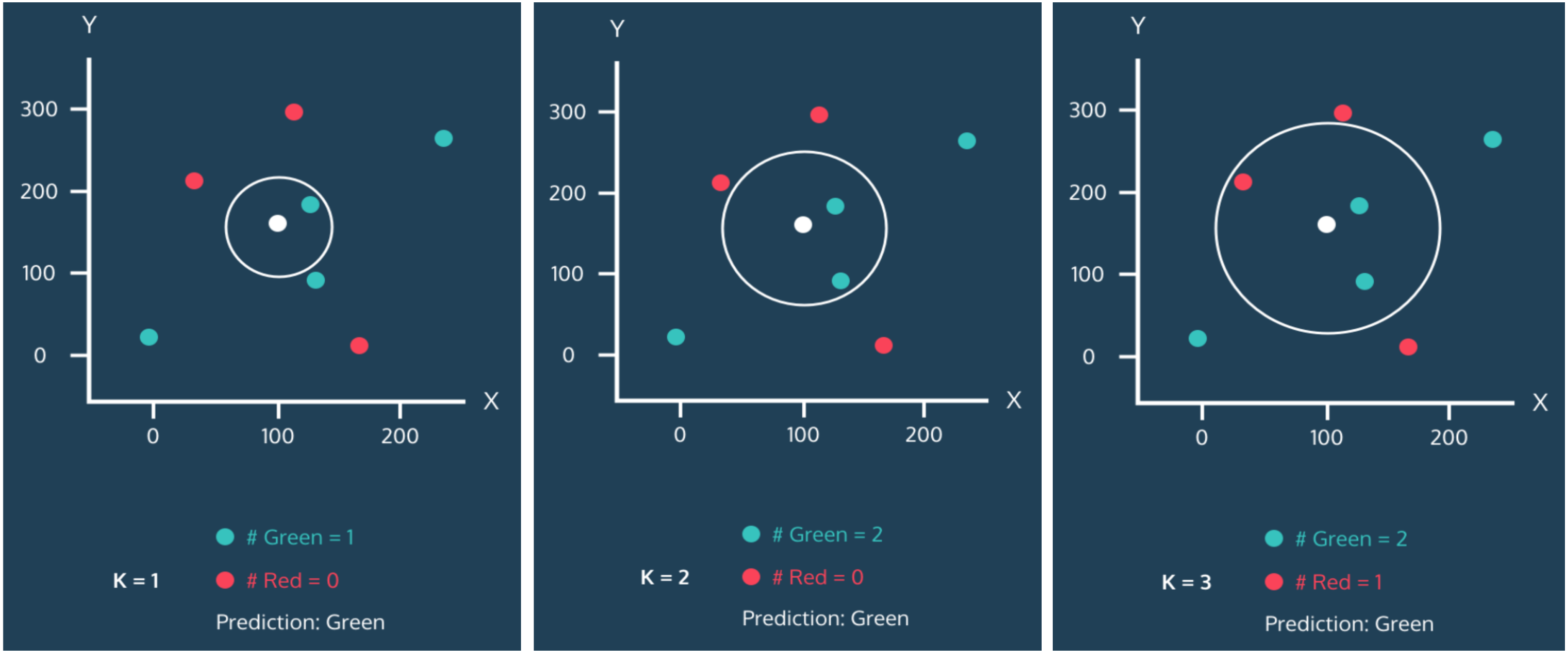
Data Analysis(3)

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

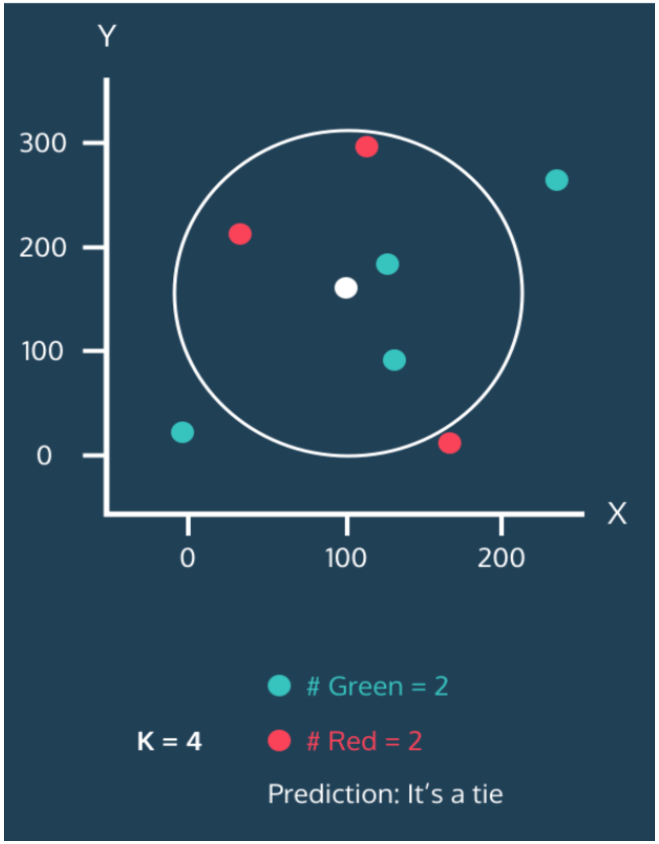
Prof. Ju Yeon Lee
(jylee@seoultech.ac.kr)

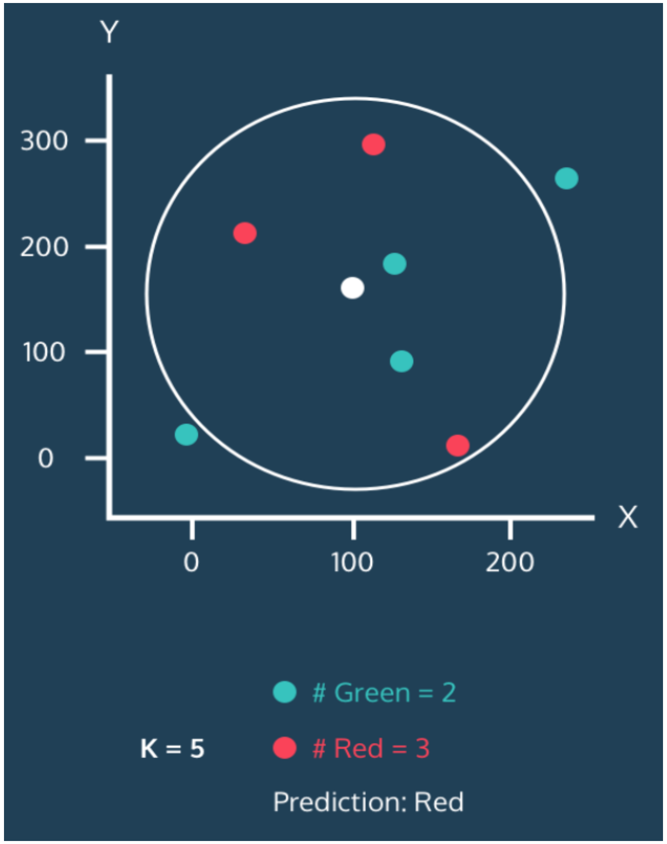
Review

KNN : K-Nearest Neighbor



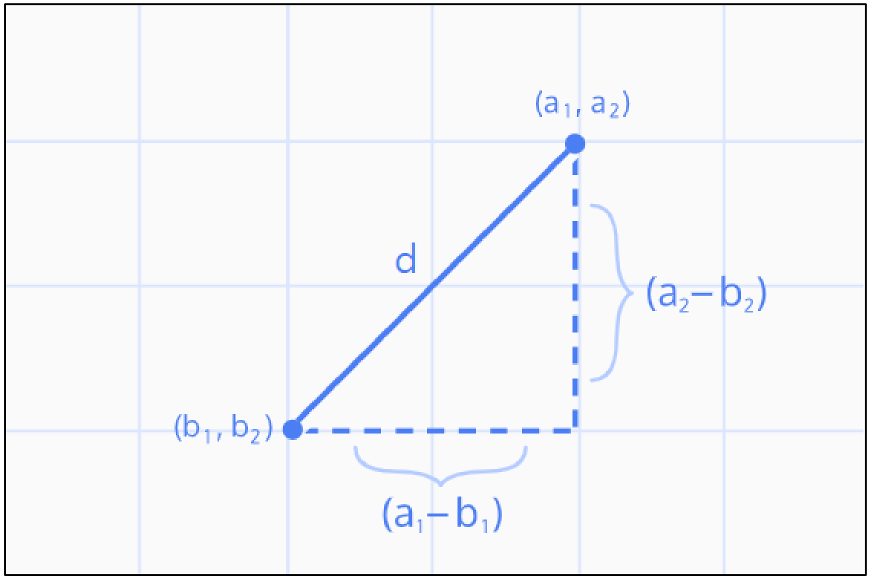
KNN : K-Nearest Neighbor





Distance Formula

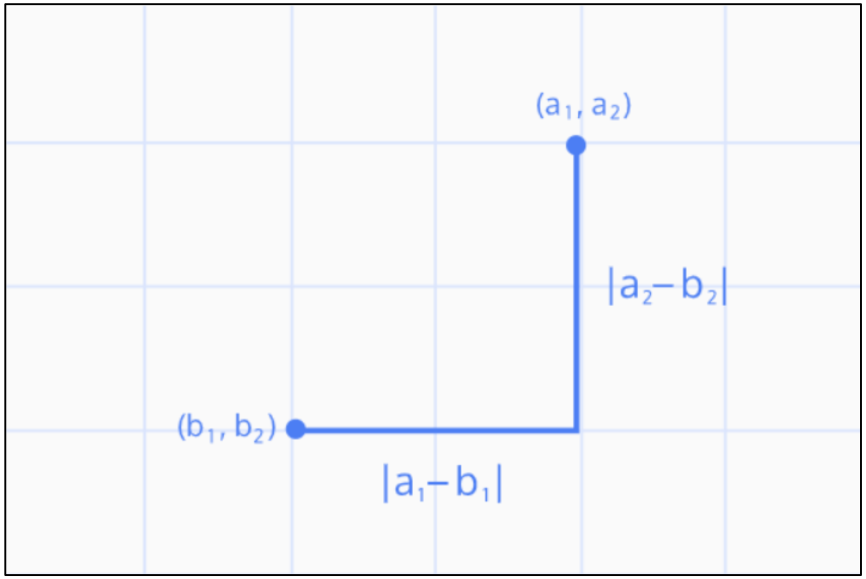
Euclidean Distance



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

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

Manhattan Distance



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

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

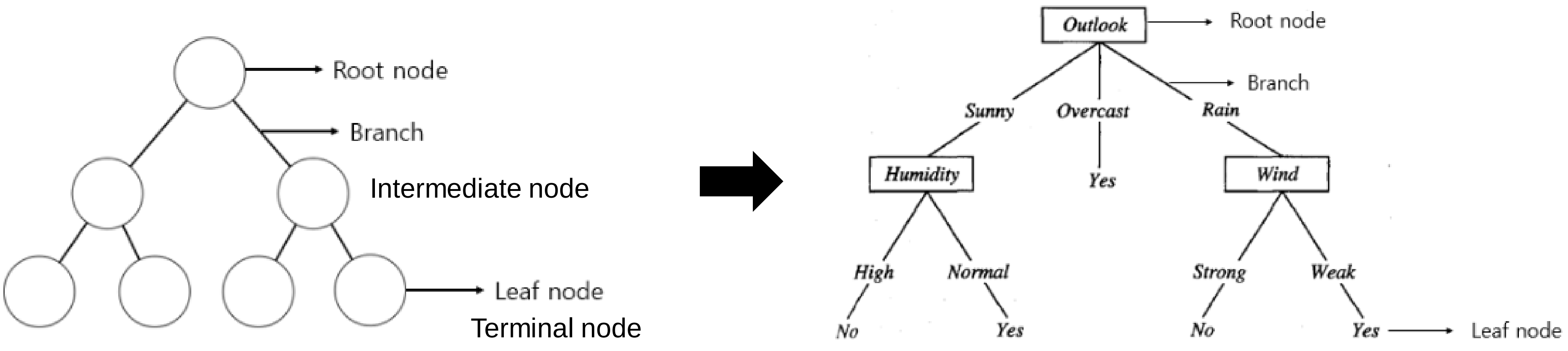
Distance Formula

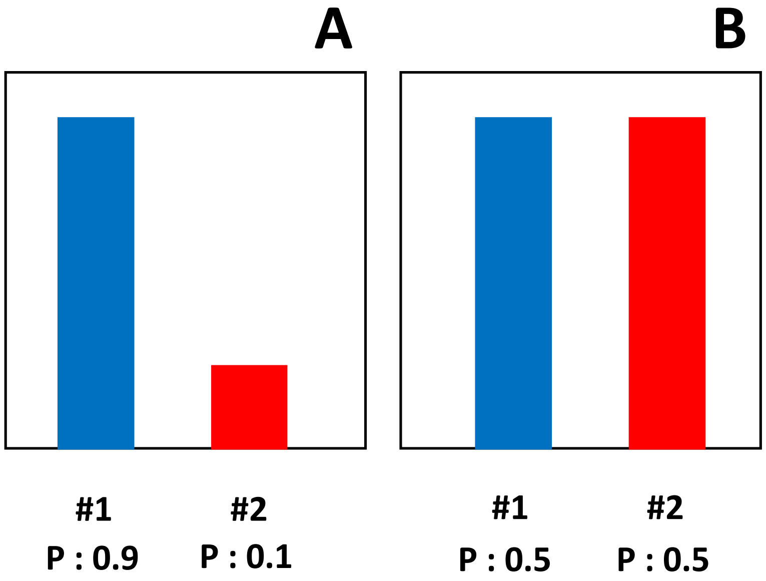
Minkowski Distance

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

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

Decision Tree



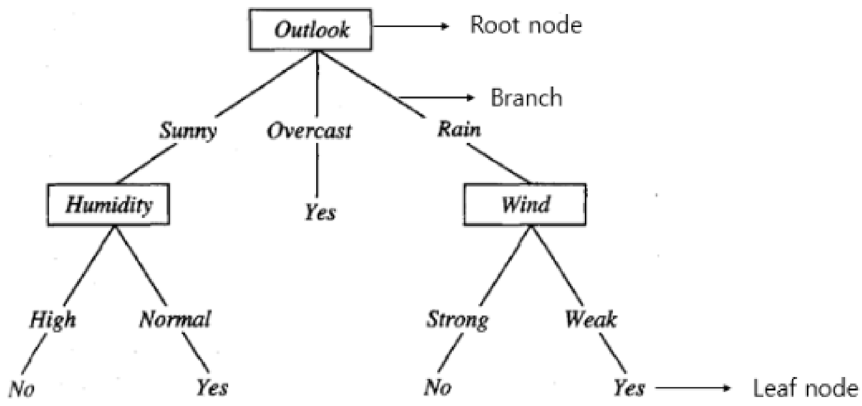


- *A : 0.9 * 0.1 = 0.09*
- *B : 0.5 * 0.5 = 0.25*

- **Gini Impurity :** $Gini=1-\sum_{i=1}^{C}(p_i)^2$

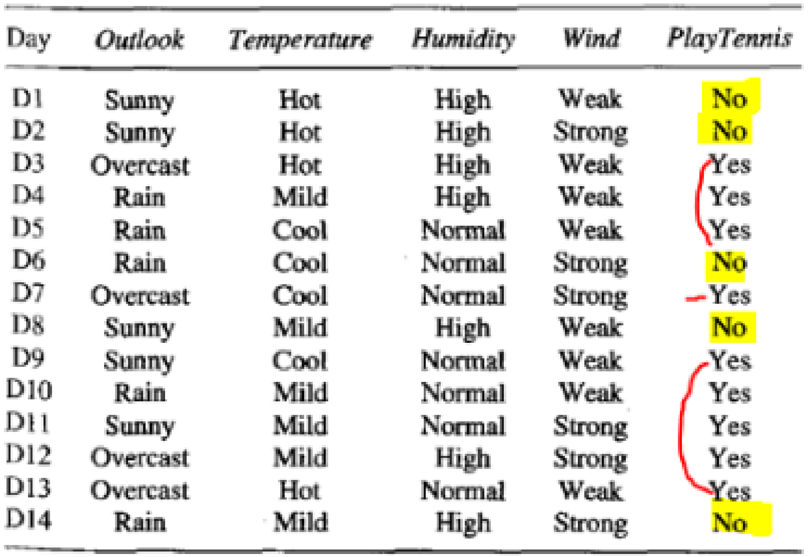
- **Entropy Index :** $H(X)=-\sum_{i=1}^{n}p_i\log_2 p_i$

Decision Tree



$$Ent(D) = -\sum_{k=1}^{|Y|} p_k \, log_2 p_k$$

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

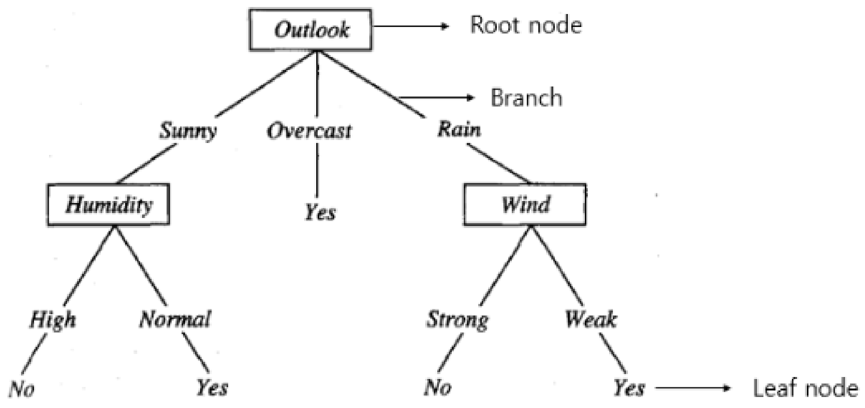


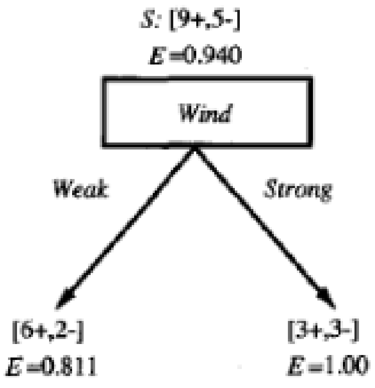
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$\begin{aligned} Values(Wind) &= Weak, Strong \\ S &= [9+, 5-] \\ S_{Weak} &\leftarrow [6+, 2-] \\ S_{Strong} &\leftarrow [3+, 3-] \end{aligned}$$

$$\begin{aligned} Gain(S, Wind) &= Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S) - (8/14) Entropy(S_{Weak}) \\ &\quad - (6/14) Entropy(S_{Strong}) \\ &= 0.940 - (8/14)0.811 - (6/14)1.00 \\ &= 0.048 \end{aligned}$$

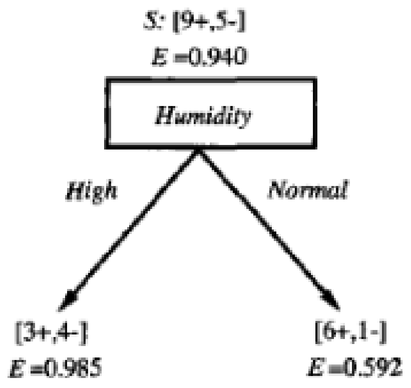
Decision Tree





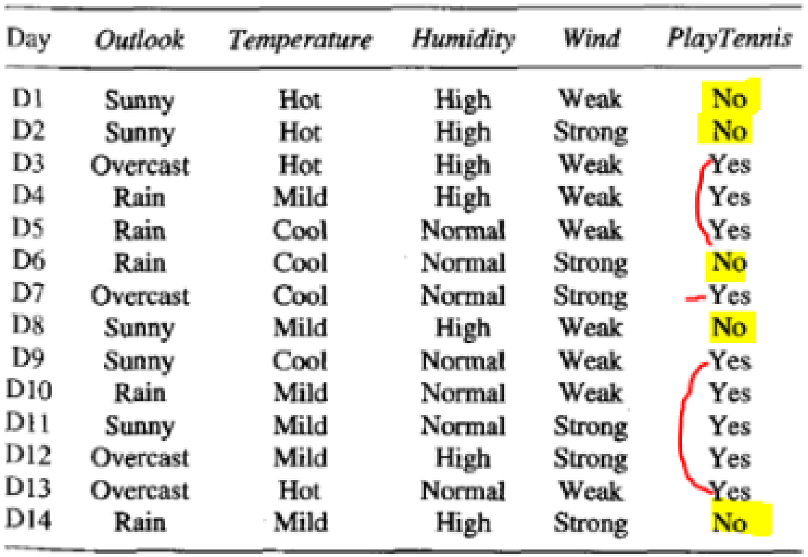
Gain (S, Wind)

$= .940 - (8/14).811 - (6/14)1.0$
$= .048$



Gain (S, Humidity)

$= .940 - (7/14).985 - (7/14).592$
$= .151$



Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$\begin{aligned} Gain(S, Outlook) &= 0.246 \\ Gain(S, Humidity) &= 0.151 \\ Gain(S, Wind) &= 0.048 \\ Gain(S, Temperature) &= 0.029 \end{aligned}$$

Sampling

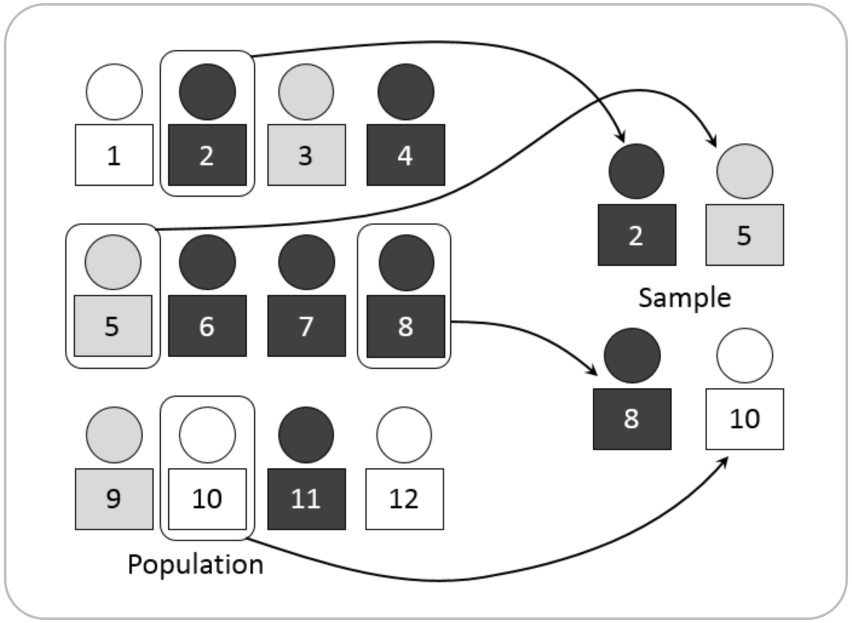
Sampling?

"Sampling" is the selection of a subset or a statistical sample (termed sample for short) of individuals from within a statistical population to estimate characteristics of the whole population

Source : Wikipedia

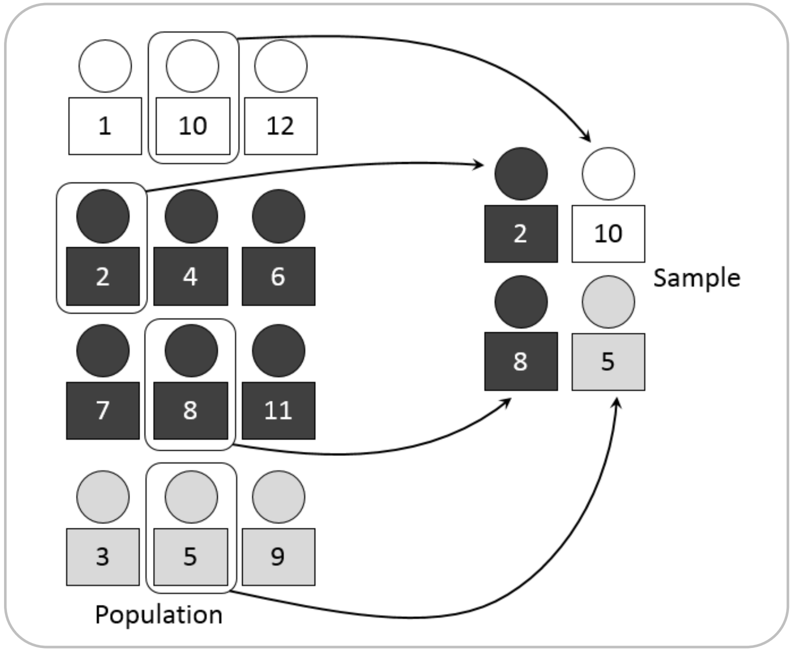
- **Sampling Type :**
  - ✓ Balanced sampling : Balance between data classes
    → **Simple random sampling (단순 임의 샘플링)**
  - ✓ Imbalanced/unbalanced sampling : Imbalance between data classes
    → **Stratified sampling (층화 추출)**
    → **Systematic sampling (계통 추출)**

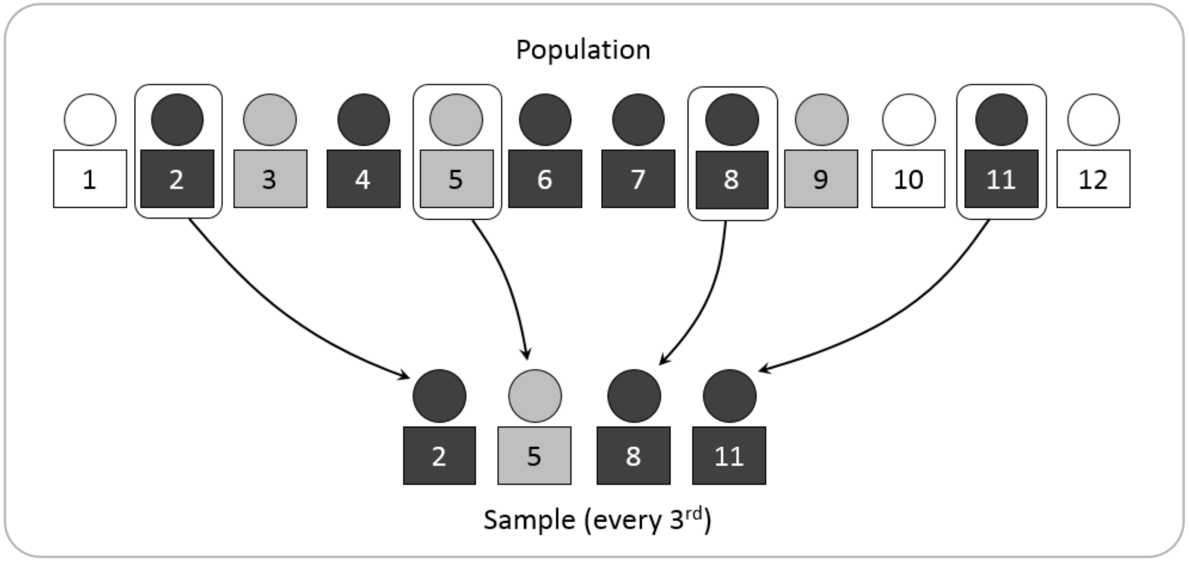
- **Simple random sampling (단순 임의 샘플링)**
  All subsets of a sampling frame have an equal probability of being selected. Each element of the frame thus has an equal probability of selection



Sampling?

- **Stratified sampling (층화 추출)**
  When the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected

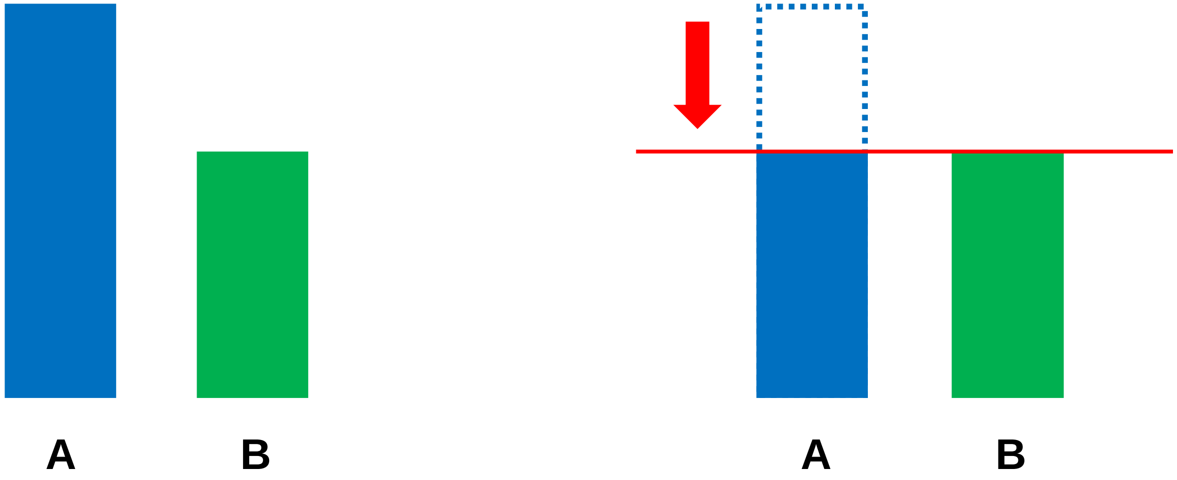




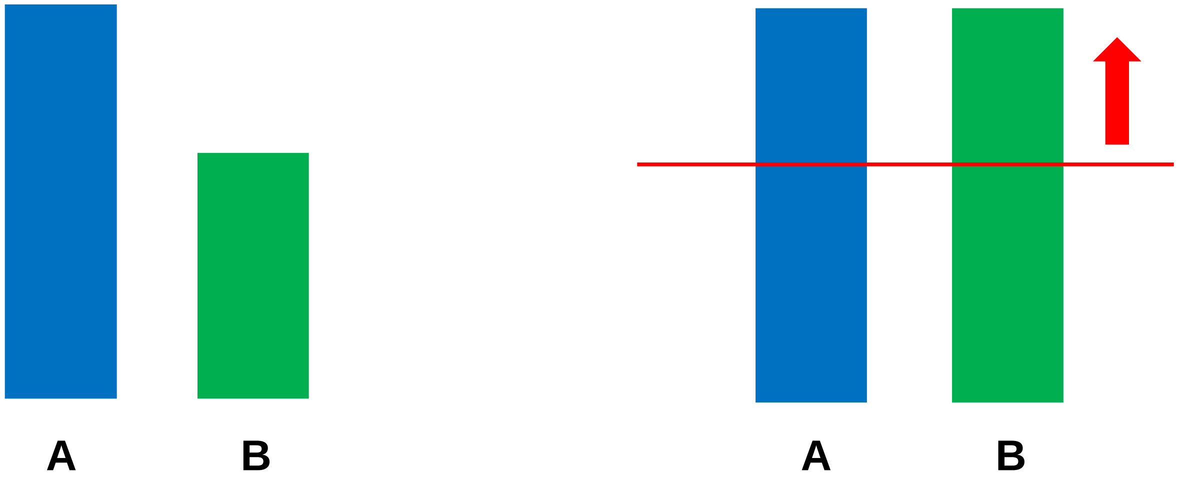
- **Systematic sampling (계통 추출)**
  Systematic sampling (also known as interval sampling) relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list

Imbalanced sampling

1. Under/Down Sampling

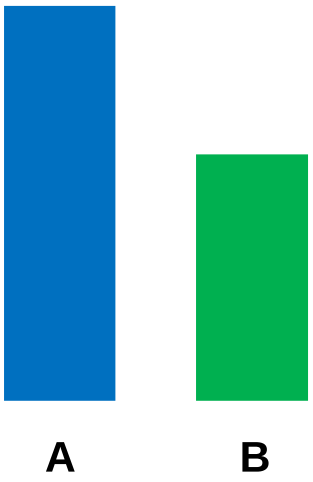


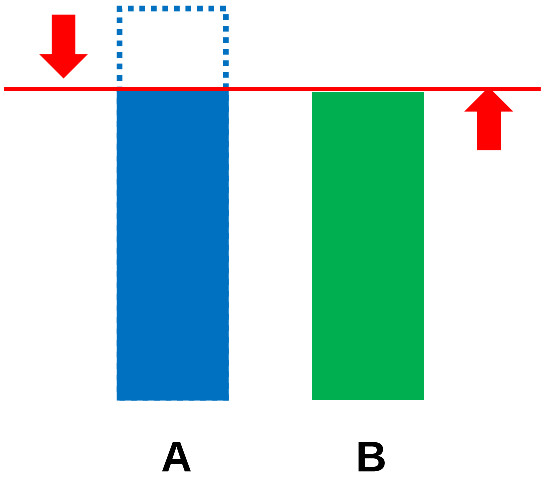
2. Over/Up Sampling



Imbalanced sampling

3. Combination Sampling : Under + Over



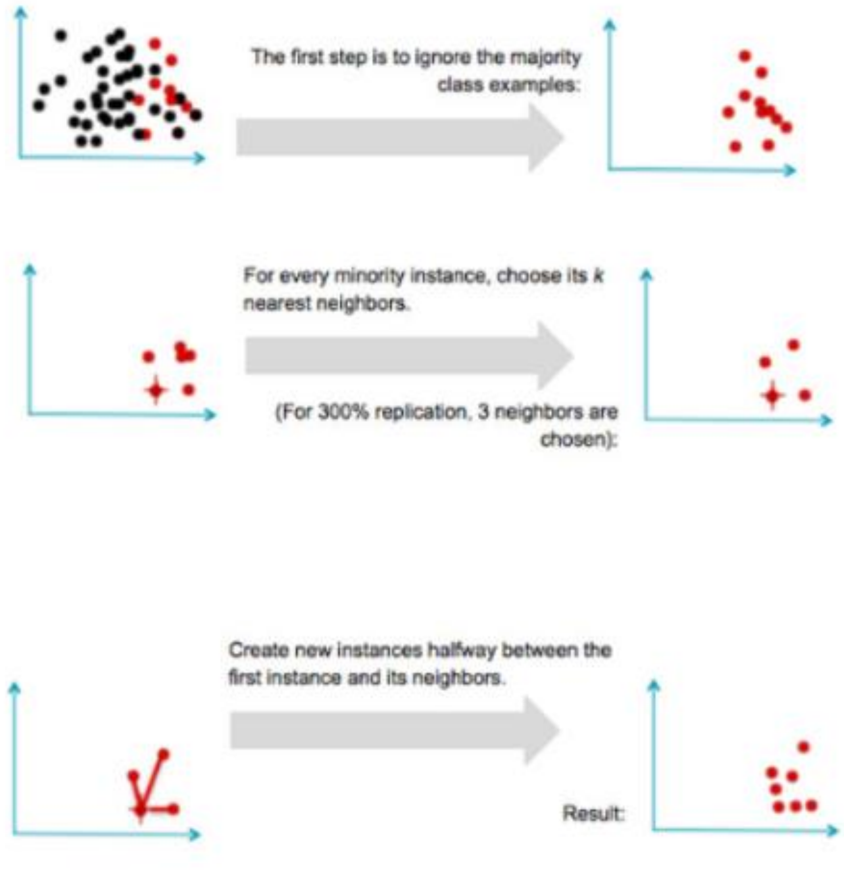


SMOTE

SMOTE(Synthetic Minority Over-Sampling Technique) :

generate new data between neighboring minority classes from random minority class data

For numerical features



The first step is to ignore the majority class examples:

For every minority instance, choose its *k* nearest neighbors.

(For 300% replication, 3 neighbors are chosen):

Create new instances halfway between the first instance and its neighbors.

Result:

SMOTENC(Synthetic Minority Over-Sampling Technique for Nominal and Continuous) :

For dataset containing numerical and categorical features

However, it is not designed to work with only categorical features

https://imbalanced-learn.org/stable/references/index.html

Thank you

Q & A