Introduction to Machine Learning

KU-BIG 학술부

Index

KU-BIG

1. What is Machine Learning?

Artificial Intelligence

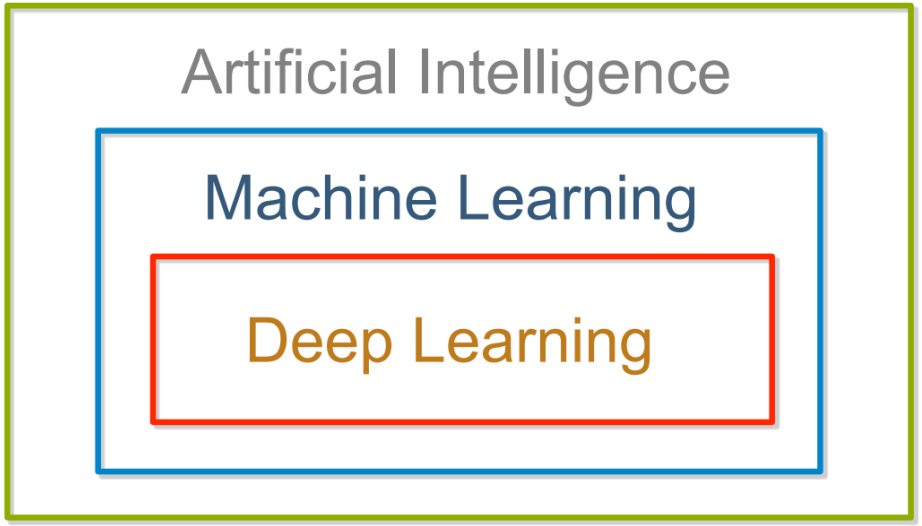
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI technique which use statistical methods to enable machines to improve with experiences.

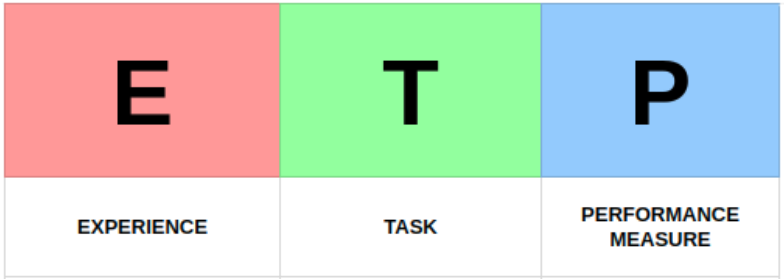
Deep Learning

Subset of ML which make the computation of muti-layer neural networks feasible.



1. What is Machine Learning?

"만약 작업 T에 대해 기준 P로 측정한 성능이 경험 E로 인해 향상되었다면, 그 프로그램은 작업 T에 대해 기준 P의 관점에서 경험 E로부터 "배웠다"라고 말할 수 있다." – Tom Mitchell



Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

1. What is Machine Learning?

[The training data set]

X	Y
{3,6,9}	3
{2,5,7}	2
{2,3,5}	1

training

[The ML model]

Q. X = {4,7,8}

A. Y = ????

KU-BIG

1. What is Machine Learning?

Supervised Learning

- Regression
- Classification

Semi- Supervised Learning

Unsupervised Learning

- Clustering
- Visualization
- Dimensionality Reduction
- Association Rule

Reinforcement Learning

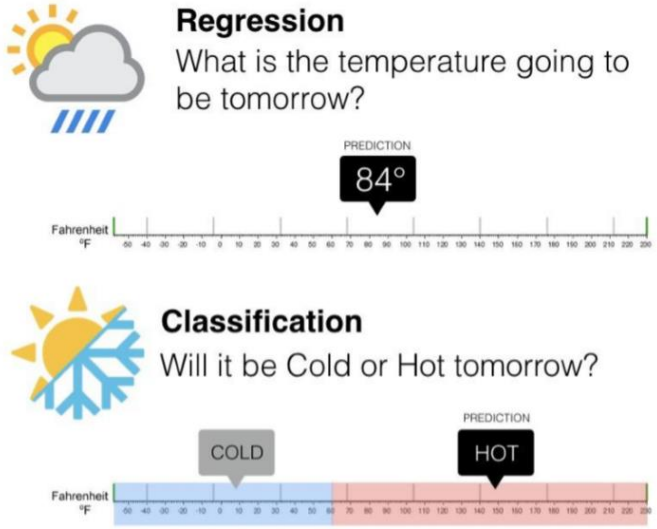
1. What is Machine Learning?

Supervised Learning

ex) KNN / Linear Regression / Logistic Regression / SVM / Decision Tree

X	y
10	90
9	80
3	50
2	30

Both X(predictor variable) and y(targe variable) are given



Regression

- Output variable is continuous
- Predict y(the target variable) with given X(the predictor variable)
- Ex) temperature, value, price etc.

Classification

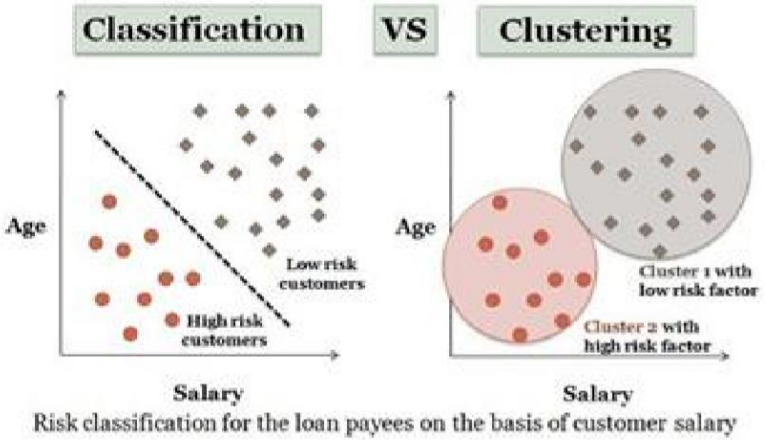
- Output Variable is discrete
- Binary Classification [1 or 0]
- Multi-label Classification [a \ b \ c \ d]

KU-BIG

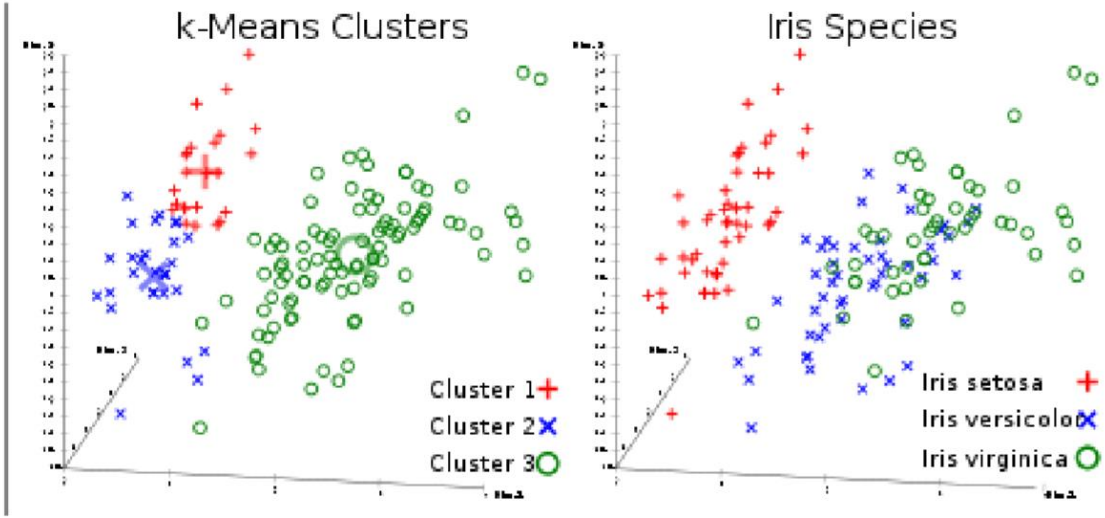
1. What is Machine Learning?

Unsupervised Learning

ex) Clustering / Dimensionality Reduction / Visualization



Risk classification for the loan payees on the basis of customer salary

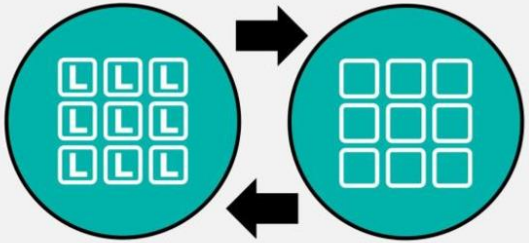


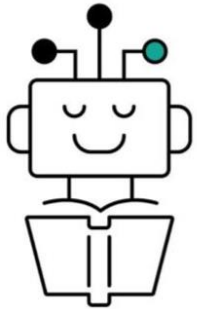
KU-BIG

1. What is Machine Learning?

Semi- Supervised Learning

Semi-supervised learning uses a combination of supervised labeled training data with unsupervised methods:





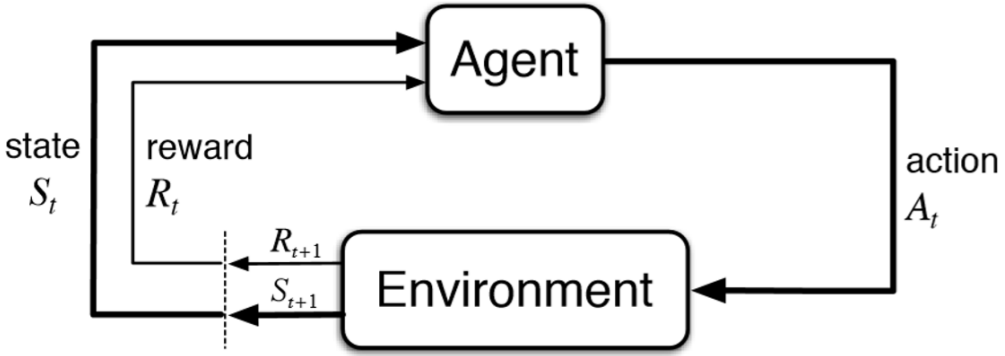
Semi-Supervised Learning

1. Load labeled input training data.
2. Model is trained on the data.
3. Present unlabeled raw data.
4. Algorithm infers classifiers for unlabeled data on its own.
5. High-confidence data is added to labeled training data set.
6. Algorithm progressively adapts and learns.

- The Label(pre-determined output value) is partially given

1. What is Machine Learning?

Reinforcement Learning



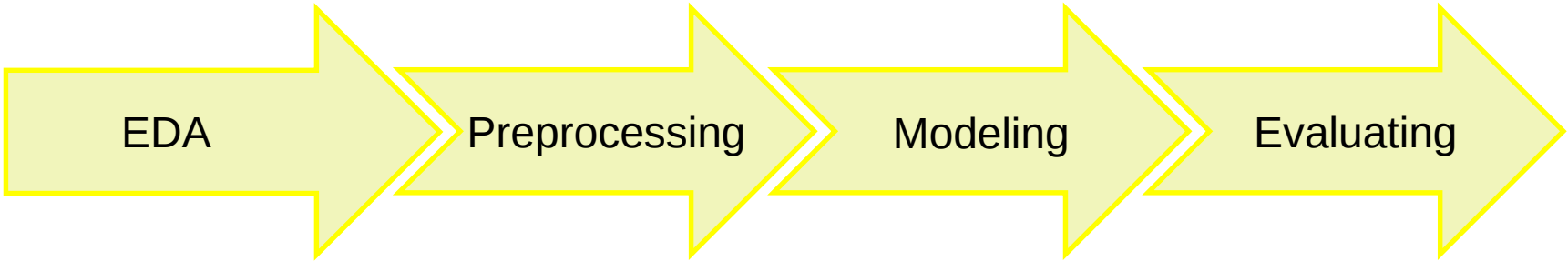
Reinforcement Learning

- Database of every professional players
- Use Reinforcement Learning method for the training
- Every decision made for the optimal outcome

New generation : AlphaGo [zero]

- No input of human data anymore
- Endless re-generation of the random training data
- Strategies that was never seen before

2. Outline of Machine Learning



- ✓ 상관관계 파악
- ✓ 기초 통계량 분석

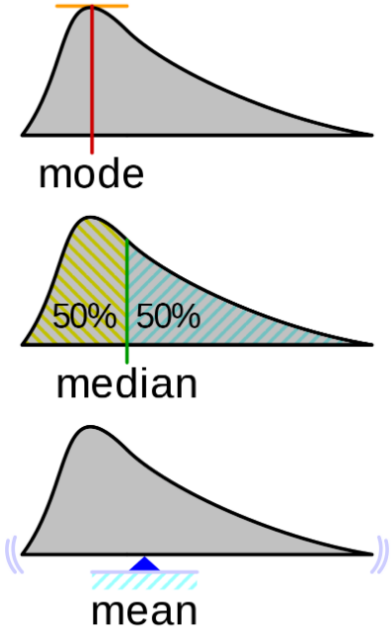
- ✓ 이상치/결측치 처리
- ✓ 정규화
- ✓ 데이터 불균형 처리
- ✓ 차원 축소

- ✓ 적합한 모델 선택

- ✓ hyperparameter tuning
- ✓ 성능 평가

3. Descriptive Statistics

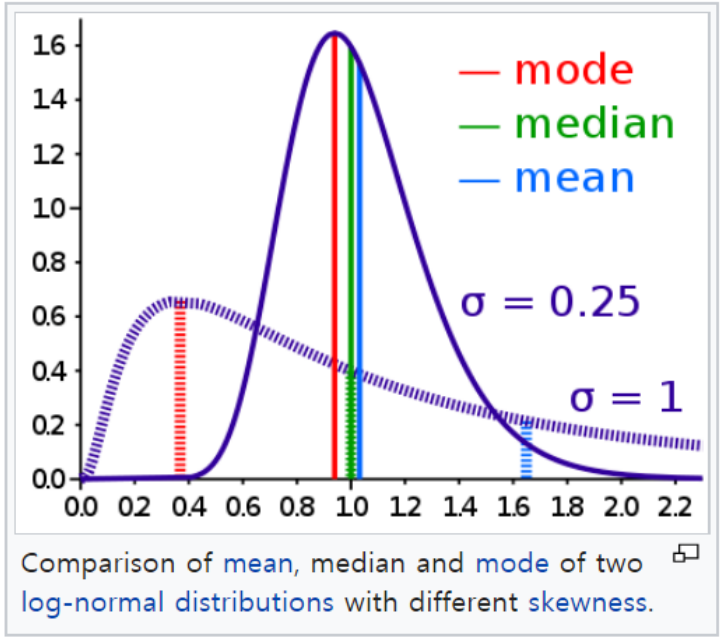
(1) Center : Mean, Median, Mode



가장 많이 관측되는 수

$$\mathrm{P}(X \leq m) \geq \frac{1}{2} \quad \wedge \quad \mathrm{P}(X \geq m) \geq \frac{1}{2}$$

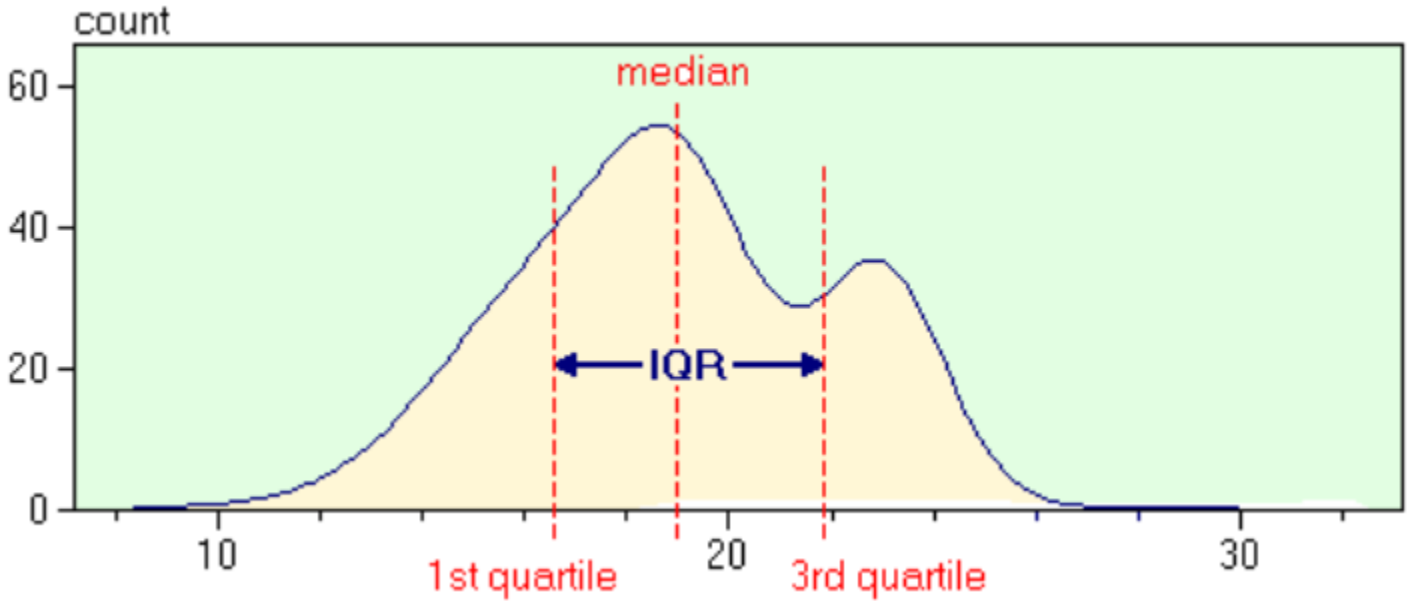
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$



Comparison of mean, median and mode of two log-normal distributions with different skewness.

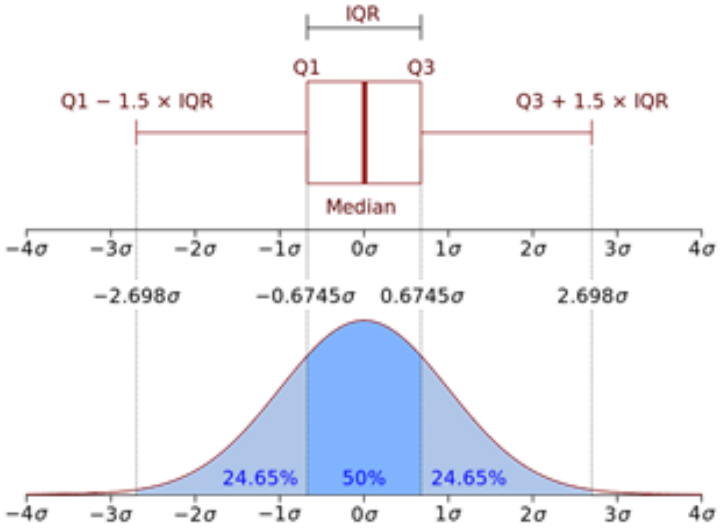
3. Descriptive Statistics

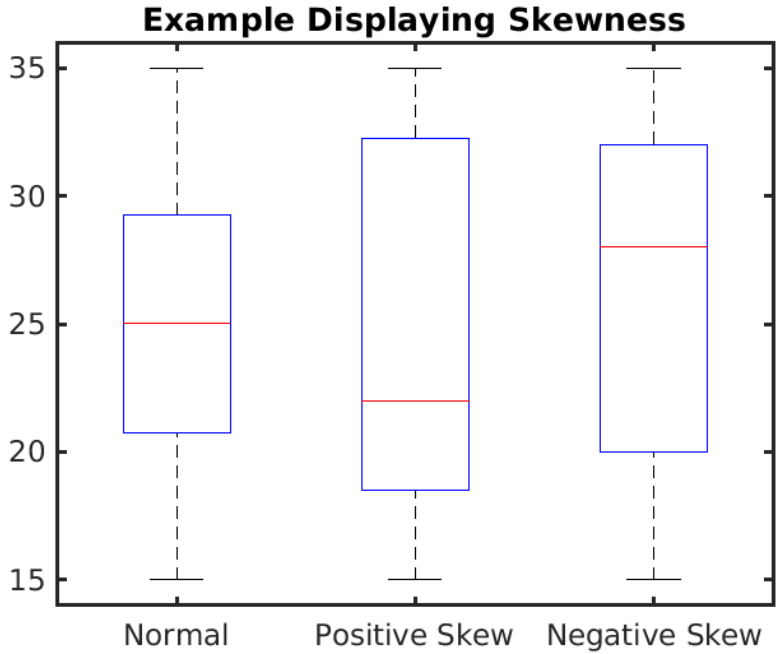
(2) Dispersion : Range, Quartile, IQR



3. Descriptive Statistics

(3) Graphics – ex) Boxplot





4. Preprocessing

Outlier & Missing value

- ✓ Outlier 제거
- ✓ 결측치 포함된 행/열 제거
- ✓ 새로운 값으로 대체

Dimensional Reduction

- ✓ PCA(Feature Extraction)
- ✓ Feature Selection

Normalization

- ✓ 최대-최소 정규화
- ✓ Z-score 정규화
- ✓ 십진스케일 정규화

Imbalanced Data

- ✓ Undersampling
- ✓ Oversampling