

머신 러닝의 이해

Understanding Machine Learning

머신 러닝의 이해

CONTENTS

1. 머신 러닝의 개요	02
1-1 인공지능(Artificial Intelligence)이란?	
1-2 머신 러닝의 목적	
1-3 머신 러닝의 구성	
1-4 머신 러닝의 프로세스	
1-5 머신 러닝의 학습 방법	
2. 머신 러닝의 분류	09
2-1 지도 학습(Supervised Learning)	
2-2 비지도 학습(Unsupervised Learning)	
3. 머신 러닝 모델 평가	29
3-1 수치 예측 모델 평가	
3-2 범주 예측 모델 평가	
3-3 클러스터링 모델 평가	
4. 데이터 분할	34
4-1 데이터셋(Dataset)의 종류	
4-2 K-겹 교차 검증(K-fold Cross Validation)	

1. 머신 러닝의 개요

인공지능이란?

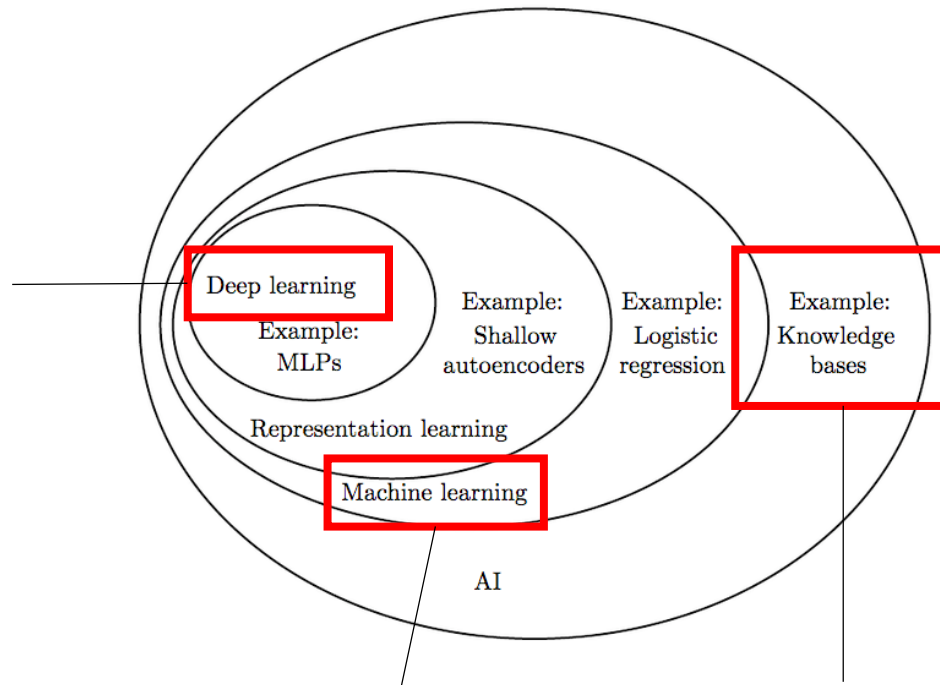
Artificial Intelligence

“의사결정이 가능한 기계”

$AI \supset ML \supset DL$

딥 러닝 (Deep Learning)

많은 데이터를 바탕으로
의사결정이 가능한 규칙과 적절한
특징(feature)들을 기계가 생성



머신 러닝 (Machine Learning)

많은 데이터를 바탕으로 의사결정이
가능한 규칙을 기계가 생성

규칙 기반 전문가 시스템

(Rule-based Expert System)
전문가의 지식을 바탕으로 의사결정이
가능한 규칙을 사람이 생성

1. 머신 러닝의 개요

머신 러닝의

Machine Learning

목적

머신 러닝의 주 목적은

데이터의 알려진 속성들을 학습하여 예측 모델을 만드는 데 있다.

예측 Prediction

수치를 예측하는 것

분류 Classification

미리 정해진 카테고리 중 어디에 속하는지 판별

클러스터링 Clustering

같이 자주 발생하는 연관성, 패턴 찾기

연관관계 분석 Association Analysis

비슷한 성격의 항목들을 그룹으로 만들기

추천 Recommender

대부분 분석의 결론은 추천의 형태를 갖는다

통계와 머신 러닝의 차이는?

- 통계는 원인을 찾기 위한 목적 ⇒ (현상의 이해와 설명, 데이터 분포, 특성 파악)
- 머신 러닝은 예측을 위한 목적

머신 러닝의 Machine Learning 구성

머신 러닝은 학습, 모델, 알고리즘으로 구성된다.

Learning

데이터를 기반으로 규칙을 만드는 과정 (분류 기준 생성)

Model

분류 예측 추정 등의 목적에 부합하는 데이터 처리가 가능한 집합체

예) KNN, Linear 모델, Decision Tree, SVM, PCA, NMF, K-Means, DBSCAN, CNN, RNN, GAN, LSTM, GRU

Algorithm

모델을 최적화 하기 위한 학습방법

예) Loss Function and Optimization
: 최소제곱법(Least Squared), 경사하강법(Gradient Descent),
역전파(Backpropagation), 엔트로피 최소화(Entropy)

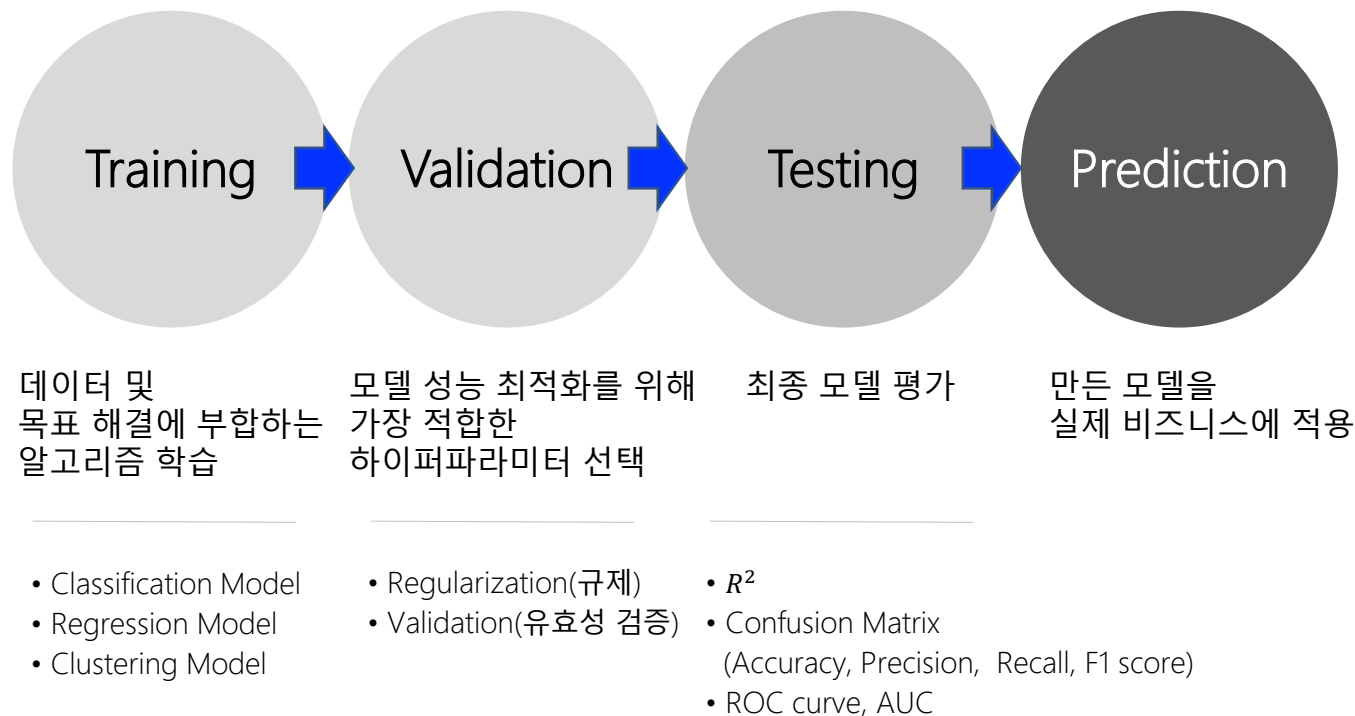
1. 머신 러닝의 개요

머신 러닝의

Machine Learning

프로세스

1. 데이터 기반 학습을 통해 적절한 **알고리즘을 선택**하고
2. **모델을 완성**하여
3. **새로운 데이터를 예측**한다.



1. 머신 러닝의 개요

머신 러닝의 Machine Learning 프로세스

변수 Variable, Feature, Attribute, Factor, Field, Column, ...

현상들을 설명/표현하는 요소

- Predictor variables(예측변수)
- Input variables(입력변수)
- Independent(독립변수)
- Target variables(타겟변수)
- Output variables(출력변수)
- Dependent variables(종속변수)

현상을 측정하는 단위

- Point(포인트)
- Sample(샘플)
- Instance(인스턴스)
- Record(레코드)
- Observation(관측치)

id	X_1	X_2	...	X_p	Y
1	x_{11}	x_{12}	...	$x_{1,p}$	y_1
2	x_{21}	x_{22}	...	$x_{2,p}$	y_2
...
n	$x_{n,1}$	$x_{n,2}$...	$x_{n,p}$	y_n

1. 머신 러닝의 개요

머신 러닝의 Machine Learning 프로세스

머신 러닝의 분석 패턴



학습알고리즘

: 오차값이 최소화 되도록 반복 학습

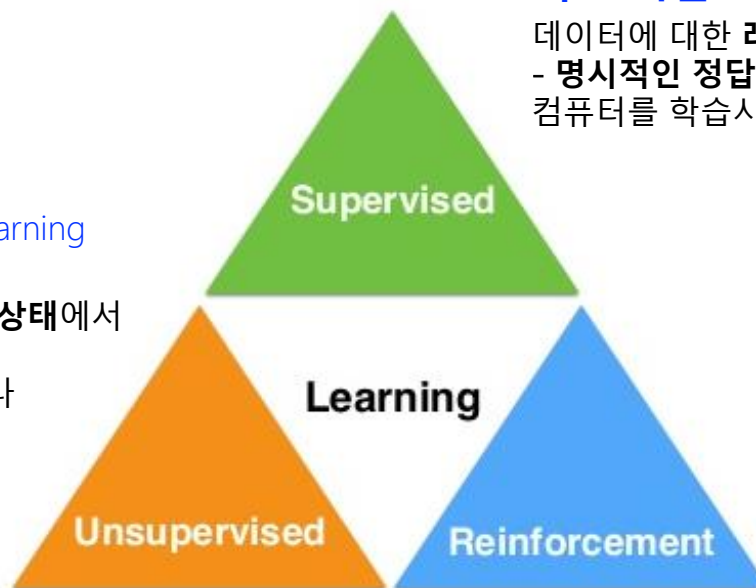
2. 머신 러닝의 분류

머신 러닝의 Machine Learning 학습 방법

머신 러닝의 학습 방법에는 지도 학습, 비지도 학습, 강화 학습이 있다.

비지도 학습 Unsupervised Learning

데이터에 대한 레이블(Label)
- 명시적인 정답이 주어지지 않은 상태에서
컴퓨터를 학습시키는 방법
데이터의 숨겨진 특징(Feature)이나
구조를 발견



지도 학습 Supervised Learning

데이터에 대한 레이블(Label)
- 명시적인 정답이 주어진 상태에서
컴퓨터를 학습시키는 방법

강화 학습 Reinforcement Learning

행동심리학에서 영감을 받았으며, 어떤 환경 안에서 정의된
에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중
보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법

2. 머신 러닝의 분류

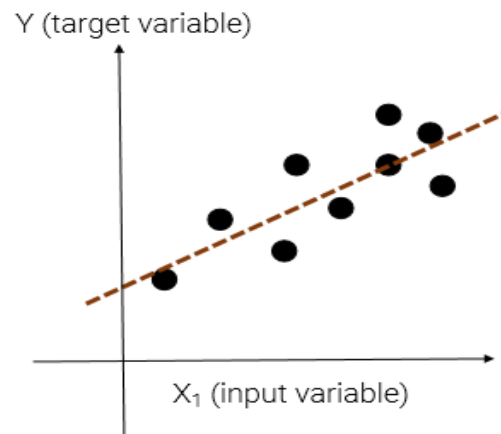
머신 러닝의 Machine Learning 학습 방법

지도 학습(Supervised Learning)

회귀 Regression

타겟 변수 Y가 연속형(Continuous),
범주형(Real Number) 일 때

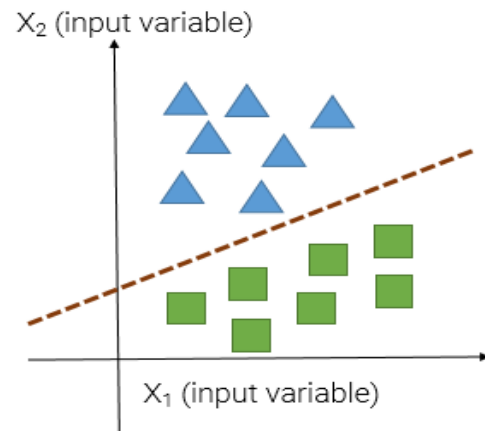
예) - 내일 KOSPI 증가 예측
- 다음 달 매출액 예측



분류 Classification

타겟 변수 Y가 이산형(Discrete),
범주형(Categoria) 일 때

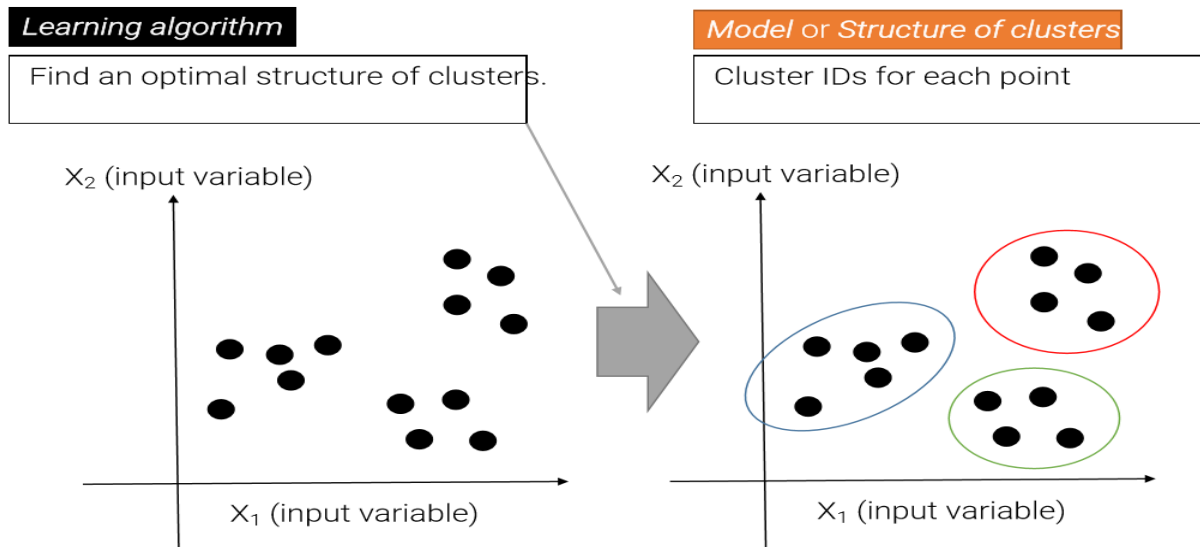
예) - 제품의 불량/ 정상 예측
- 메일의 스팸 예측
- 얼굴 인식



2. 머신 러닝의 분류

머신 러닝의 Machine Learning 학습 방법

비지도 학습(Unsupervised Learning)



군집화 Clustering

유사한 포인트들끼리 모아 군집 구조를 만드는 방법

분포 추정 Density Estimation

관측된 샘플의 확률 분포를 추정하는 방법

연관 규칙 분석 Association Rule Mining

아이템 간의 연관 규칙을 확률 기반으로 평가

잠재 요인 추출 Extracting Latent Factors

데이터 내 잠재되어 있는 새로운 변수/요인 추출

2. 머신 러닝의 분류

머신 러닝의

Machine Learning

학습 방법

머신 러닝 모델의 분류

구분	지도학습	비지도학습	강화학습
분류	KneighborsClassifier Logistic Regression Linear SVC Naïve bayes Decision Tree RandomForest GradientBoosting SVM Feed-Forward Network CNN RNN LSTM	MinMaxScaler StanadScaler PCA NMF t-SNE K-means Agglomerative Clustering DBSCAN Autoencoders	
회귀	KNeighborsRegressor Linear regression Ridge regression Rasso regression Elastic-Net regression		
기타			Q-Learning Deep-Q-Learning

2. 머신 러닝의 분류

2-1. 지도 학습

선형 회귀 모델(Linear Regression Model)

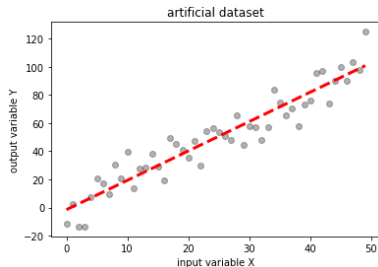
수치 예측

Regression Model

연속형 타겟 변수(continuous target variable) 와
여러 입력 변수들(input variable)의 관계를 만드는 모델

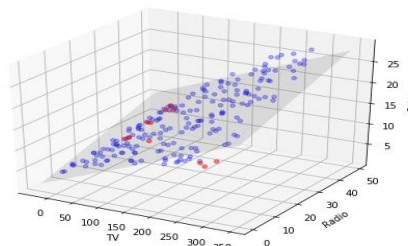
Simple Linear Regression

$$\hat{y} = w_0 + w_1 x_1$$



Multiple Linear Regression

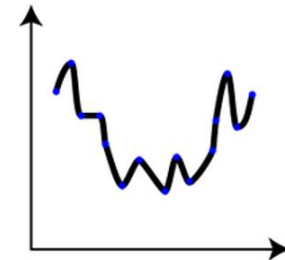
$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$



*3차원 이상일 때는 hiperplane

Polynomial Linear Regression

$$\hat{y} = w_0 + w_1 x_1^6 + w_2 x_1^5 + w_3 x_1^4 + \dots$$



$$Y = \underbrace{w_0}_{\text{Intercept (절편)}} + \underbrace{w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n}_{\text{Coefficients (계수, 가중치, 웨이트)}} + \underbrace{\epsilon}_{\text{Error (오차, 잔차, 손실)}}$$

2. 머신 러닝의 분류

2-1. 지도 학습

수치 예측

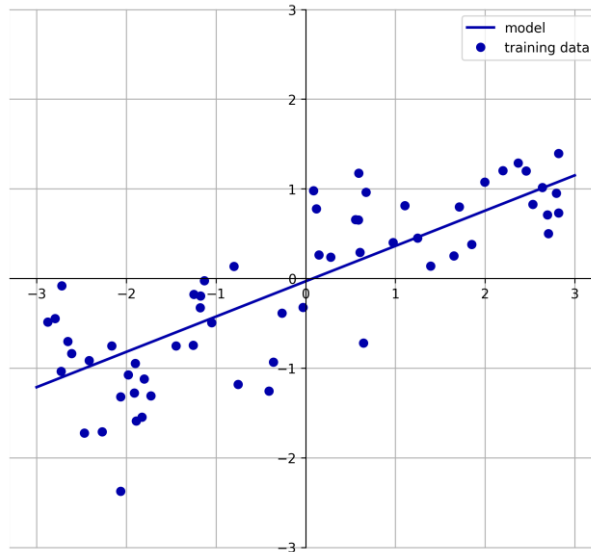
Regression Model

선형 회귀 모델(Linear Regression Model) - 학습 방법

손실 함수의 값을 최소화 하는 계수 찾기

손실 함수(Loss Function) = 기존 값과 예측 값의 차이

$$RSS(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$$



* Simple Linear Regression Case

2. 머신 러닝의 분류

2-1. 지도 학습

범주 예측

Classification Model

로지스틱 회귀(Logistic Regression)

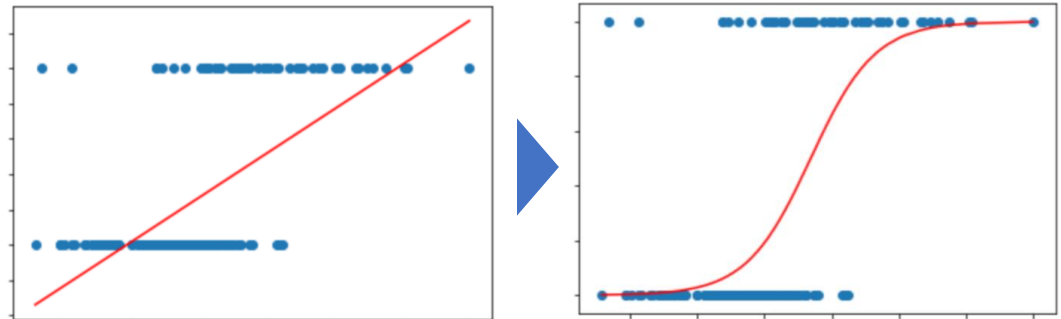
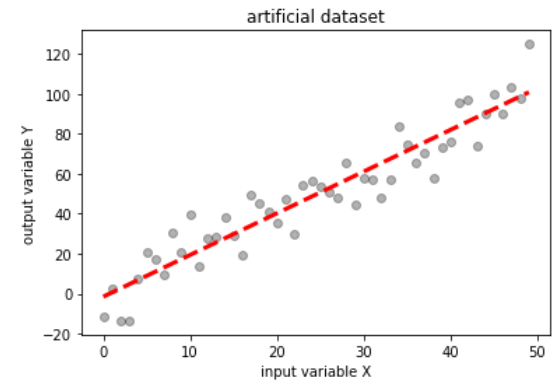
회귀선을 이용해 분류에 활용 → Sigmoid 함수 이용

Linear Regression

선형 함수의 회귀 최적선을 찾는 것

Logistic Regression

Logistic 함수의 최적선을 찾고 반환값을 확률로 간주하는 것



2. 머신 러닝의 분류

2-1. 지도 학습

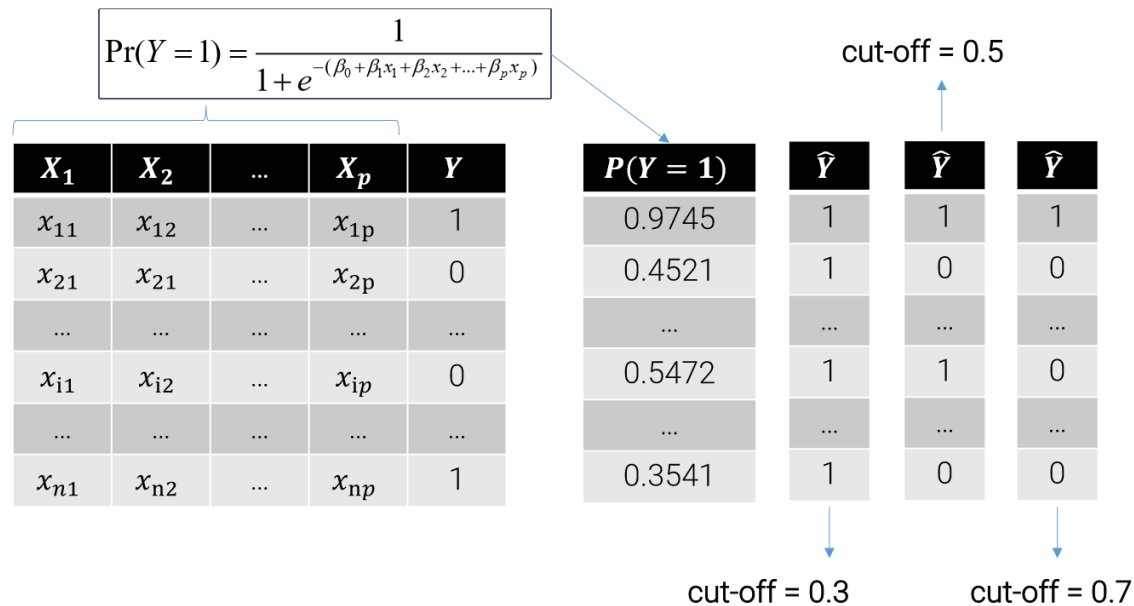
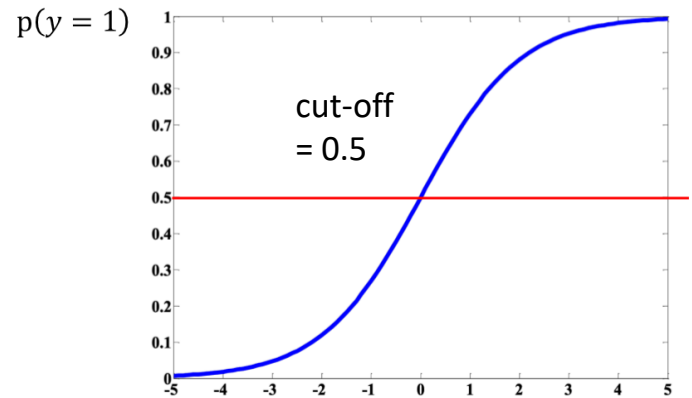
범주 예측

Classification Model

로지스틱 회귀(Logistic Regression) – 예측(Prediction)

cut-off value 이용

- 일반적인 선택 : cut-off = 0.5
- 예측 성능을 높이기 위하여 검증 데이터 (Test set)를 고려하여
최적의 cut-off value를 찾는 방법도 있음



2. 머신 러닝의 분류

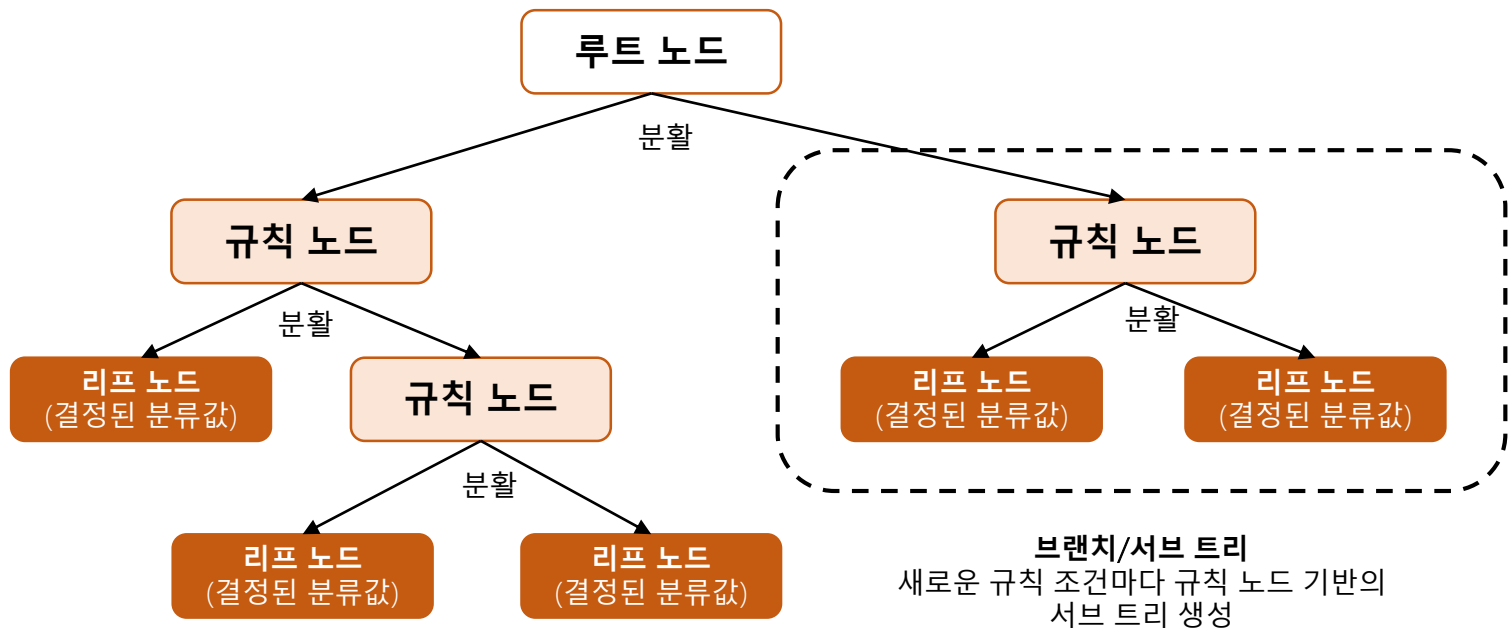
2-1. 지도 학습

의사결정나무 모델(Decision Tree Model)

범주 예측

Classification Model

결정 트리(Decision Tree)는 데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리(Tree) 기반의 분류 규칙을 만드는 것 → 따라서 데이터의 어떤 기준을 바탕으로 규칙을 만들어야 가장 효율적인 분류가 될 것인가가 알고리즘의 성능을 크게 좌우한다.



2. 머신 러닝의 분류

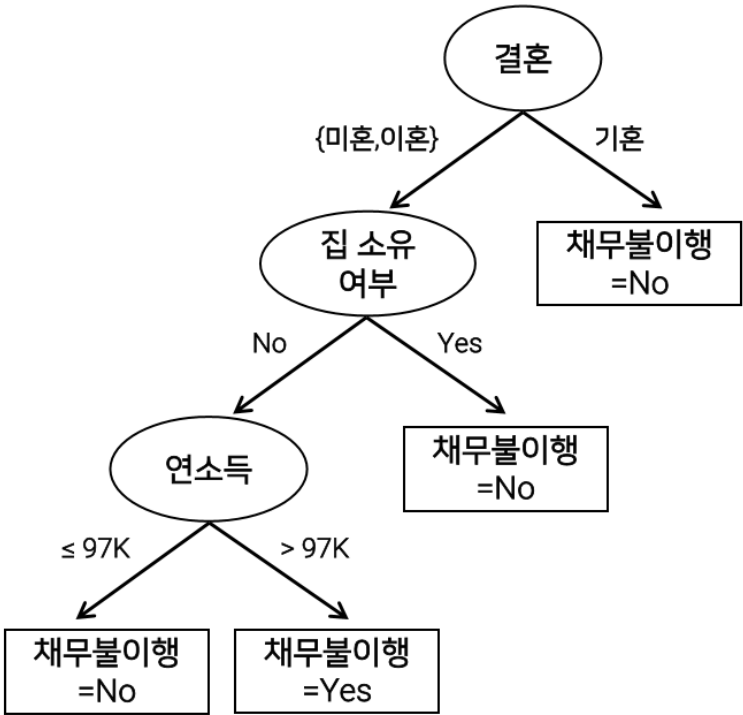
2-1. 지도 학습

의사결정나무 모델(Decision Tree Model) - CART(Classification and Regression Tree)

범주 예측

Classification Model

다음과 같이 새로운 데이터가 주어지면, 미리 학습 데이터를 이용하여, 구축한 의사결정나무 모델을 이용하여 새로운 데이터의 채무불이행 여부를 예측



ID	집 소유	결혼	연소득(K)	채무 불이행	예측 클래스
11	No	미혼	55	?	No
12	Yes	기혼	80	?	No
13	Yes	미혼	110	?	No
14	No	기혼	95	?	No
15	No	이혼	300	?	Yes

2. 머신 러닝의 분류

2-1. 지도 학습

범주 예측

Classification Model

의사결정나무(Decision Tree Model) – 앙상블 학습(Ensemble Learning)

여러 개의 분류기(Classifier)를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 수행

앙상블 학습 유형:

구분	지도학습	비고
보팅 Voting	서로 다른 알고리즘이 같은 데이터 세트에 대해 학습하고 예측한 결과를 보팅	서로 다른 알고리즘 기반
스태킹 Stacking	스태킹은 여러가지 다른 모델의 예측 결과값을 다시 학습데이터로 만들어 다른 모델로 재학습시켜 결과를 예측하는 방법	
배깅 Bagging	단일 결정 트리로 데이터 샘플링을 서로 다르게 가져가면서 학습을 수행해 보팅	결정 트리 알고리즘 기반
부스팅 Boosting	여러 개의 분류기가 순차적으로 학습하면서 앞에서 학습한 분류기가 틀린 데이터에 대해서는 가중치를 부여하면서 학습과 예측을 진행	

2. 머신 러닝의 분류

2-1. 지도 학습

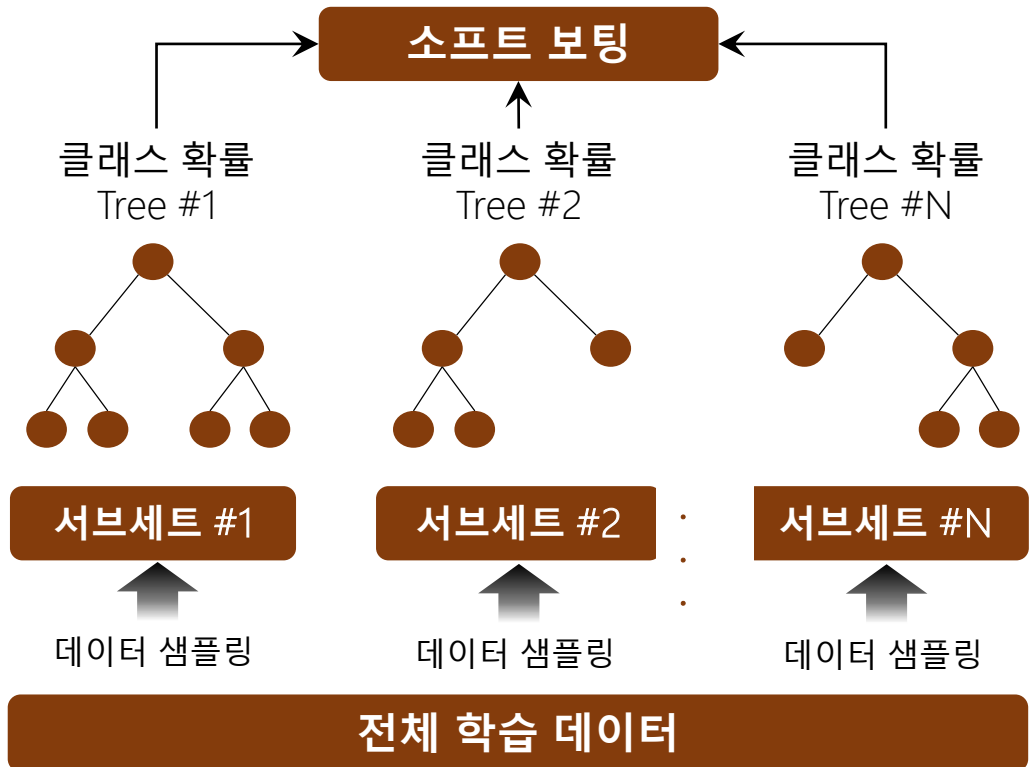
범주 예측

Classification Model

랜덤 포레스트(Random Forest)

여러 개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행한 뒤 최종적으로 모든 보팅을 통해 예측 결정

최종 클래스 값 결정



2. 머신 러닝의 분류

2-1. 지도 학습

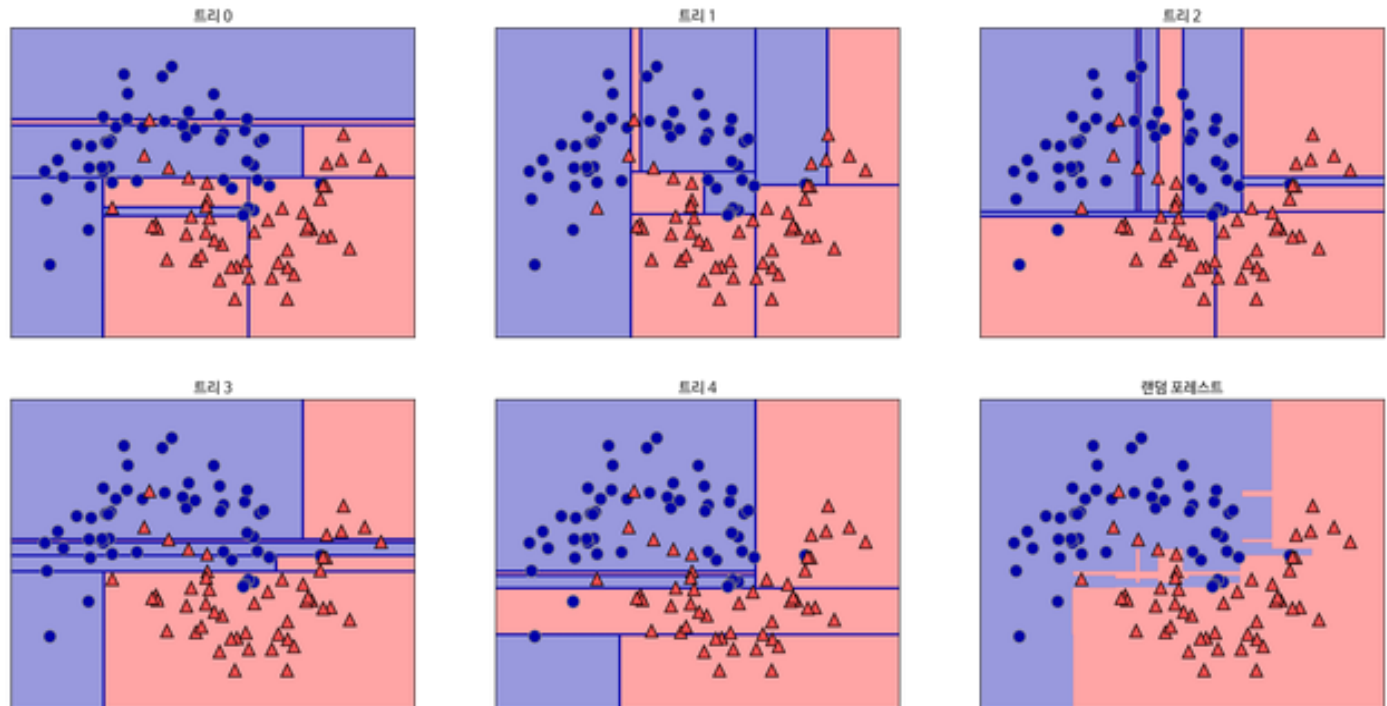
범주 예측

Classification Model

랜덤 포레스트(Random Forest) - Hyperparameter

- **매개변수**($n_{\text{estimators}}$) : 결정 트리의 개수 (디폴트 10개)
- **max_feature** : 결정 트리에 사용된 max_features 파라미터와 동일
- **max_depth**나 **min_sample_leaf**와 같이

결정 트리에서 과적합을 개선하기 위해 사용되는 파라미터가 랜덤 포레스트에도 똑같이 적용



2. 머신 러닝의 분류

2-1. 지도 학습

범주 예측

Classification Model

나이브 베이즈 분류(Naïve Bayes Classification)

나이브 베이즈 분류기(Naïve Bayes Classifier)란, Bayes 정리에 기반을 두는 분류기로 가장 확률이 높은 곳으로 단순 분류

Naïve Bayes Classification 개념

- n 개의 특성을 가지는 데이터 벡터 : $X = (x_1, x_2, \dots, x_n)$
- K 개의 가능한 확률적 결과들 (클래스)의 확률 :

$$P(C_k | x_1, x_2, \dots, x_n) = P(C_k | X) = \frac{P(C_k) \cdot P(X | C_k)}{P(X)}$$

- 각 x_i 값들을 독립으로 가정하면 $= P(C_k) \prod_{i=1}^n P(x_i | C_k)$
- 클래스 예측 값 \leftarrow 최대 확률을 가지는 클래스

$$\hat{y} = \arg \max_{k \in \{1, \dots, K\}} P(C_k) \prod_{i=1}^n P(x_i | C_k)$$

2. 머신 러닝의 분류

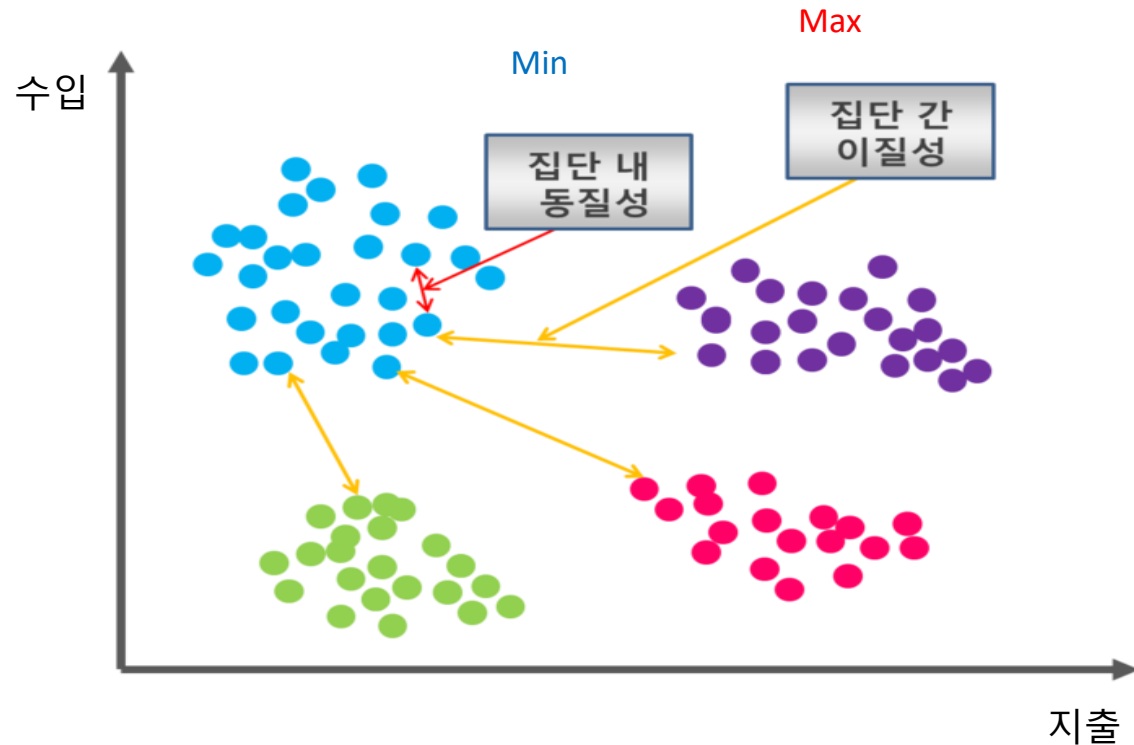
2-2. 비지도 학습

군집 분석

Clustering

군집 분석(Clustering)의 원리

거리가 가까운 데이터끼리 묶어 줌 (거리 distance 감소 = 유사도 similarity 증가)
서로 다른 배타적인 집단으로 나누는 것



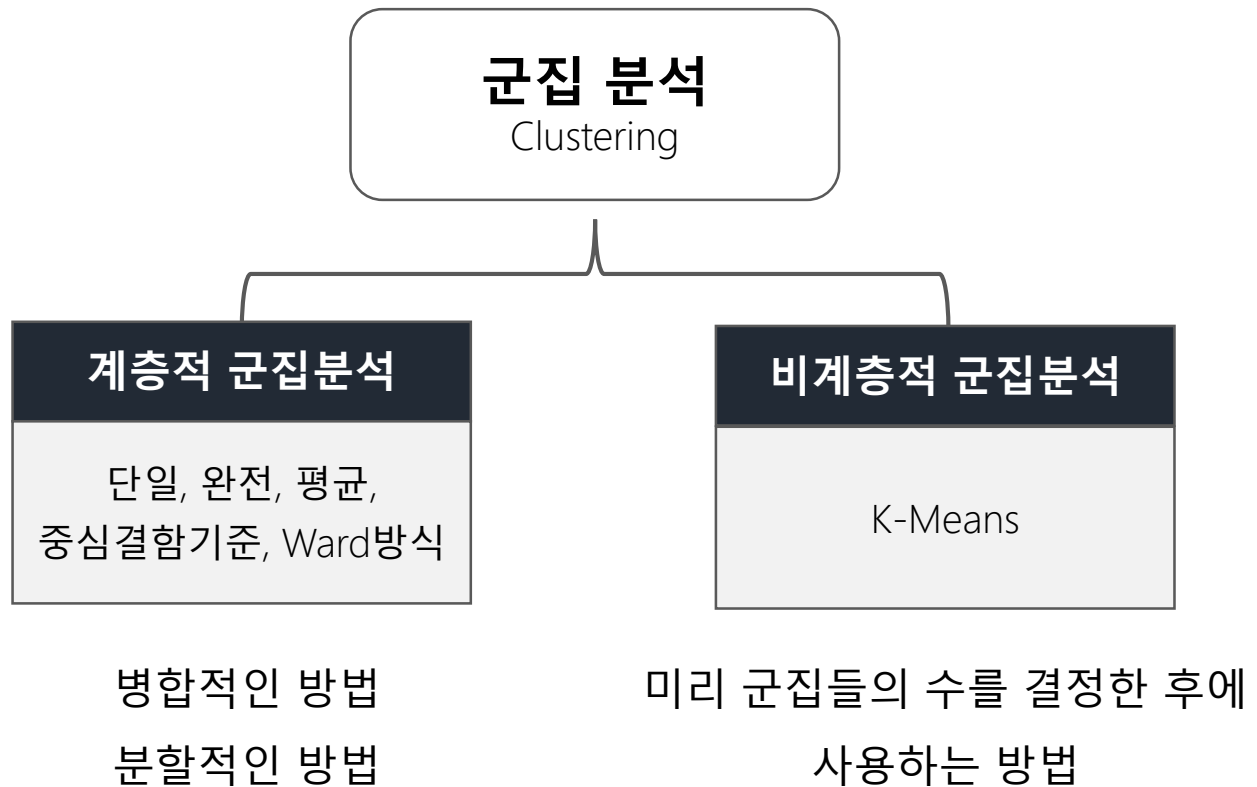
2. 머신 러닝의 분류

2-2. 비지도 학습

군집 분석

Clustering

군집 분석(Clustering)에는 계층적(Hierarchical Clustering) 방법과 비계층적(Non-Hierarchical Clustering) 방법이 있다.



2. 머신 러닝의 분류

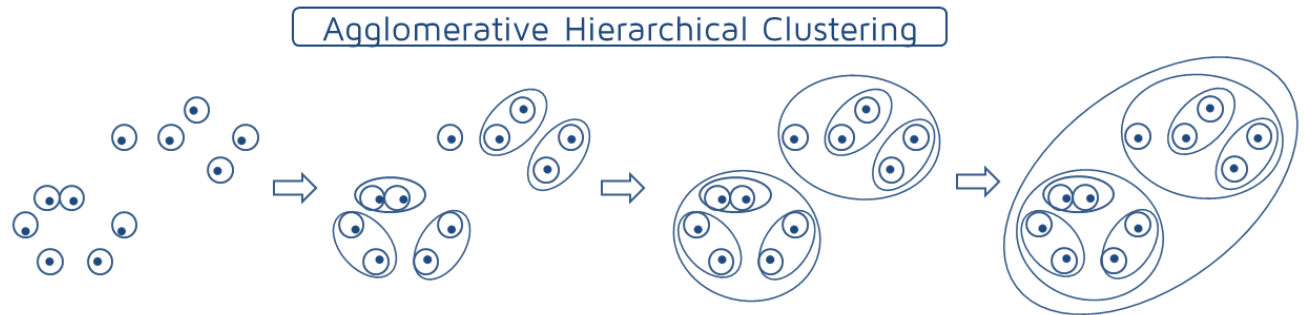
2-2. 비지도 학습

군집 분석

Clustering

계층적 군집 분석 - 병합 군집화(Agglomerative Clustering)

- 1) 정확히 하나의 레코드로 구성된 군집들로 시작
- 2) 종료 조건을 만족할 때까지 가장 가까운 두 군집들을 점진적으로 병합해 나감



(이미지: tds, [Hierarchical clustering Clearly Explained](#))

병합 군집화 알고리즘 :

- N개의 군집으로 시작
- 가장 근접한 두 개의 레코드들은 하나의 군집으로 병합(merge)
- 매 단계에서, 가장 거리가 짧은 두 개의 군집들이 병합됨.
(단일 레코드들이 기존의 군집에 추가되거나, 기존의 군집 두 개가 묶이는 것 의미)

2. 머신 러닝의 분류

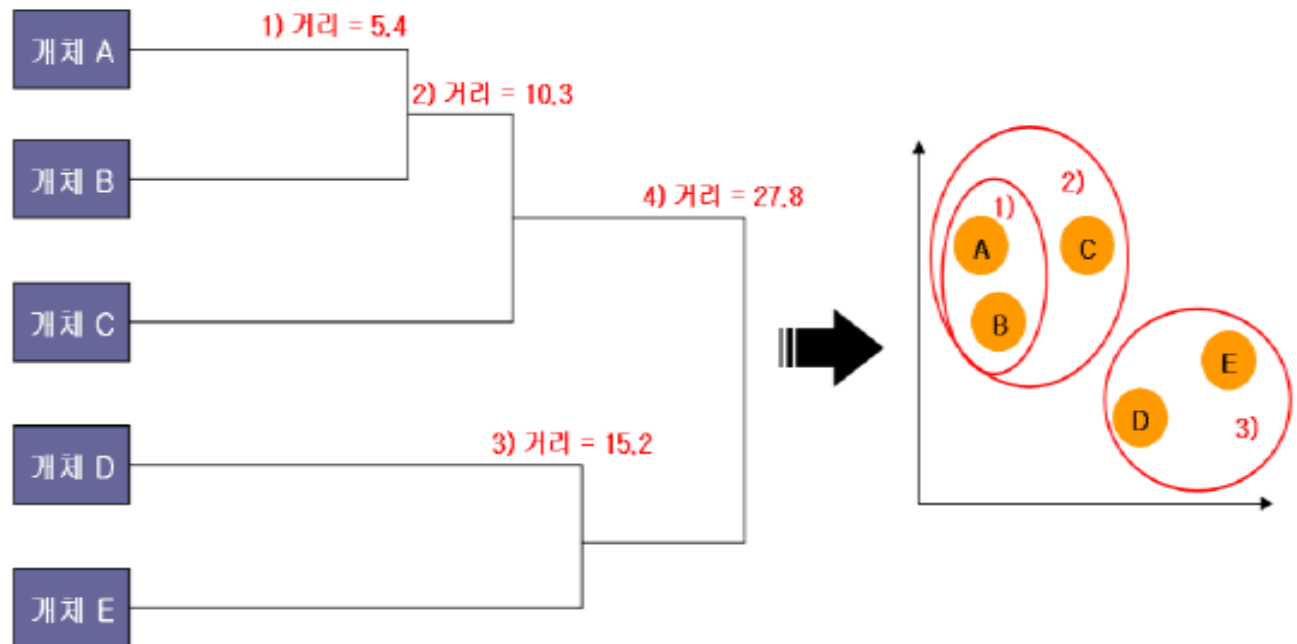
2-2. 비지도 학습

군집 분석

Clustering

계층적 군집 분석 – 덴드로그램(Dendrogram)

군집화 과정을 간략하게 나타내는 나무 형태의 도표



2. 머신 러닝의 분류

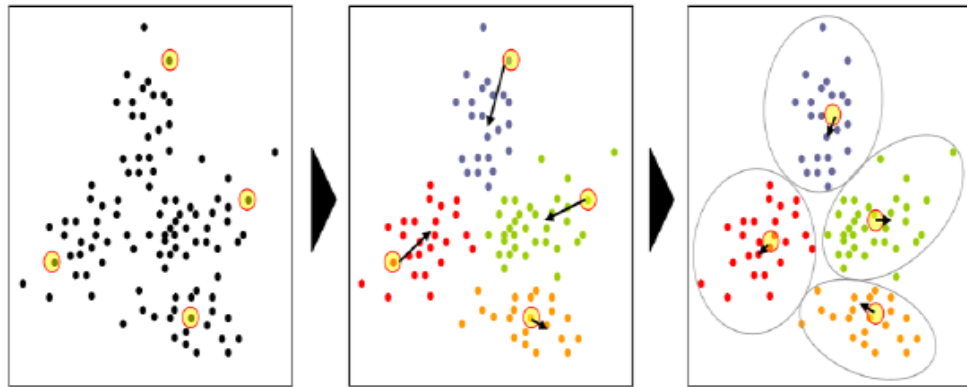
2-2. 비지도 학습

군집 분석

Clustering

비계층적 군집 분석 – K 평균 군집화(K-means Clustering)

군집의 중심이 되는 seed(씨드) 집합을 선택하여 그 seed 점과 거리가 가까운 개체들을 그룹화 하는 방법



K 평균 군집화 알고리즘 :

- K 개의 관측 값을 선택하여 중심점(centroid)으로 정함
- 각 관측 값들을 '가장 가까운' 중심에 해당하는 군집에 할당
- 새로운 군집에 할당된 관측 값들로 새로운 중심을 계산
- 2)과 3)의 과정을 군집의 중심에 변화가 없을 때까지 반복

※ K 평균 군집분석의 단점: 사전에 군집 수에 대한 예측이 필요하고 처음 선정한 seed 점들에 따라서 군집의 분류가 달라질 가능성 有

2. 머신 러닝의 분류

2-2. 비지도 학습


차원 축소

Dimensionality
Reduction

**차원 축소(Dimensionality Reduction)는, 데이터의 의미를 잘 표현하는
특징(Feature)을 추려내는 것**


특징 추출(Feature Extraction)

- 원본 특징을 기반으로 새로운 특징 벡터를 생성함
- 모든 feature를 온전히 잘 설명하는 원래 벡터보다 작은 feature 벡터로 나타냄

$(X_1, X_2, X_3, X_4, X_5)$  (Z_1, Z_2, Z_3)

특징 선택(Feature Selection)

- 전체 입력된 feature 중에서 가장 의미 있는 feature들 만을 선택
- 원본 데이터에서 불필요한 feature(변수)들을 제거

$(X_1, X_2, X_3, X_4, X_5)$  (X_1, X_3, X_5)

2. 머신 러닝의 분류

2-2. 비지도 학습

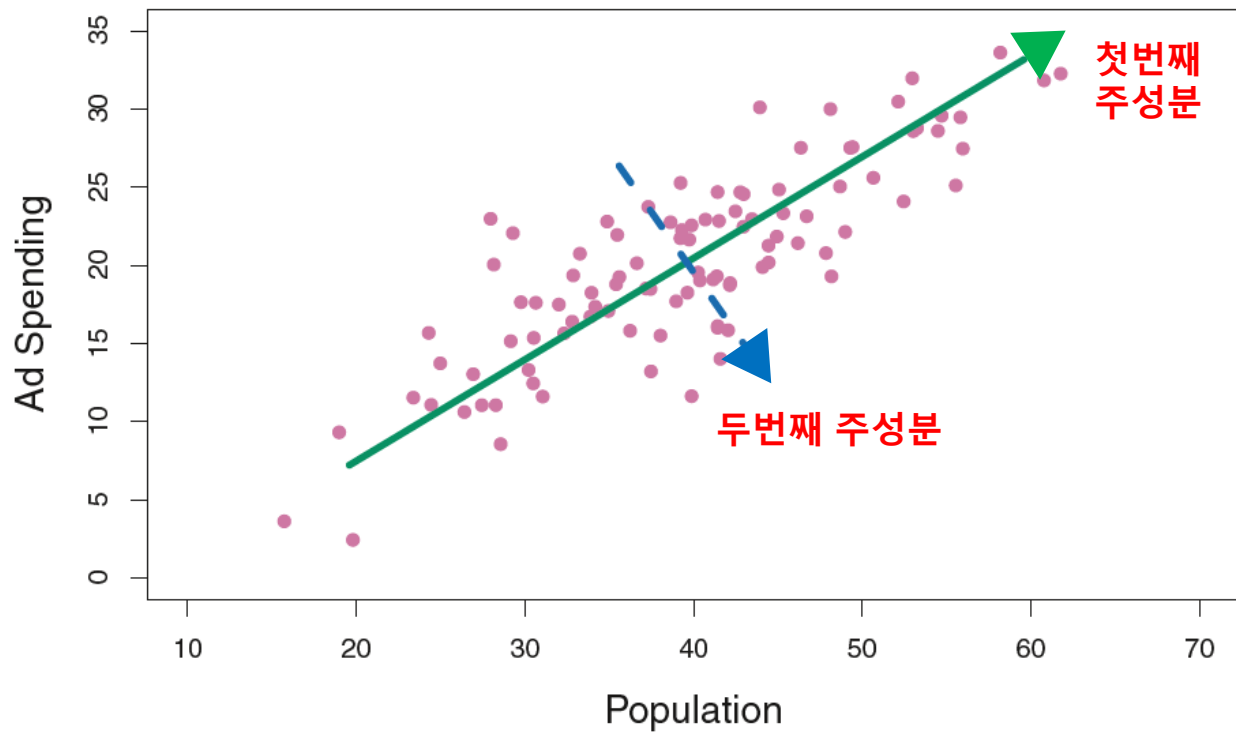
차원 축소

Dimensionality
Reduction

주성성분분석(Principal Component Analysis, PCA)

PCA는 분포의 주성분(principal component)을 분석해 주는 방법이며, 주성분은 그 방향으로 가장 분산이 큰 벡터를 의미한다.

데이터 주성분 구하기:



3. 머신 러닝 모델 평가

수치 예측

Regression Model

모델 평가

회귀 모델(Regression Model)에 대한 성능 평가지표

평가 지표	설명	수식
MAE	Mean Absolute Error이며 실제 값과 예측 값의 차이를 절대값으로 변환해 평균한 것	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $
MSE	Mean Squared Error이며 실제 값과 예측 값의 차이를 제곱해 평균한 것 *MAE값이 같은데 MSE가 클 경우 편차가 더 큼을 나타낸다.	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
RMSE	MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 것이 RMSE(Root Mean Squared Error)다.	$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
R ²	분산 기반으로 예측 성능을 평가합니다. 실제 값의 분산 대비 예측값의 분산비율을 지표로 하며, 1에 가까울수록 예측 정확도가 높다. *R ² = 0.91인 경우, 전체 데이터 변동성의 91%를 선형회귀 모델이 설명	$\frac{\sum (\hat{y}_i - \bar{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$

3. 머신 러닝 모델 평가

범주 예측

Classification Model

모델 평가

분류 모델(Classification Model)에 대한 성능 평가지표

혼동 행렬(Confusion Matrix): 데이터의 실제 클래스와 모델에 의해 예측된 클래스를 비교하는 행렬

		PREDICTIVE VALUES		
		POSITIVE (1)	NEGATIVE (0)	
ACTUAL VALUES	POSITIVE (1)	TP = 3	FN = 1	4
	NEGATIVE (0)	FP = 2	TN = 4	6
		5	5	

PRECISION (green box around TP and FP)

RECALL (red box around TP and FN)

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

$$\text{Recall} = \text{Sensitivity} = TP / (TP + FN) = TP / (\text{Actual Yes})$$

$$\text{Precision} = TP / (TP + FP) = TP / (\text{Predicted Yes})$$

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

3. 머신 러닝 모델 평가

범주 예측

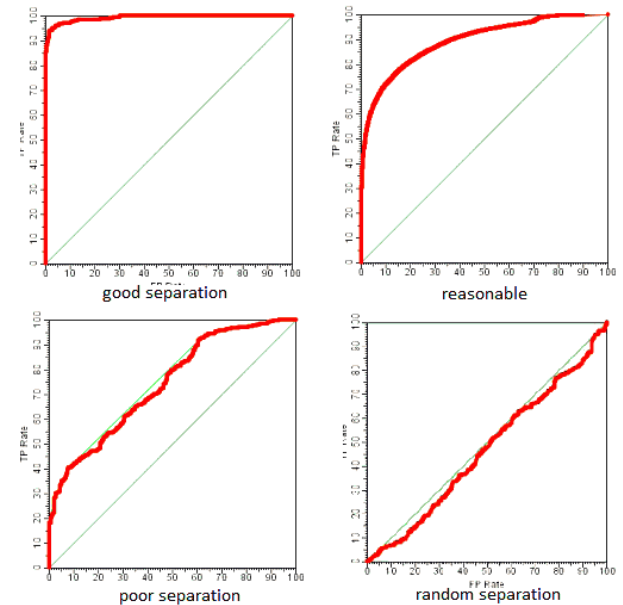
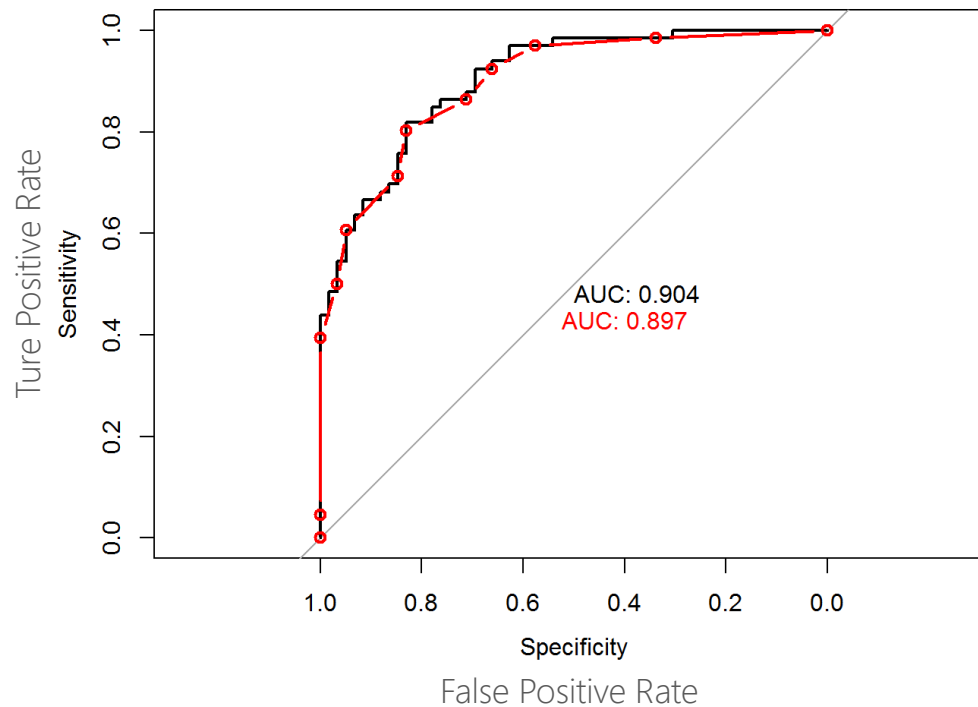
Classification Model

모델 평가

분류 모델(Classification Model)에 대한 성능 평가지표

ROC(Receiver Operating Characteristic) Curve : FPR과 TPR을 x, y축으로 놓은 그래프

AUC(Area Under the Curve) Score : Curve 아래의 면적을 계산한 것, AUC는 0.5~1의 값을 가지며, 값이 1에 가까울수록 성능이 좋은 모델이다.



3. 머신 러닝 모델 평가

범주 예측

Classification Model

모델 평가

분류 모델(Classification Model) - 분류 평가

Model X

		Predicted Class	
		1 (+)	0 (-)
Actual Class	1 (+)	6	4
	0 (-)	50	940

- Accuracy = $946/1000 = 0.946$
- Recall = $6 / (6 + 4) = 0.6$
- Precision = $6 / (6 + 50) = 0.107$
- F1-score = 0.18

Ideal
Model

		Predicted Class	
		1 (+)	0 (-)
Actual Class	1 (+)	10	0
	0 (-)	0	900

- Accuracy = 1
- Recall = 1
- Precision = 1
- F1-score = 1* (Best)

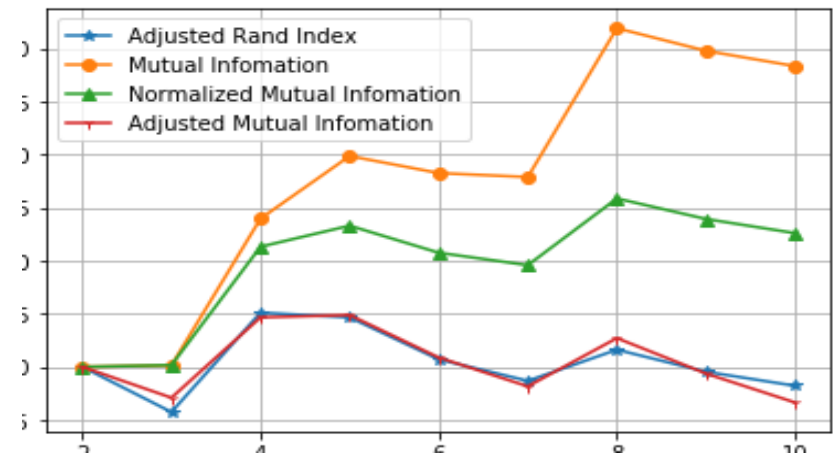
3. 머신 러닝 모델 평가

클러스터링 Clustering Model 모델 평가

군집 모델(Clustering Model) – 클러스터링 평가

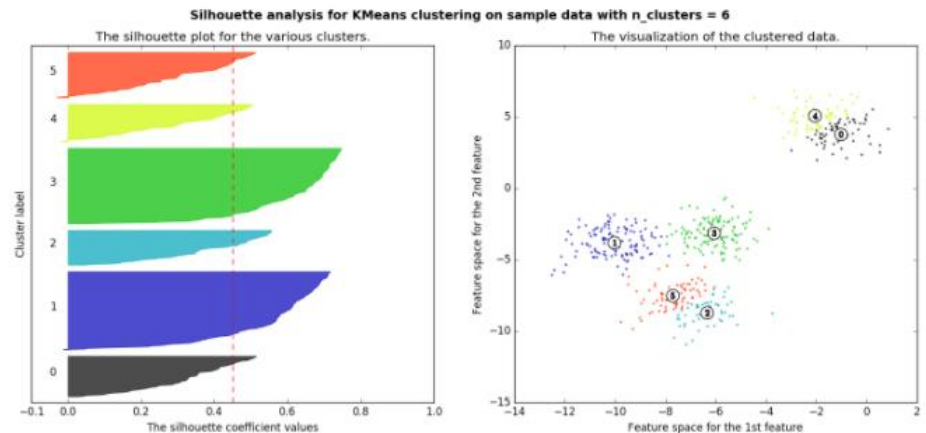
ARI(Adjusted Rand Index):

분류된 군집의 정답 필요

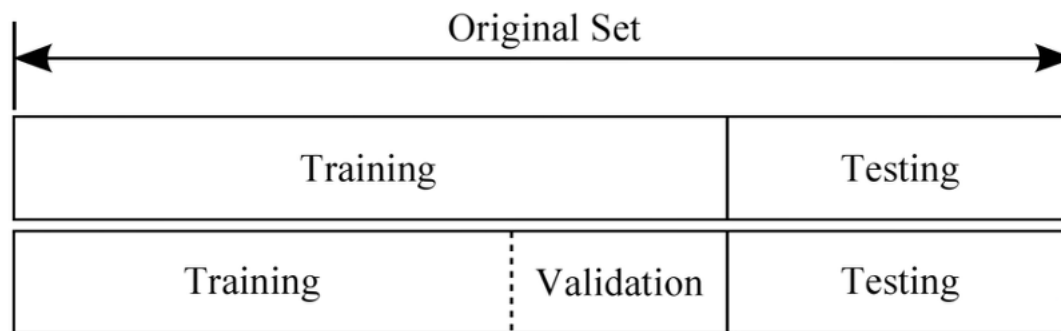


실루엣 계수(Silhouette Coefficient):

군집의 밀집 정도



데이터셋 Dataset **의** **머신 러닝을 위한 데이터는 Training, Validation, Test 셋으로 나눈다.**
종류



Training Set

- 모델 생성 및 학습에 이용

Validation Set

- 모델의 오버피팅(Overfitting) 방지
- 모델의 복잡도 축소
- 모델의 파라미터(Parameter) 탐색

Test Set

- 모델의 예측 성능(Predictive Performance) 평가

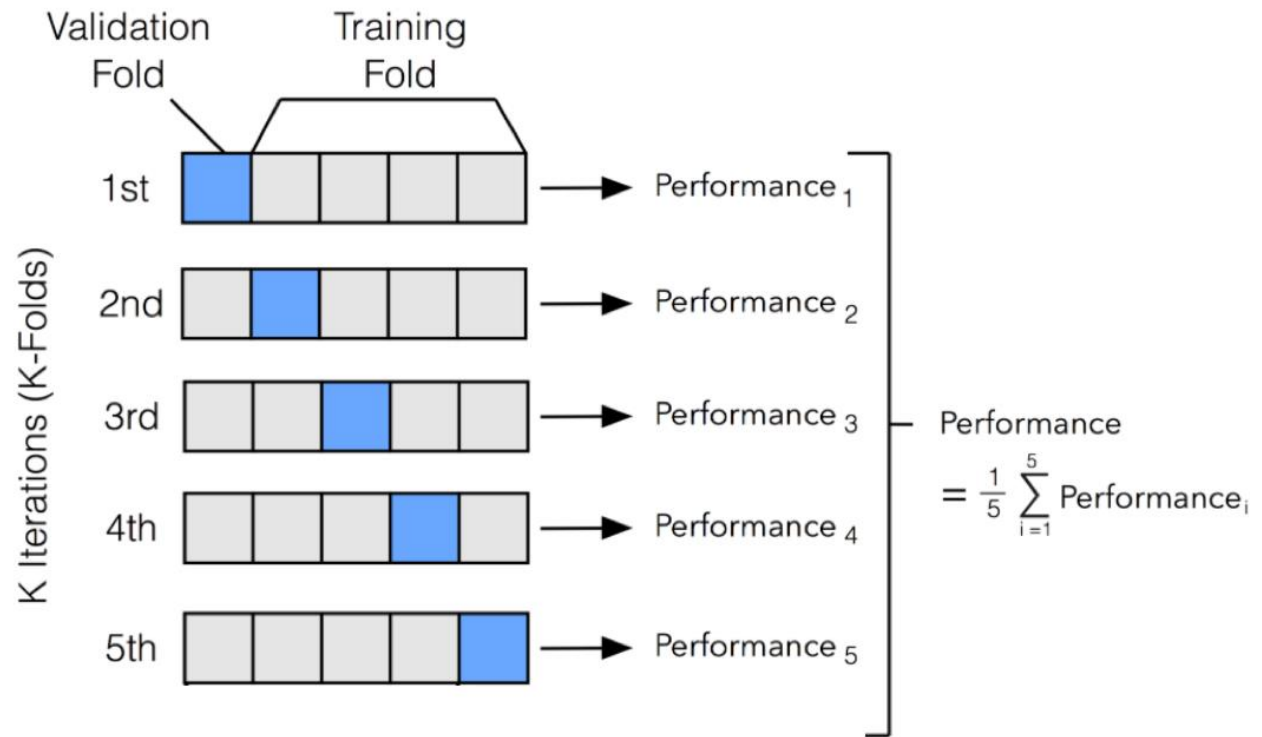
4. 데이터 분할

K-겹 교차 검증

K-fold Cross Validation

K-겹 교차 검증은 모든 데이터가 최소 한 번은 테스트셋으로 쓰이도록 한다.

- 데이터를 K개의 겹치지 않는 folds로 분리
- K개의 folds 중 하나를 Validation Set, 나머지를 Training Set으로 사용
- 하나의 파라미터 셋에 대해 k번 모델을 생성하여 모델 성능 평가



End of Document