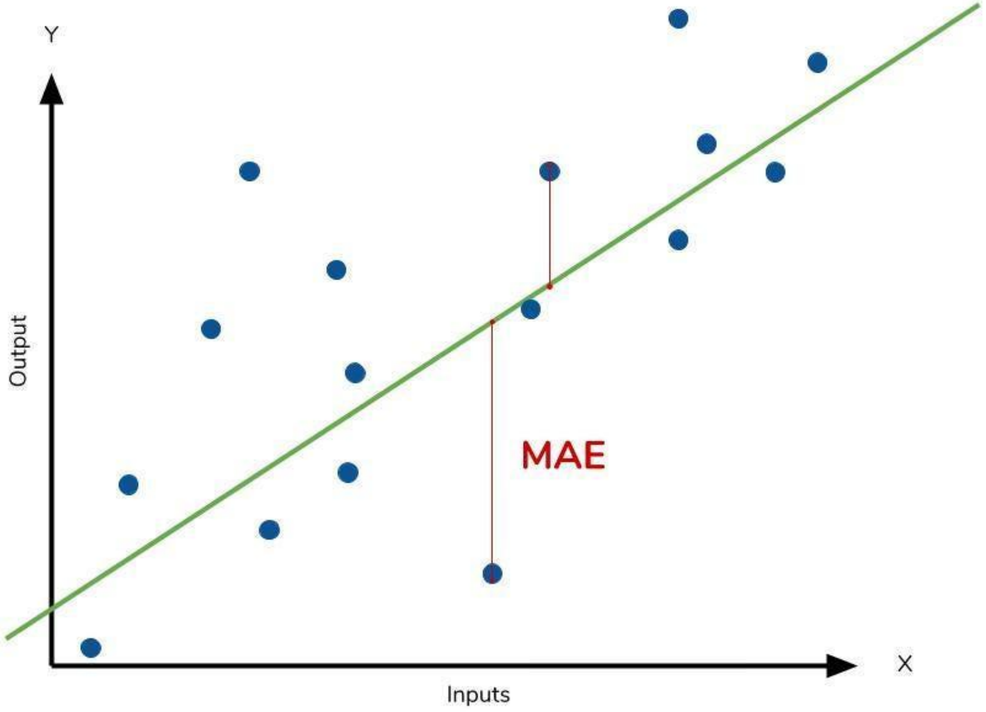
Divide by the total number of data points

Predicted output value

Actual output value

$$MAE = \boxed{\frac{1}{n}} \sum \underbrace{\left| \boxed{y} - \boxed{\hat{y}} \right|}_{\text{The absolute value of the residual}}$$

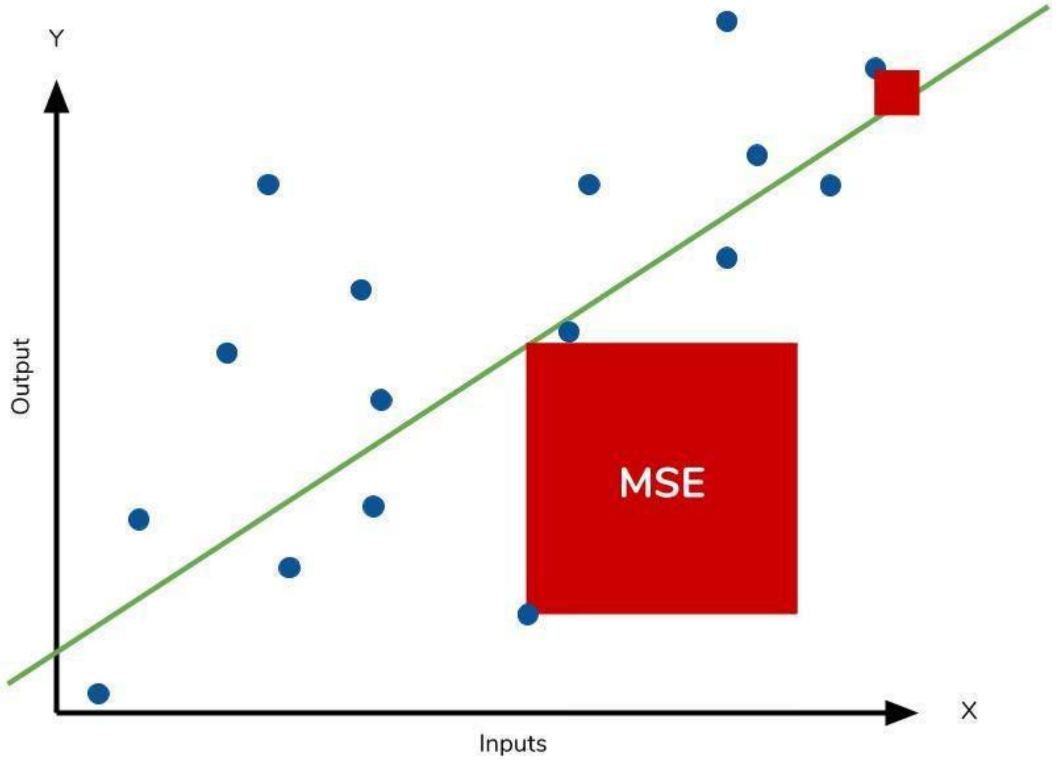
Sum of



Linear Regression

❖ Đánh giá mô hình Linear Regression

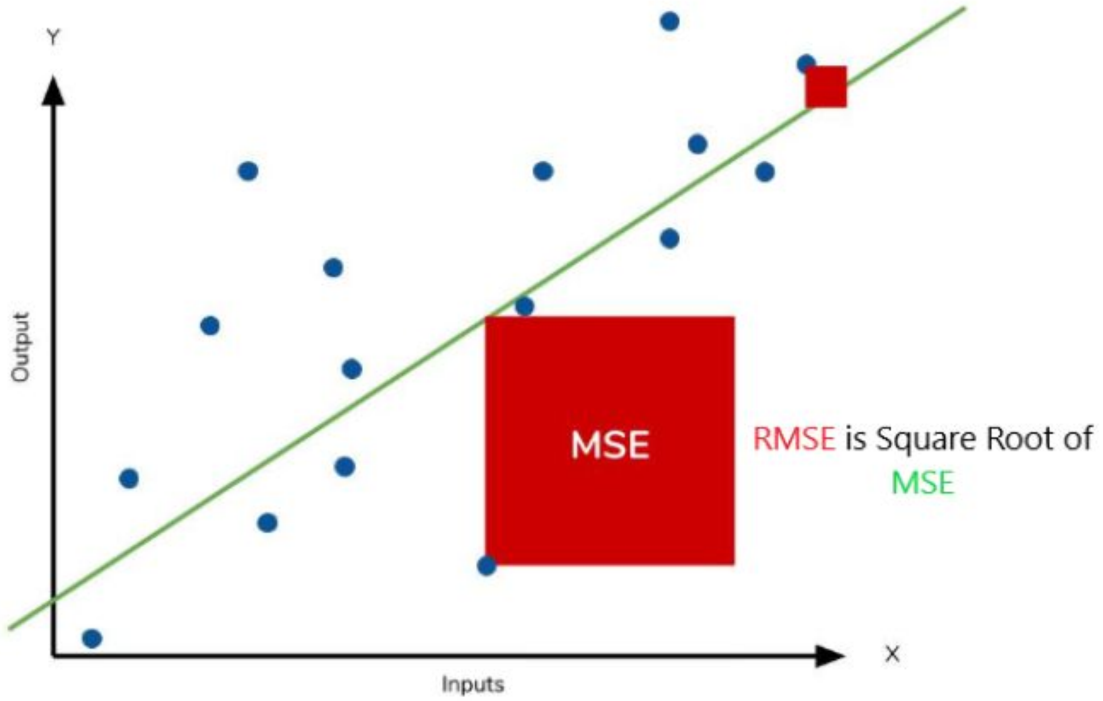
$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$



Linear Regression

❖ Đánh giá mô hình Linear Regression

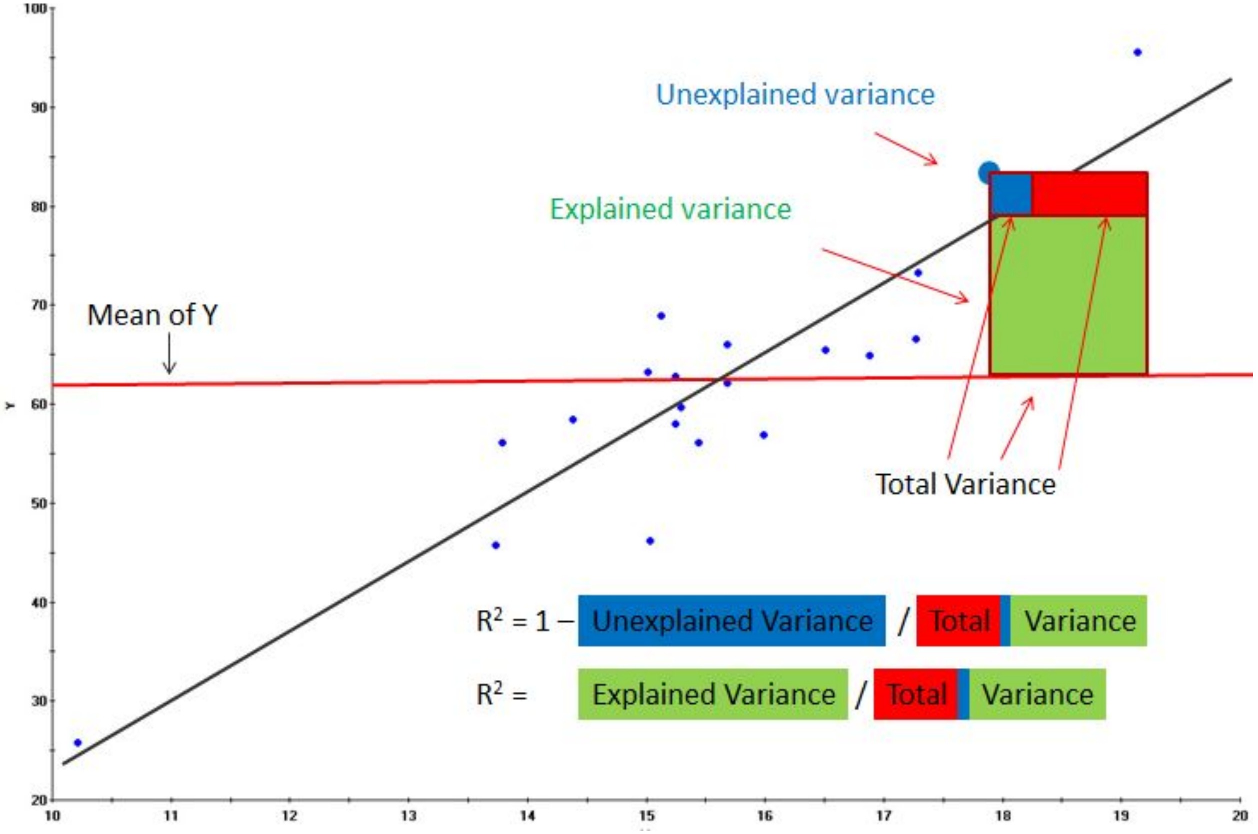
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$



Linear Regression

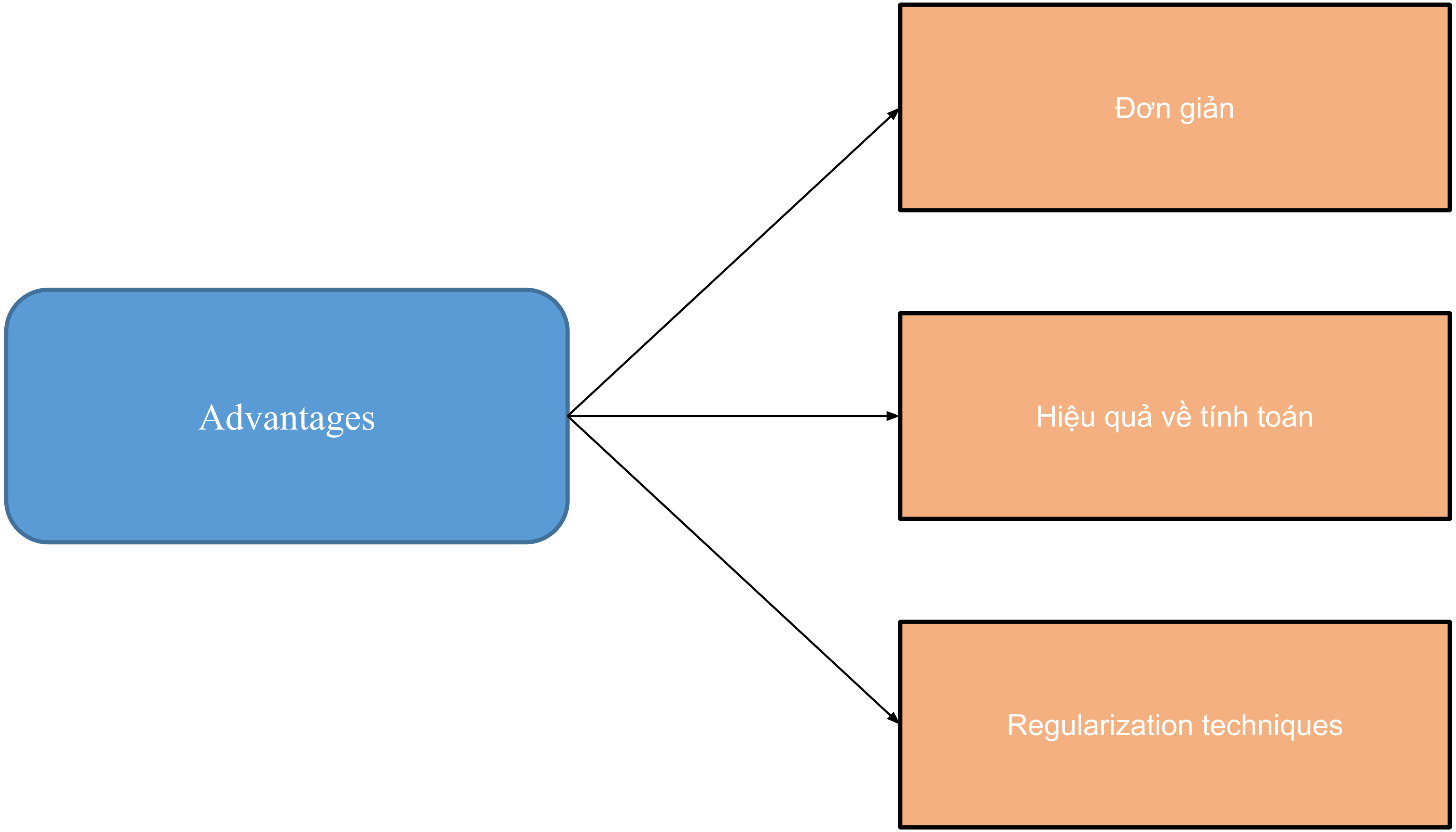
❖ Đánh giá mô hình Linear Regression

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$



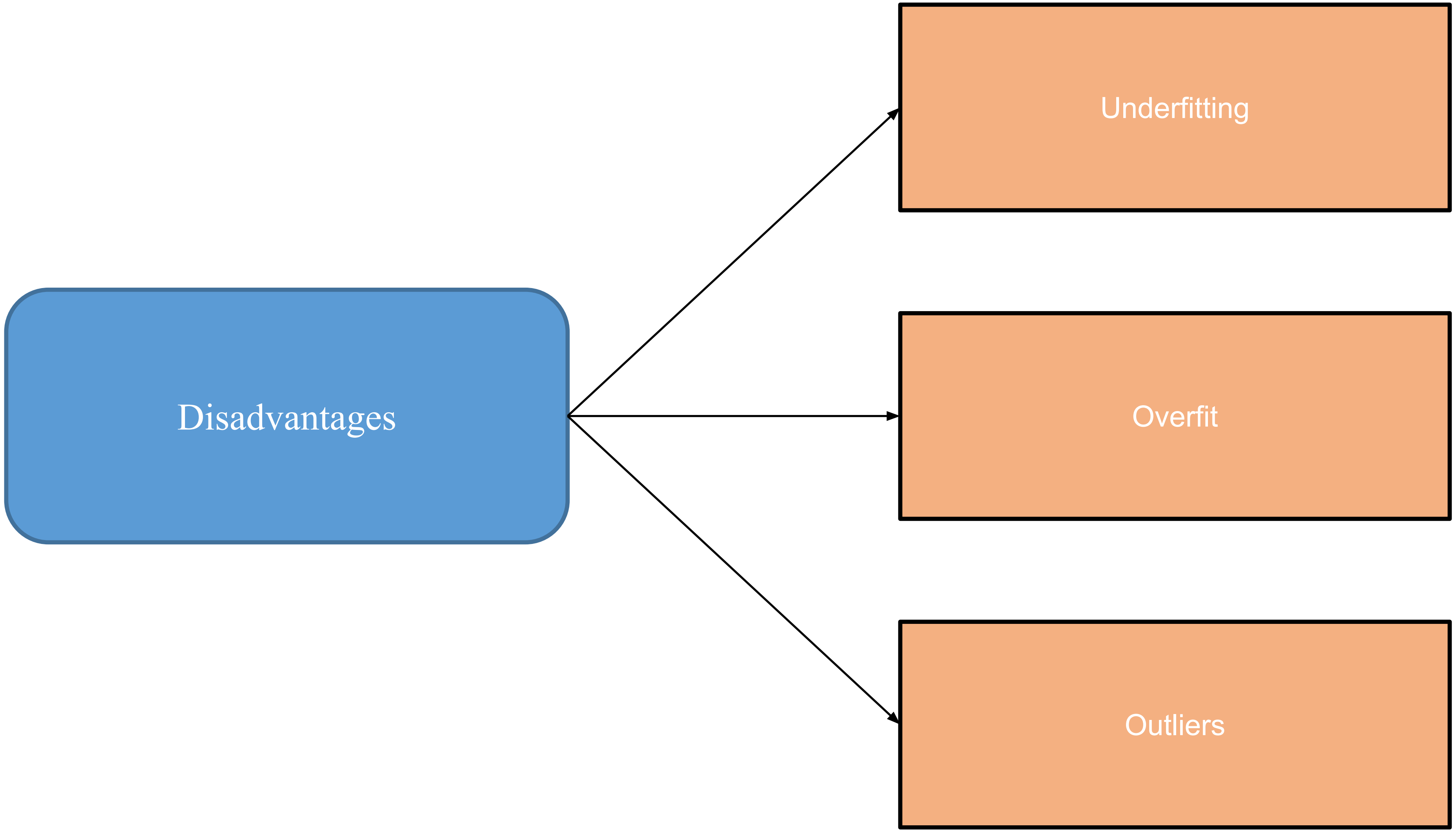
Linear Regression

❖ Linear Regression



Linear Regression

❖ Linear Regression



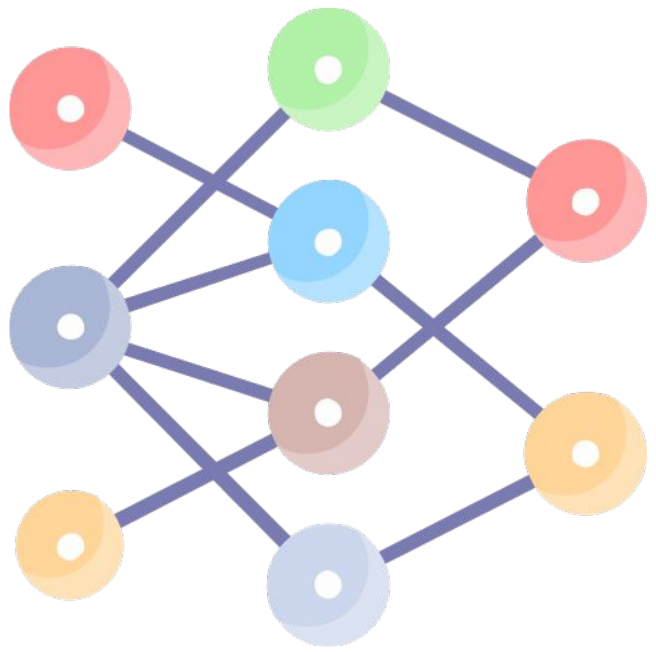
Outline

1 Regression

2 Linear Regression

3 **Regularization**

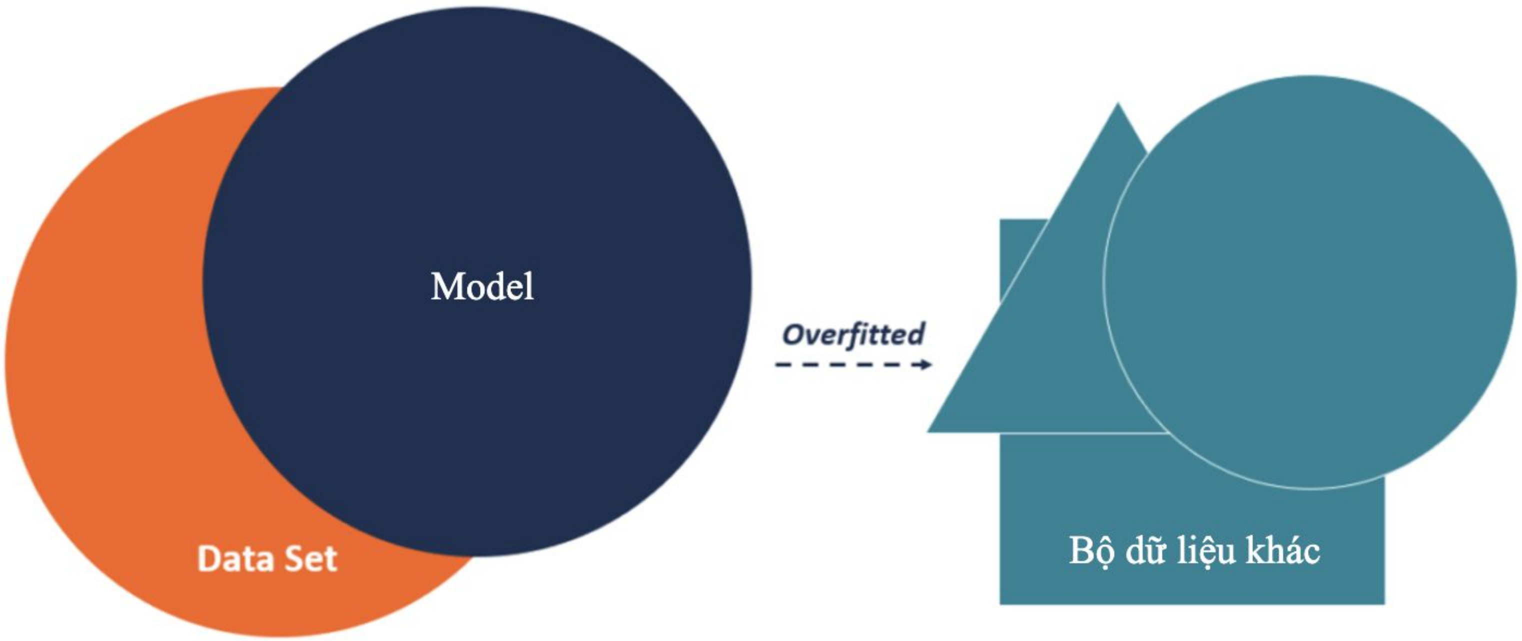
4 Other



AI VIET NAM
@aivietnam.edu.vn

Regularization

❖ Overfitting



→ Không khớp trên bộ dữ liệu khác

Regularization

❖ Linear Regression

- Problem: **Overfitting** xảy ra vì chỉ giảm thiểu **loss** được xác định trên training dataset
- Weights $\boldsymbol{w} = [w_1, w_2, \ldots, w_D]$ trở nên quá lớn để fit trên training dataset
  - Weights sẽ hoạt động không tốt trên test dataset

$R(\boldsymbol{w})$: Regularizer

- Solution: Minimize a regularized objective $\boxed{L(\boldsymbol{w}) + \lambda\, R(\boldsymbol{w})}$
  - Ngăn chặn **Weights** $\boldsymbol{w}$ trở nên quá lớn
  - Giải thích : Minimize trên cả training error + magnitude of vector

$\lambda \geq 0$

Regularization

Linear Regression

- Regularization of Linear Models
  - Ridge Regression (L2 Regularization)
  - Lasso Regression (L1 Regularization)
  - Elastic-Net Regression (kết hợp cả L1 và L2)

Regularization

Ridge Regression

$$L_{reg}(\boldsymbol{w}) = L(\boldsymbol{w}) + \lambda\, R(\boldsymbol{w})$$

$$R(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^\top \boldsymbol{w}$$

$$\boldsymbol{w}_{ridge} = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}) + \lambda\, R(\boldsymbol{w})$$

$$= \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

$$\boldsymbol{w}_{ridge} = \left(\sum_{n=1}^{N} \boldsymbol{x}_n\, \boldsymbol{x}_n^\top + \lambda I_D\right)^{-1} \left(\sum_{n=1}^{N} y_n \boldsymbol{x}_n\right)$$ (the optimal $\boldsymbol{w}$)

Regularization

Lasso Regression

$$L_{reg}(\boldsymbol{w}) = L(\boldsymbol{w}) + \lambda\, R(\boldsymbol{w})$$

$$R(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_{d=1}^{D} |w_d|$$

$$\boldsymbol{w}_{lasso} = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}) + \lambda\, R(\boldsymbol{w})$$

$$= \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \lambda \|\boldsymbol{w}\|_1$$

Regularization

Elastic-Net Regression

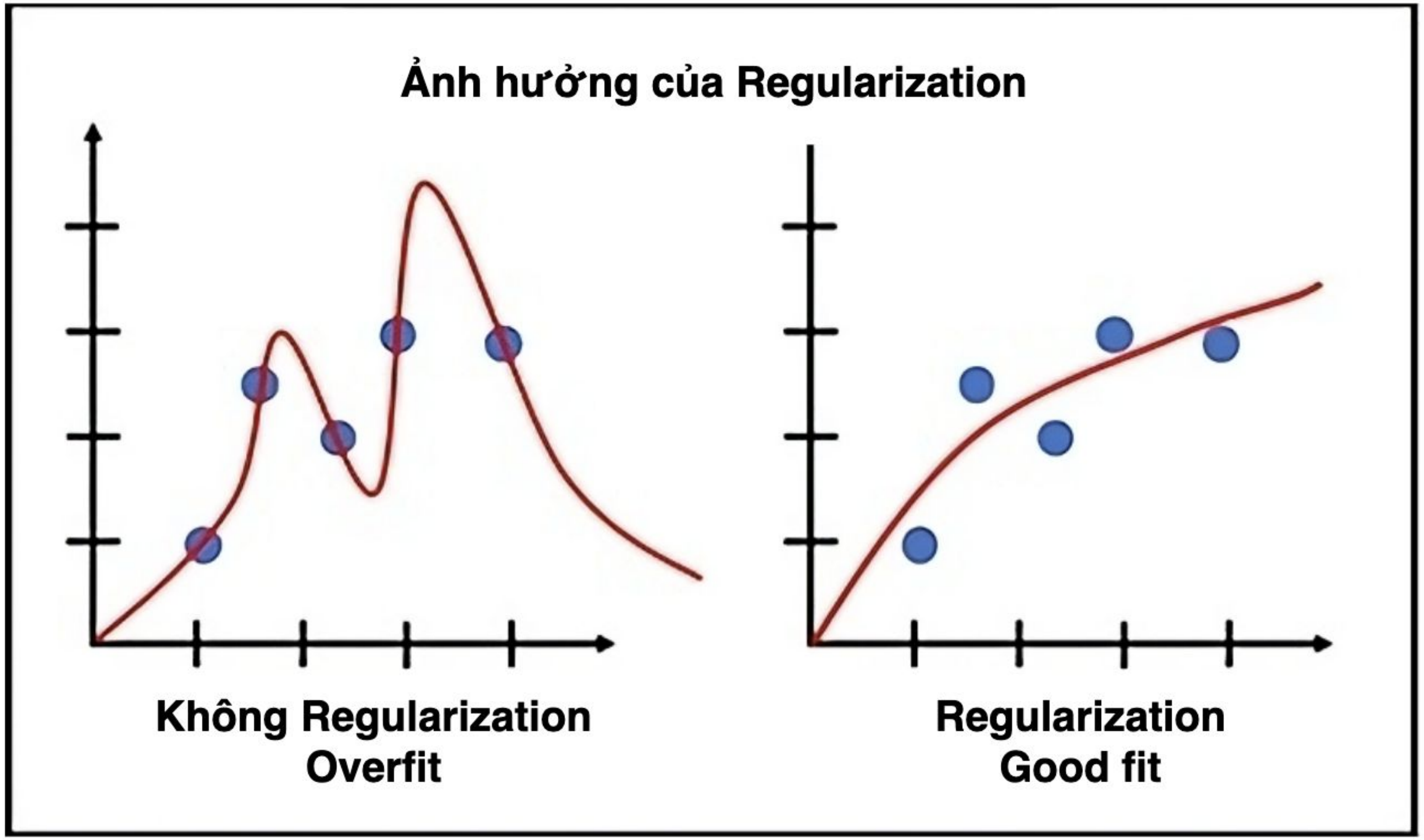
$$L_{reg}(\boldsymbol{w}) = L(\boldsymbol{w}) + \lambda_1 R_1(\boldsymbol{w}) + \lambda_2 R_2(\boldsymbol{w})$$

$$R_1(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_{d=1}^{D} |w_d|$$

$$R_2(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^\top \boldsymbol{w}$$

Regularization

❖ Linear Regression

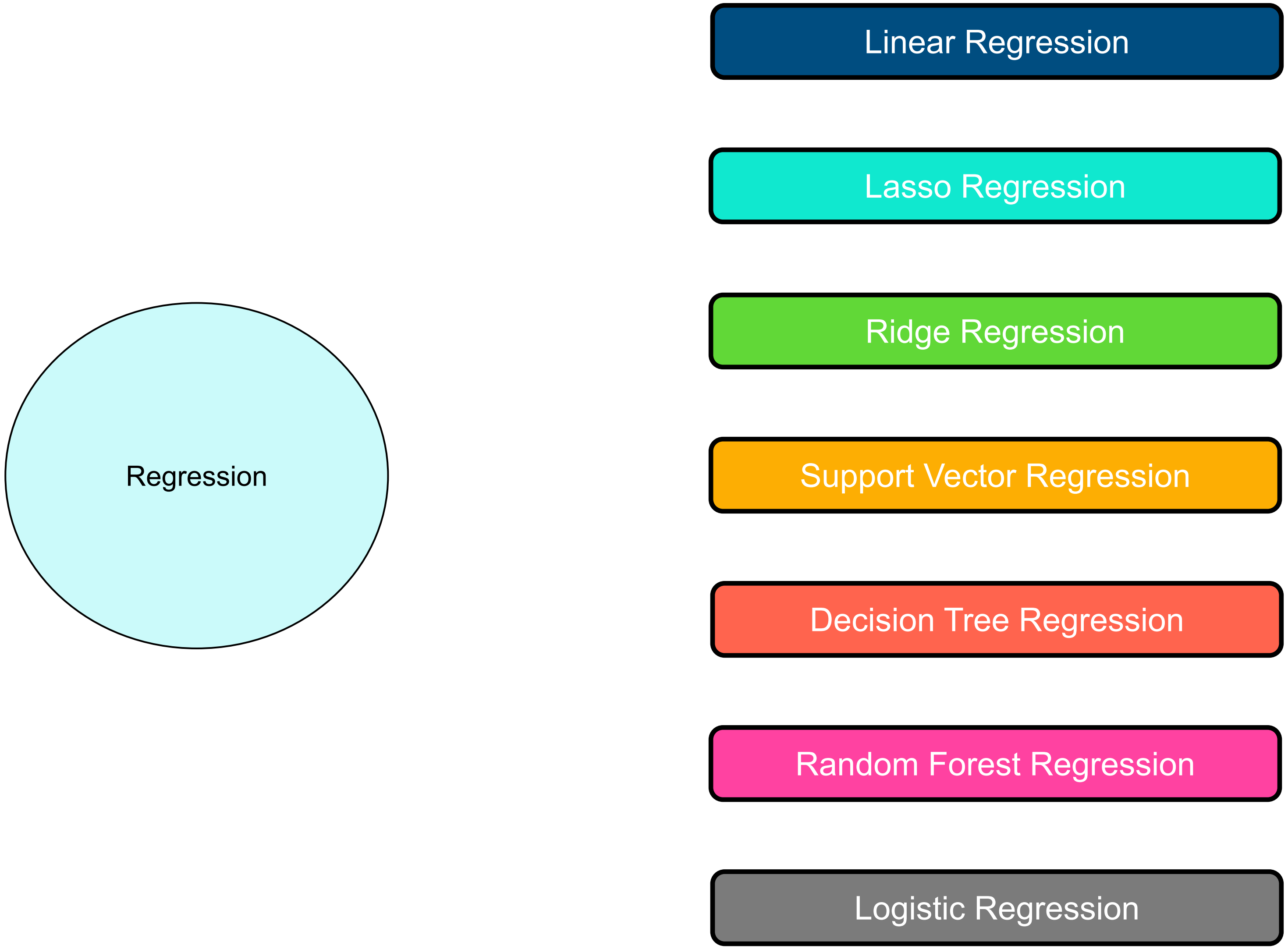


Outline

1. Regression
2. Linear Regression
3. Regulariation
4. **Other**

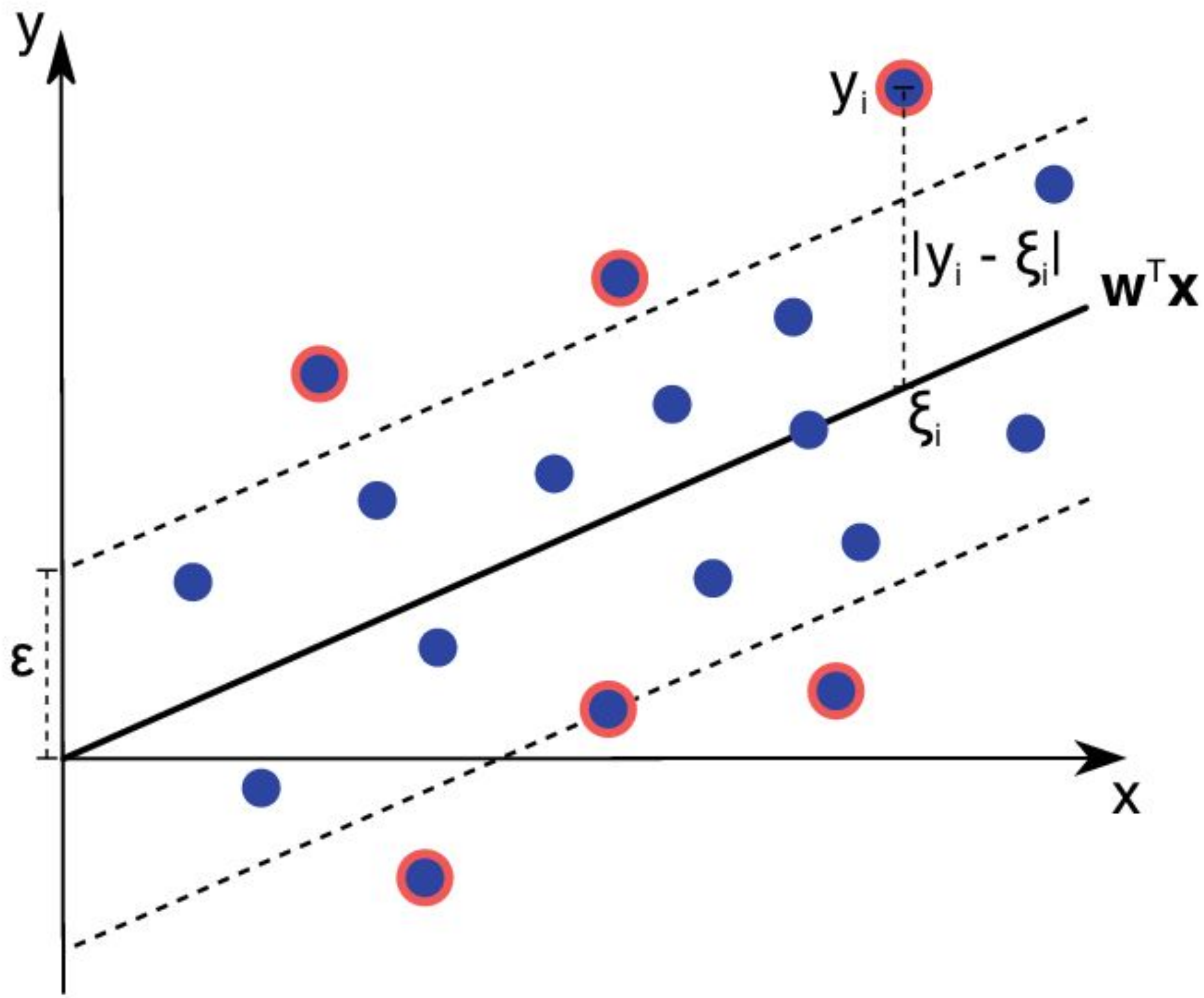
Other

❖ Types of Regression



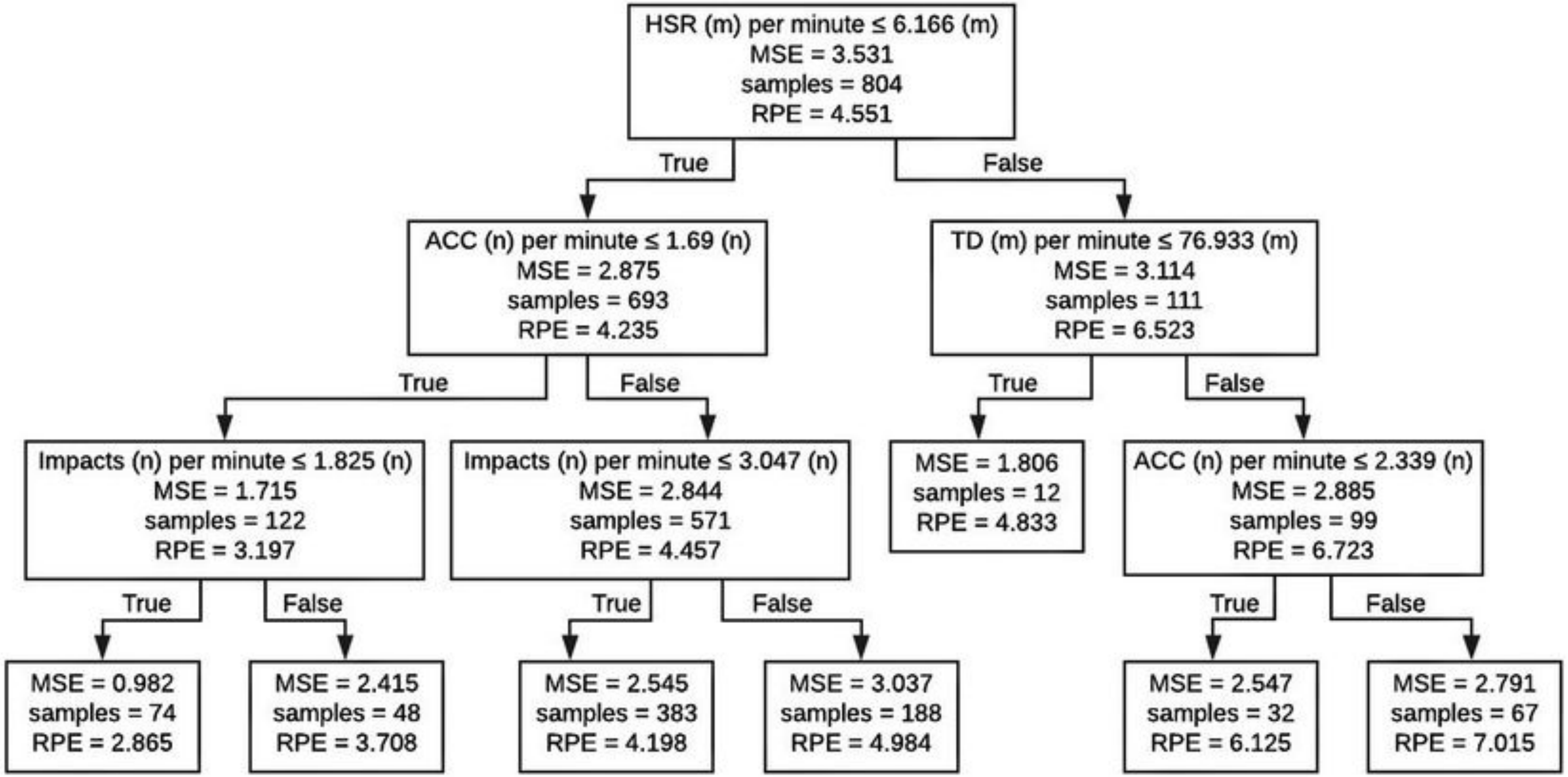
Other

❖ Support Vector Regression



Other

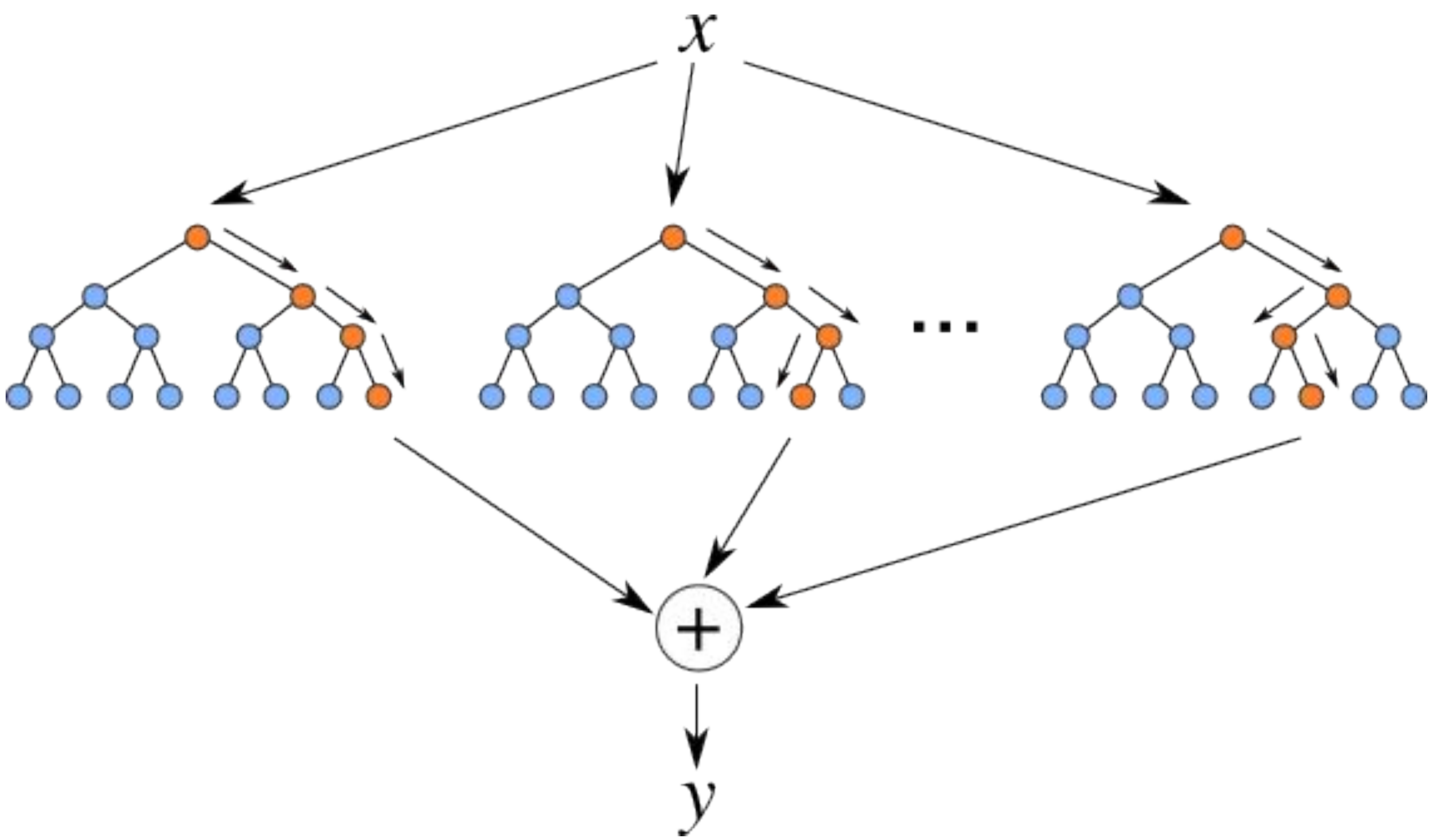
❖ Decision Tree Regression



https://www.researchgate.net/figure/Decision-tree-regression-model-for-RPE-Abbreviation-ACC-acceleration-HSR-high-speed_fig1_351024035

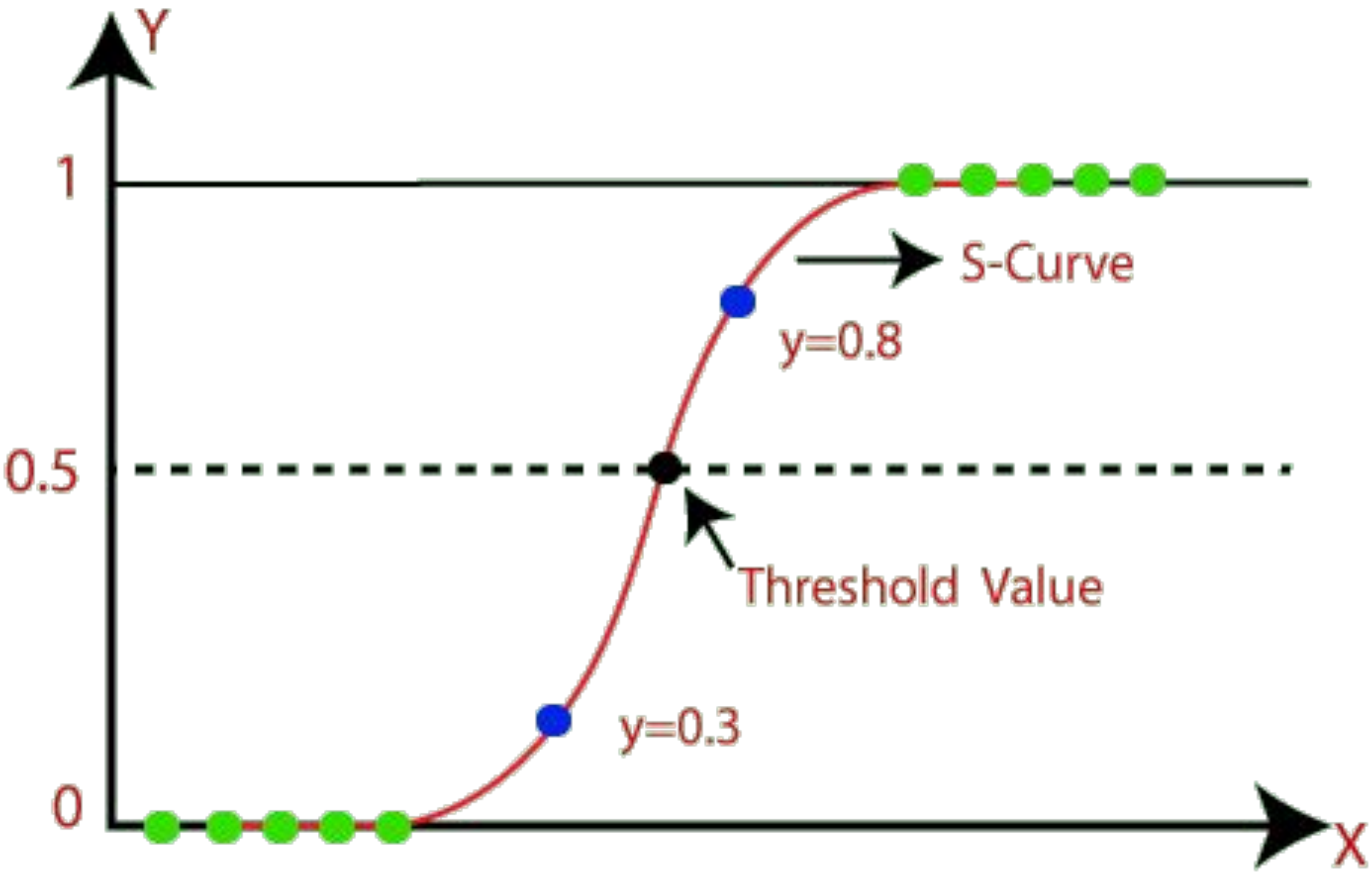
Other

❖ Random Forest Regression



Other

❖ Logistic Regression



QUIZ TIME