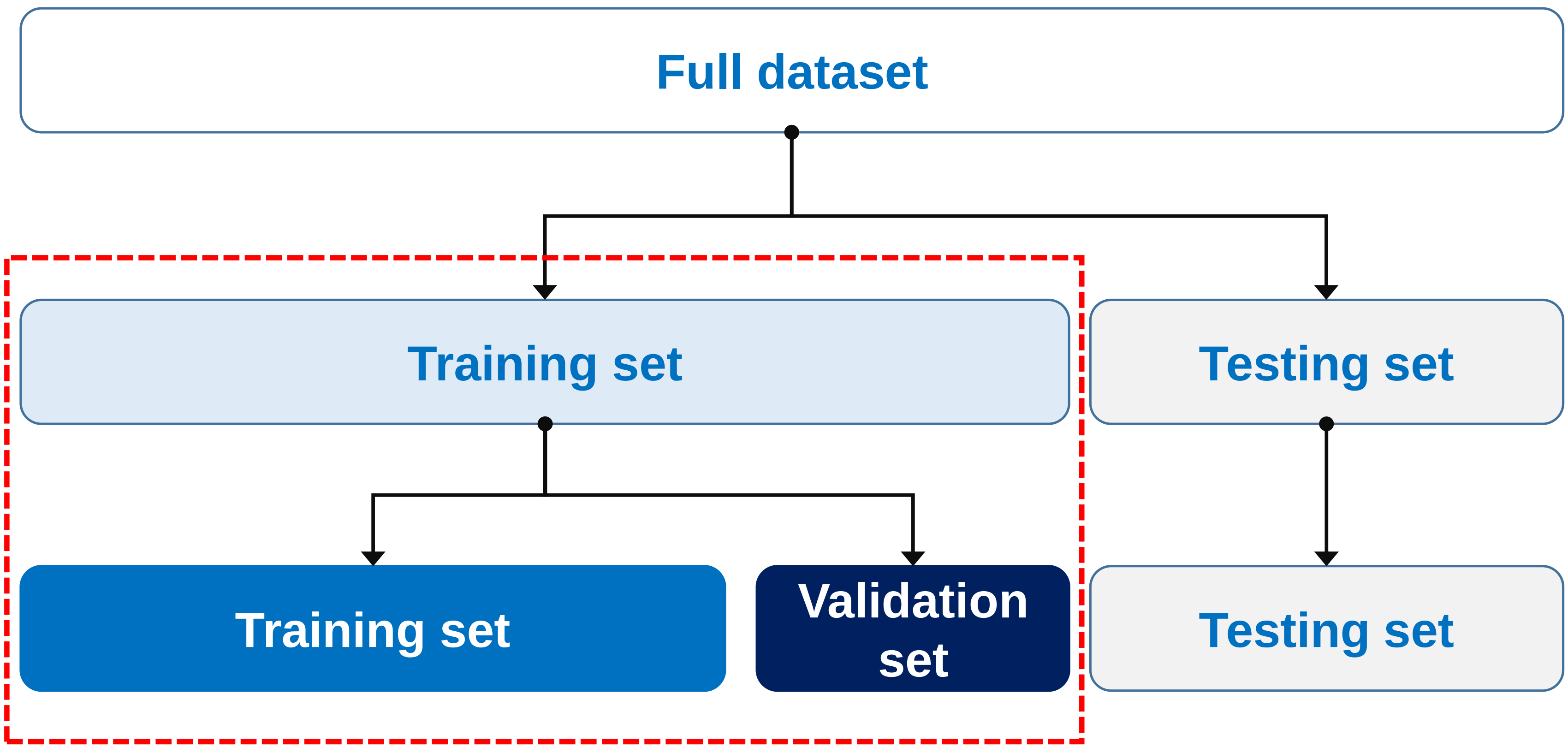
Data Analysis(2)

Dept. of Mechanical System Design Engineering,
Seoul National University of Science and Technology

Prof. Ju Yeon Lee
(jylee@seoultech.ac.kr)

Review

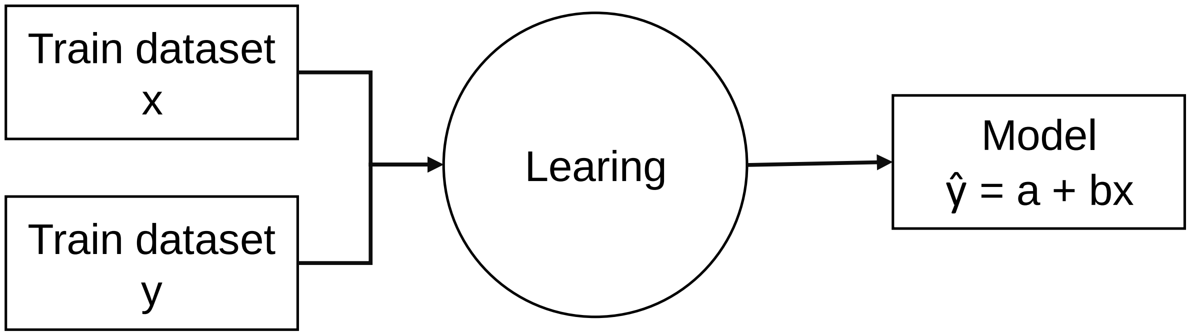
Datasets



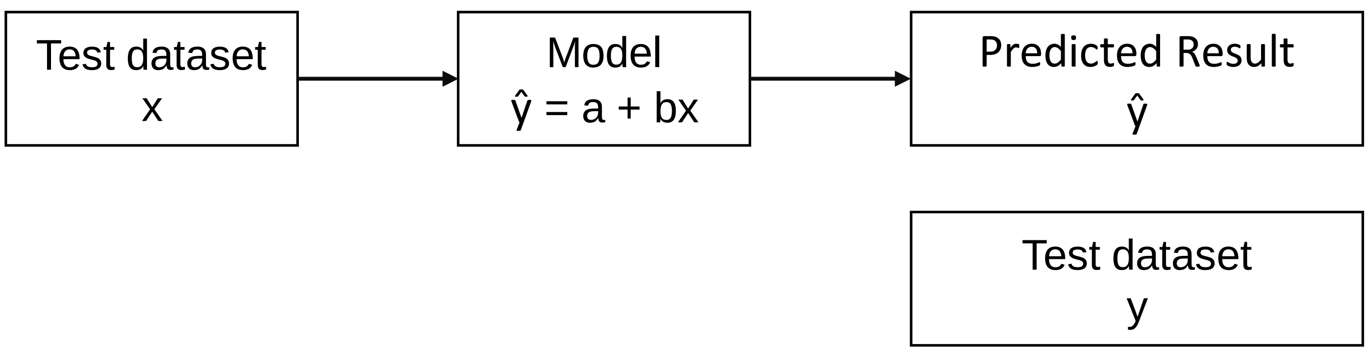
Full dataset
Training set
Testing set
Training set
Validation set
Testing set

Datasets (Supervised Learning)

- **Train dataset : dataset to find the model**



- **Test dataset : dataset to validate the model**



Model Evaluation (Regression)

- **Test dataset : dataset to validate the model**

Test dataset x → Model $\hat{y} = a + bx$ → Predicted Result $\hat{y}$

Test dataset y

Error (Data analysis) or residual (Statistics) : y – ŷ

SSE (Sum of Squared Error)

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Test set | Predicted value | Actual value

MSE (Mean Squared Errors)

$$MSE = \frac{SSE}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Test set | Predicted value | Actual value

RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/n}$$

$$= \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

Test set | Predicted value | Actual value

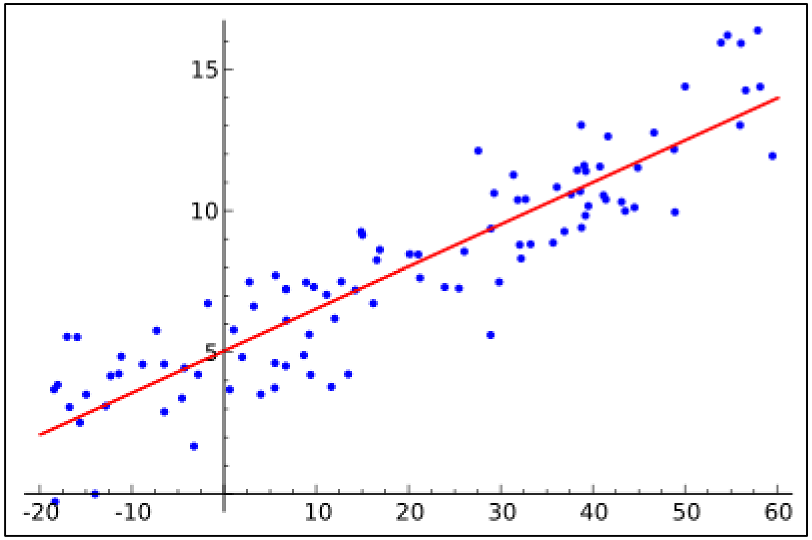
Linear Regression

- **Simple linear regression :**
  The very simplest case of a single scalar predictor variable x and a single scalar response variable y

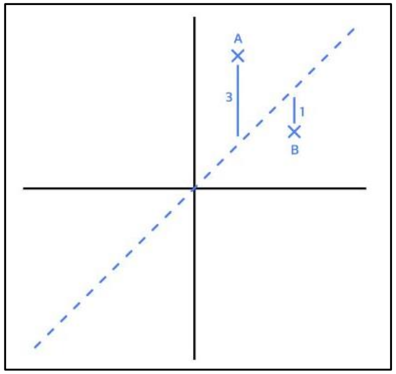
$$\hat{y} = b_0 + b_1 x_1$$

- **Multiple linear regression :**
  generalization of simple linear regression to the case of more than one independent variable; a special case of general linear models, restricted to one dependent variable

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

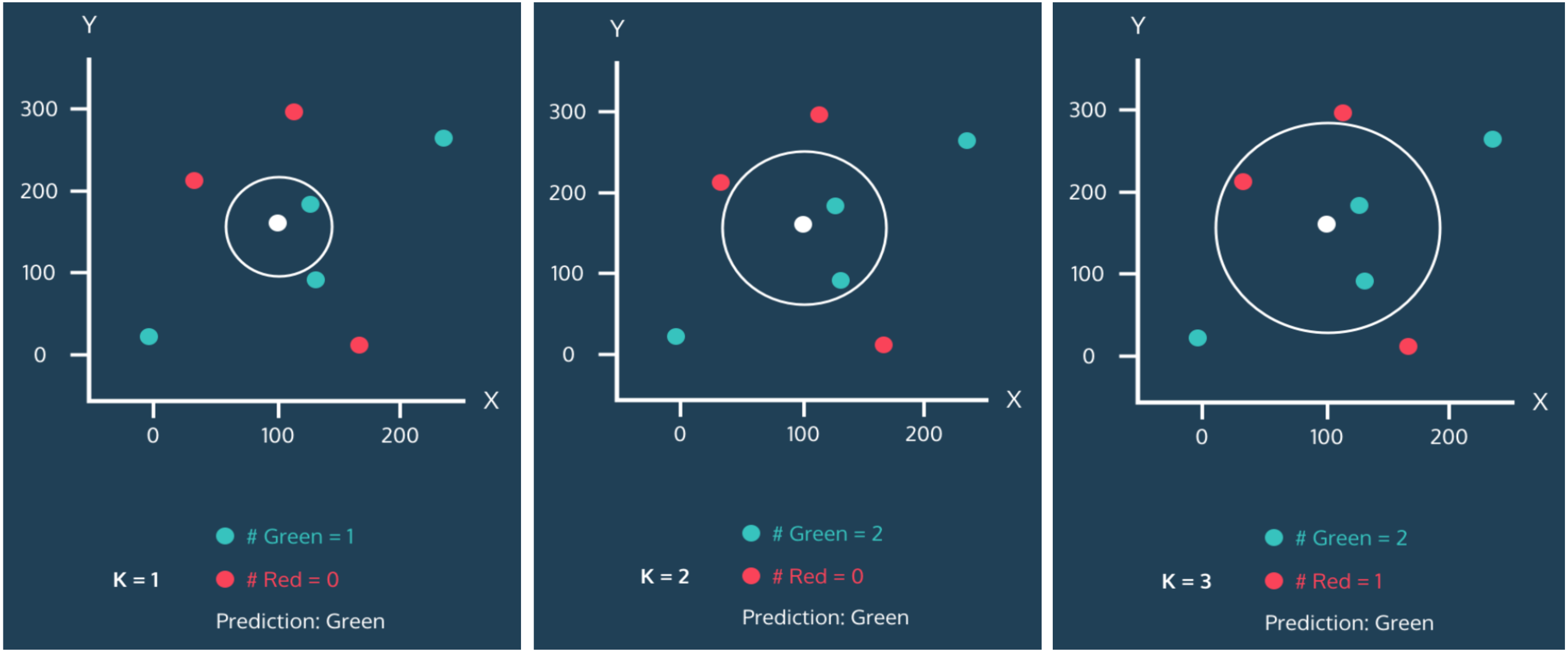


$$MSE = \frac{SSE}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

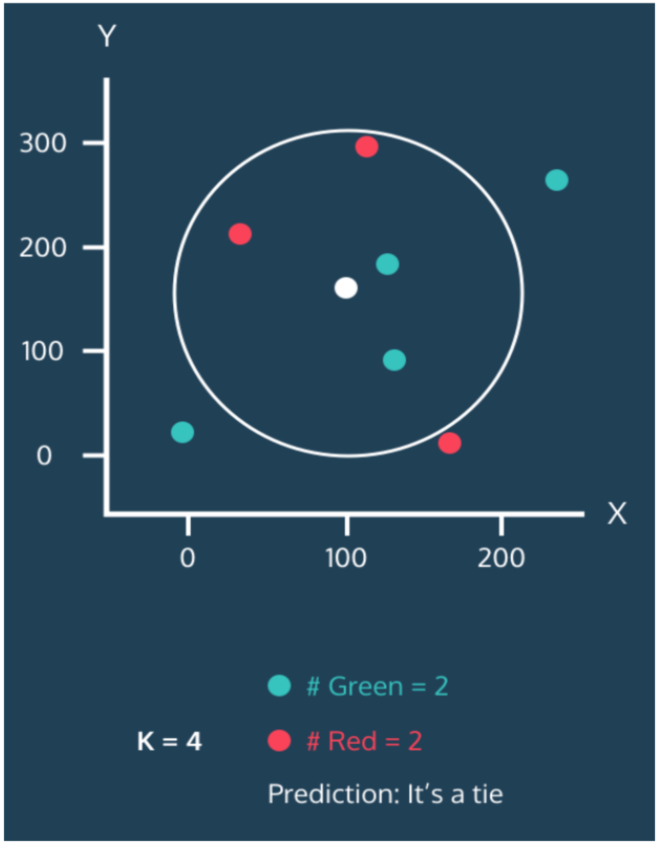


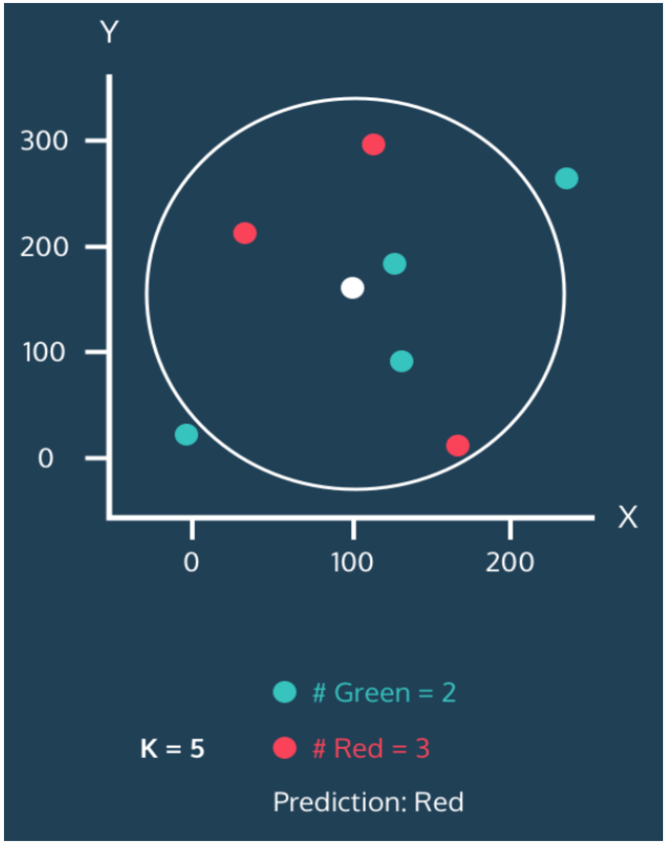
Applying Machine Learning Algorithms for Missing Data

KNN : K-Nearest Neighbor



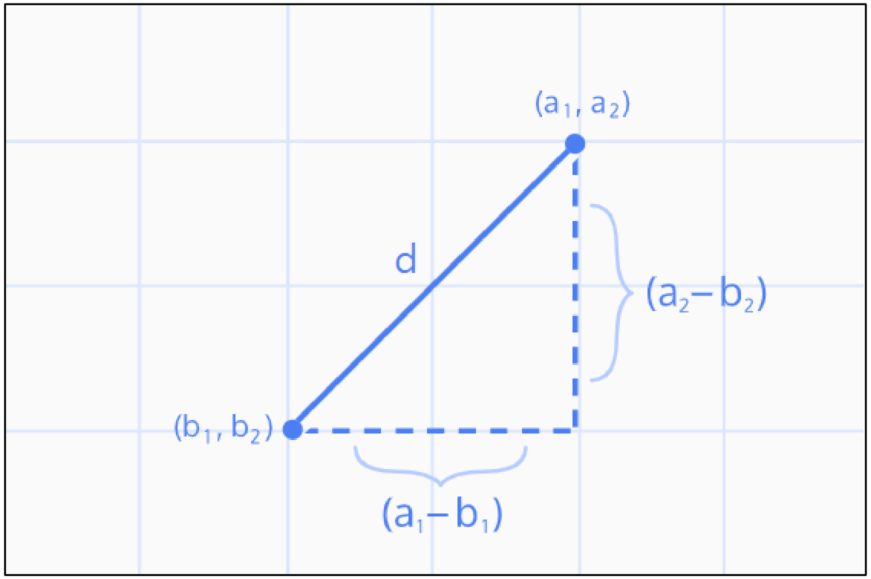
KNN : K-Nearest Neighbor





Distance Formula

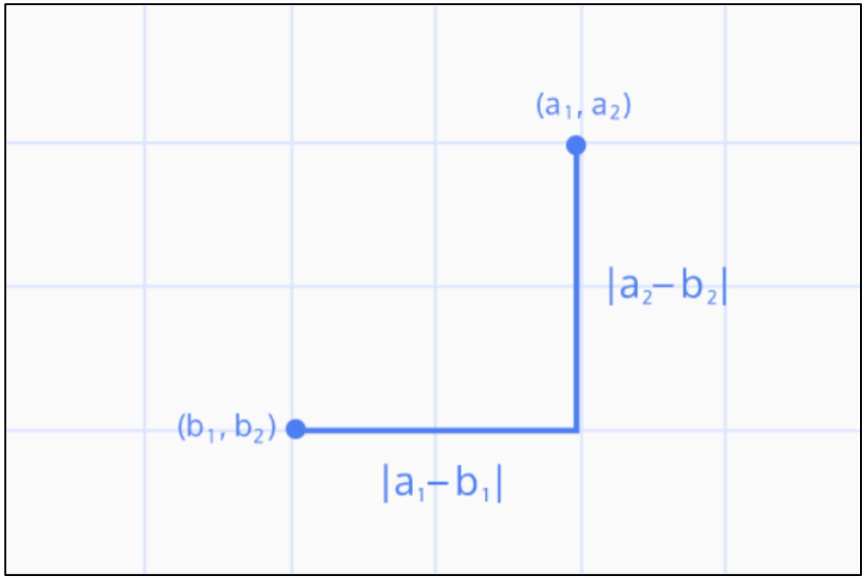
Euclidean Distance



$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Manhattan Distance



$$d = | a_1 - b_1 | + | a_2 - b_2 |$$

$$| a_1 - b_1 | + | a_2 - b_2 | + \ldots + | a_n - b_n |$$

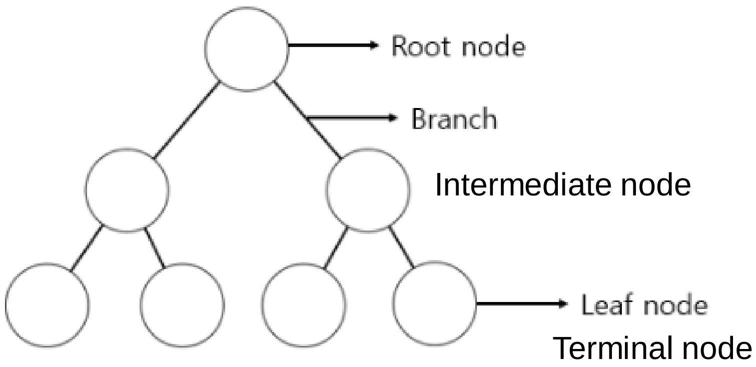
Distance Formula

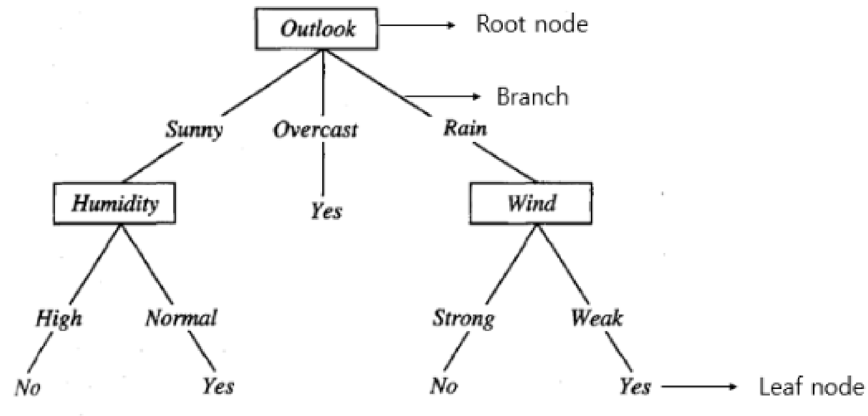
Minkowski Distance

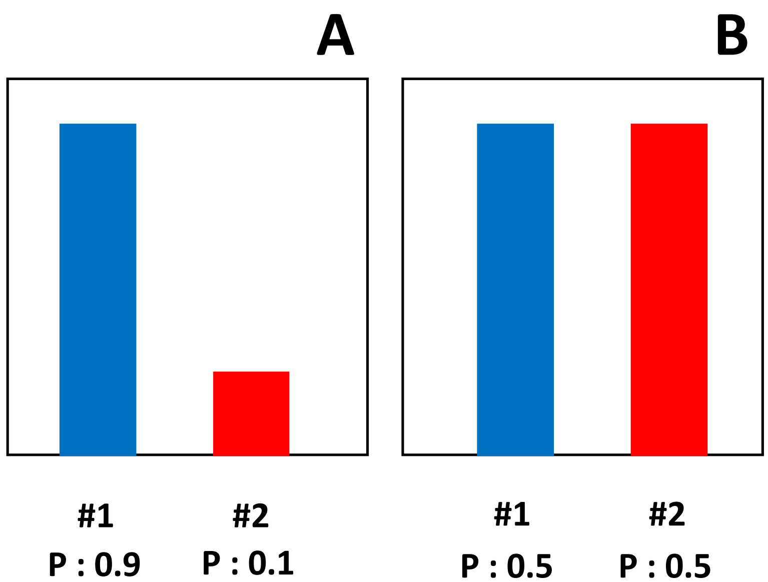
$$D(X,Y) = (\sum_i^n (|x_i - y_i|)^p)^{\frac{1}{p}}$$

- **P = 1, Manhattan Distance**
- **P = 2, Euclidean Distance**

Decision Tree







- *A : 0.9 * 0.1 = 0.09*
- *B : 0.5 * 0.5 = 0.25*
- **Gini Impurity :** $Gini = 1 - \sum_{i=1}^{C} (p_i)^2$
- **Entropy Index :** $H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$