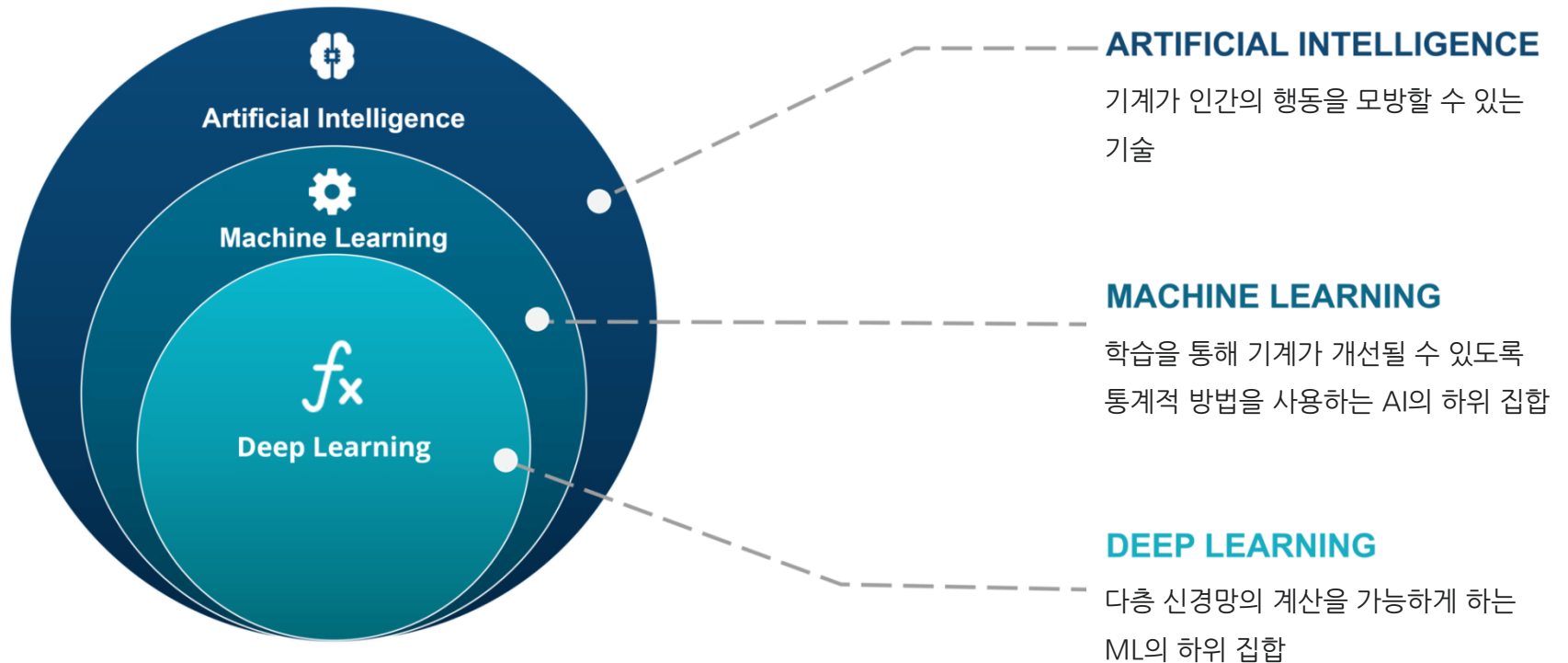


1. Machine Learning

1. Machine Learning

1-1. 머신 러닝 (Machine Learning)의 개념

AI > Machine Learning > Deep Learning



1. Machine Learning

1-2. 통계와 머신 러닝의 차이

통계와 머신러닝 모두 독립변수와 결과의 관계를 설명하는 모델을 찾기 위해 데이터를 학습한다. 그러나 결과 모델의 사용 목적이 다르다. 통계는 모델을 사용해 요인을 추론하고, 머신 러닝은 모델을 사용해 새로운 데이터가 들어왔을 때의 결과를 예측하는데 목적을 둔다.

Statistics

Machine Learning



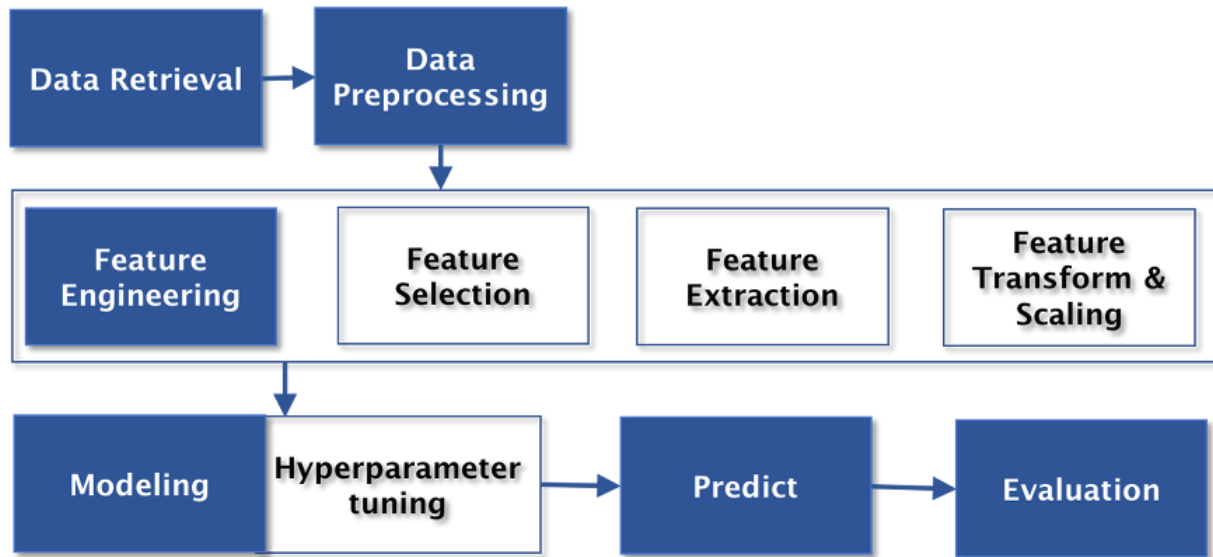
1. Machine Learning

1-3. 머신 러닝의 작업 흐름

머신 러닝의 전체 흐름에서 가장 중요한 단계는 Data Preprocessing과 Feature Engineering이다.

비정형 데이터를 정형 데이터로 변환, 결측치/이상치 처리, 중복된 데이터를 제거, 데이터 차원의 축소 등의 작업을 통해 모델링 단계에서의 계산 시간과 비용을 절약하고 모델의 성능을 향상 시킬 수 있다.

Machine Learning Pipeline



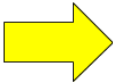
1. Machine Learning

1-4. 전처리 (Preprocessing)

데이터 분석의 안정적인 결과와 성능 향상을 위해서 주어진 데이터를 분석에 적합하게 가공하는 작업이다.
대표적인 작업으로는 필터링, 클리닝, 결측치 처리, 이상치 처리, 데이터 형태 변경 등이 있다.

- Filtering / Cleaning / Missing Value / Outlier
- 범주형 데이터 인코딩 : 레이블 인코딩 (Label Encoding) & 원핫 인코딩 (One-hot Encoding)
- Feature 스케일링 : 표준화 (Normalization) & 정규화 (Standardization)
- Data Shape : Long Data, Wide Data

One-hot Encoding

Color		Red	Yellow	Green
Red				
Red		1	0	0
Yellow		1	0	0
Green		0	1	0
Yellow		0	0	1

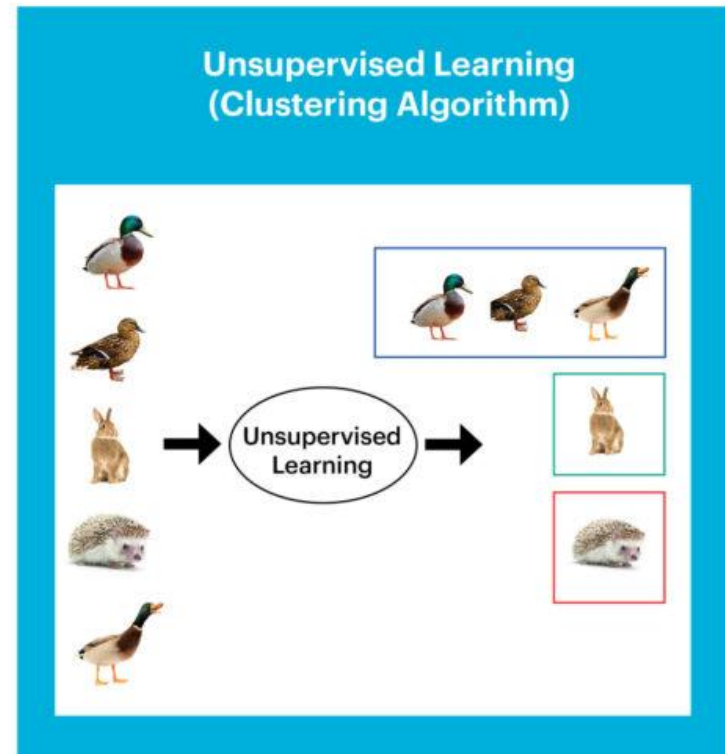
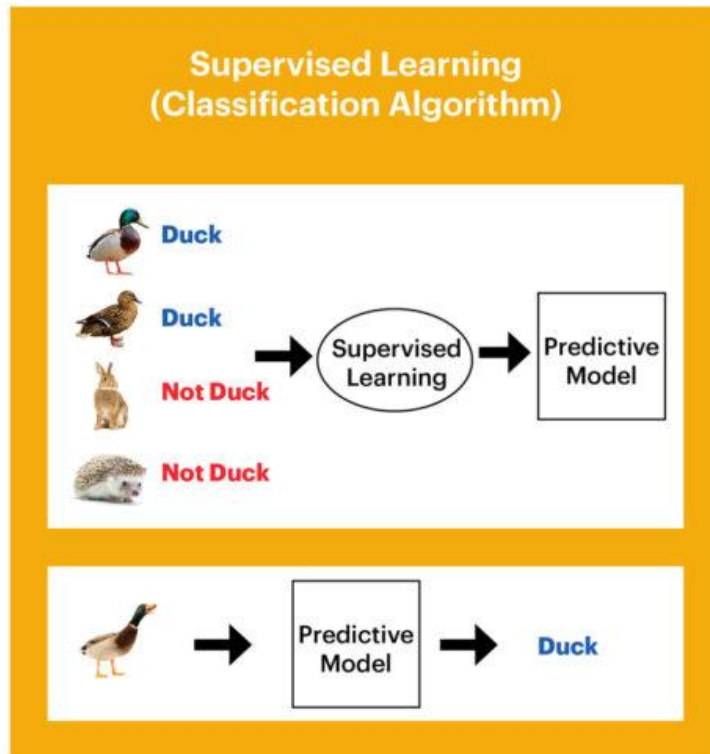
Feature Scaling

Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

1. Machine Learning

1-5. 지도 학습과 비지도 학습 (Supervised Learning & Unsupervised Learning)

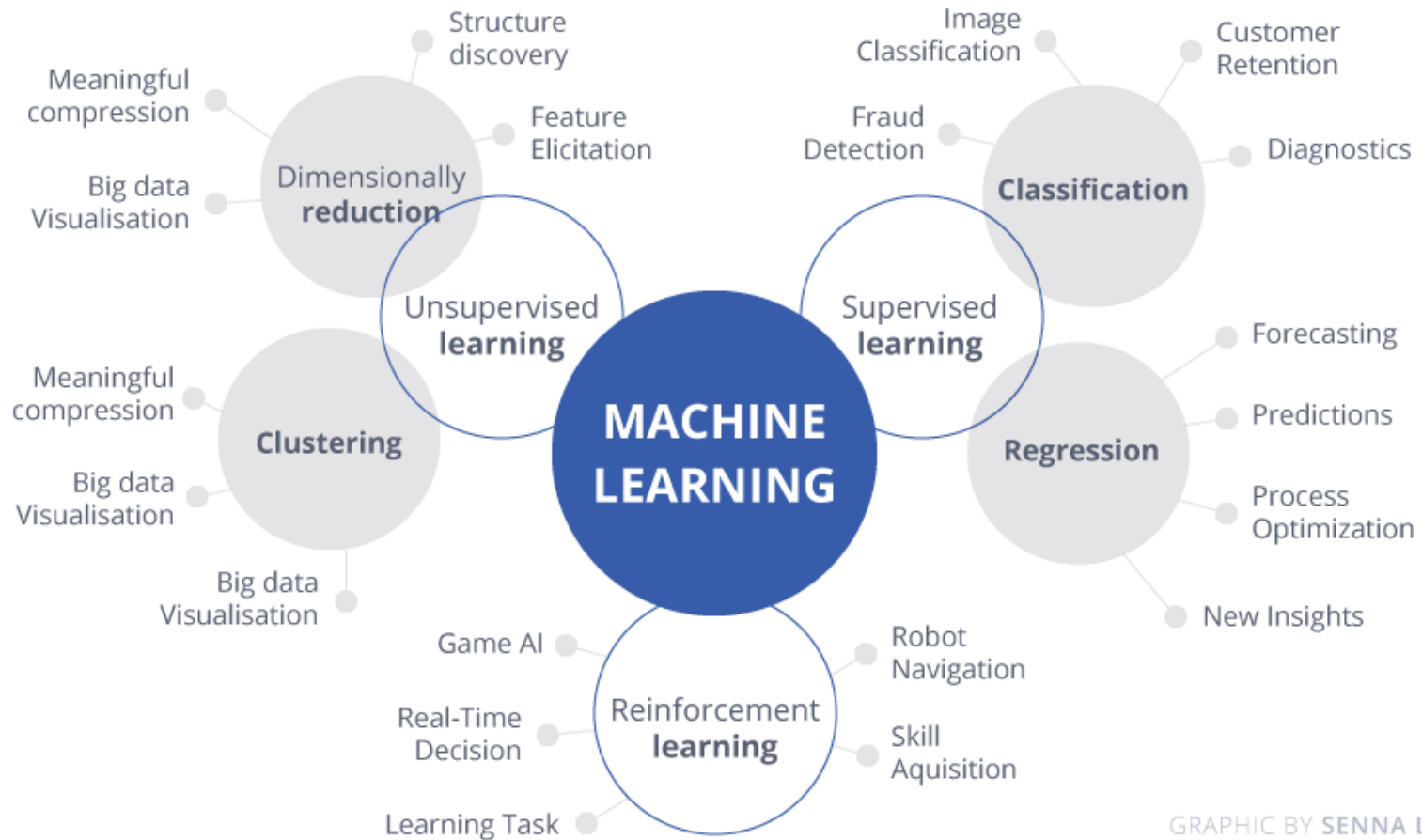
지도 학습과 비지도 학습의 가장 큰 차이점은 학습 데이터에서의 레이블(Label) 유무이다.



Western Digital.

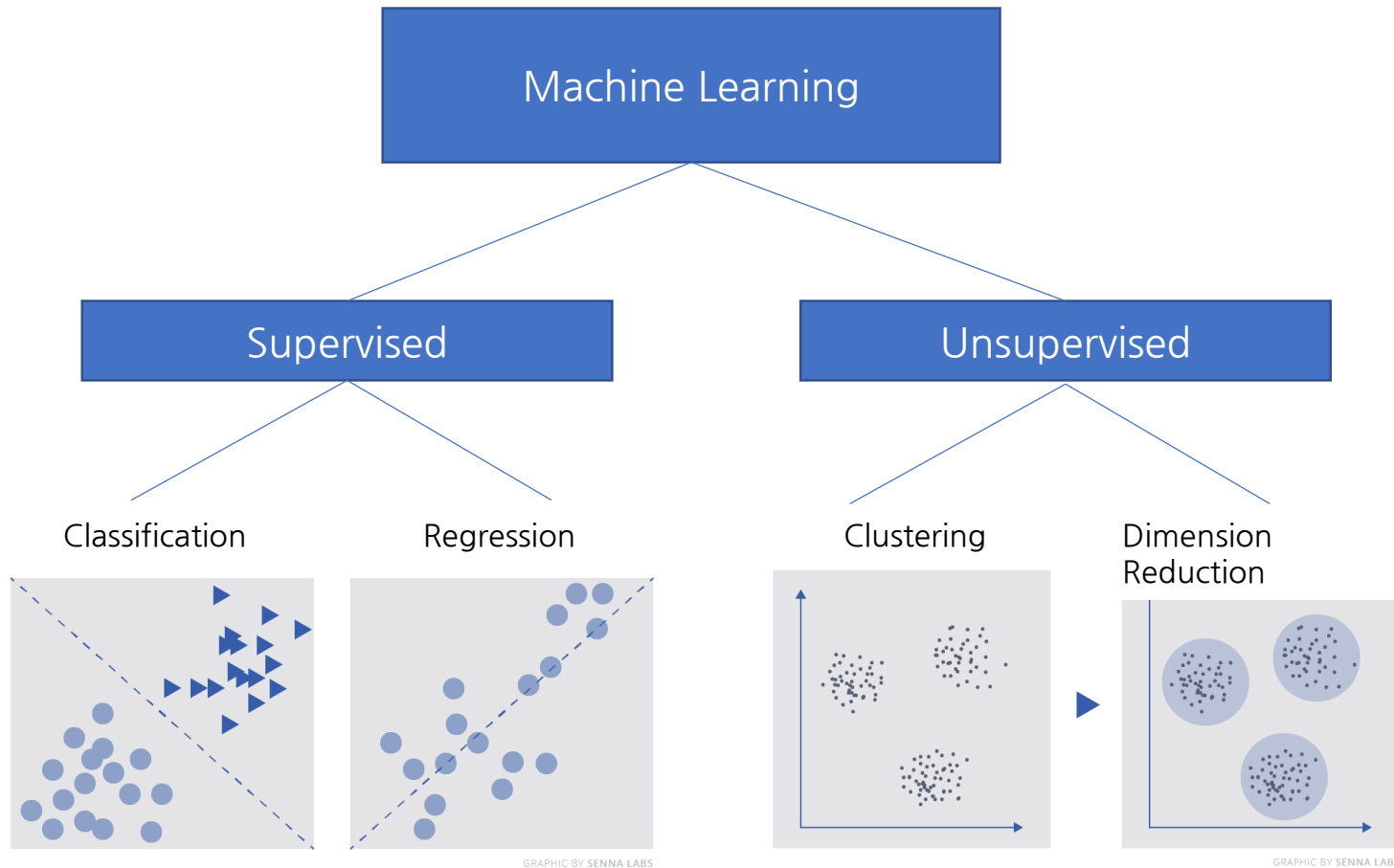
1. Machine Learning

1-5. 지도 학습과 비지도 학습 (Supervised Learning & Unsupervised Learning)



1. Machine Learning

1-5. 지도 학습과 비지도 학습 (Supervised Learning & Unsupervised Learning)



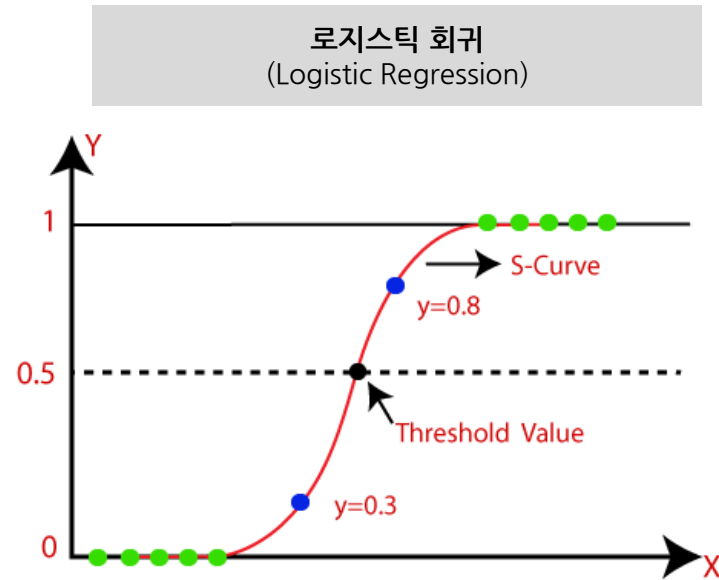
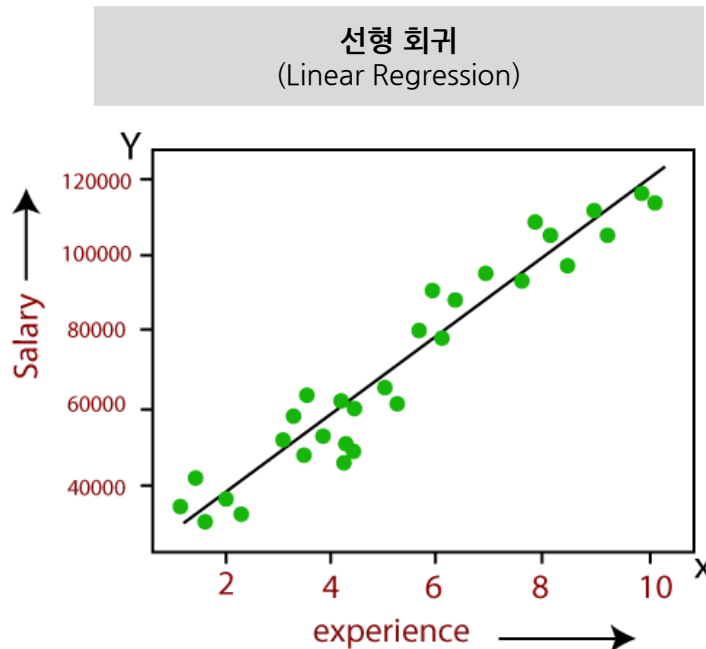
1. Machine Learning

1-6. Regression 계열 알고리즘

회귀 (Regression) 알고리즘은 레이블링 된 데이터를 기반으로 결과 값을 예측하는데 사용되며 지도학습에 해당된다. 회귀 모델은 '예/아니오'등 범주형 값을 예측하는 로지스틱 회귀 (Logistic Regression)과 온도, 주가 등 연속되는 값을 예측하는 선형 회귀 (Linear Regression)으로 구분된다.

* 참고 : Ridge / Lasso / Elastic Net

SVM (Support Vector Machine)



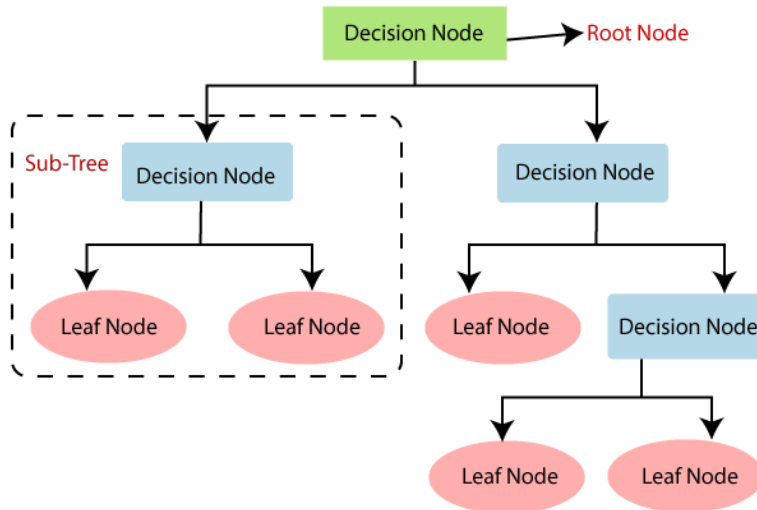
1. Machine Learning

1-7. Tree 계열 알고리즘

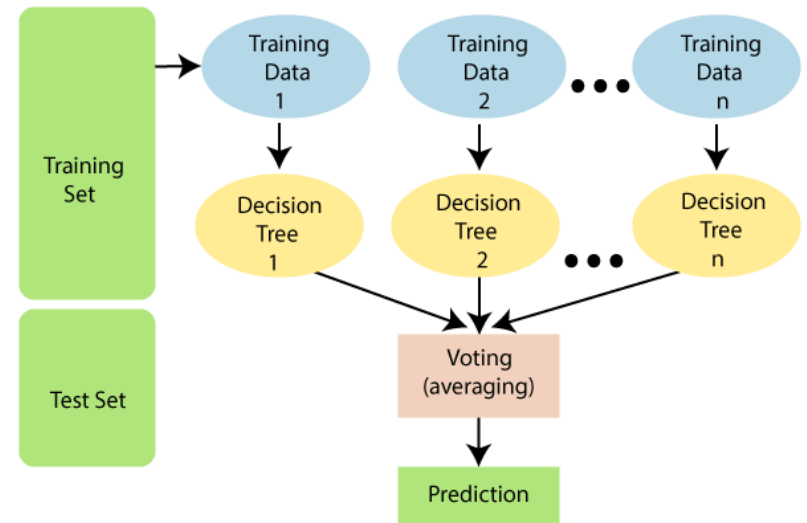
Tree 계열의 알고리즘은 지도학습 기술로 분류와 예측 문제에 모두 사용할 수 있지만 대부분 분류 문제 해결에 선호된다. 랜덤 포레스트 모델은 모델의 성능 향상을 위해 여러 Tree 모델을 결합한 앙상블 학습의 개념이다.

* 참고 : GBM / XGBoost / LightGBM

의사결정나무
(Decision Tree)



랜덤 포레스트
(Random Forest)



1. Machine Learning

1-7. Tree 계열 알고리즘

앙상블 (Ensemble)

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

앙상블 학습 유형:

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

1. Machine Learning

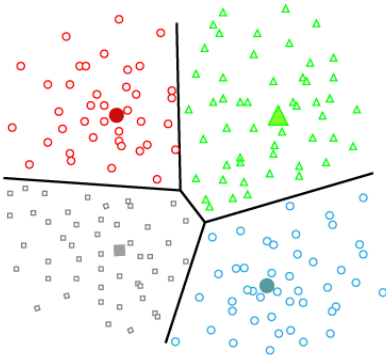
1-8. 군집 분석 (Clustering)

군집 분석은 레이블이 지정되지 않은 데이터 세트를 그룹화하는 비지도학습 기술이다.
데이터 세트에서 유사한 패턴을 찾아 그 패턴에 따라 분할한다.

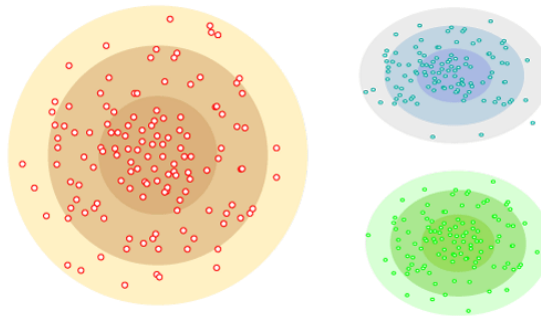
- 군집 분석 : 데이터 세트에 레이블이 '없음'
- 분류 분석 : 데이터 세트에 레이블이 '있음'

*참고 : K-Means / GMM (Gaussian Mixture Model) / DBSCAN

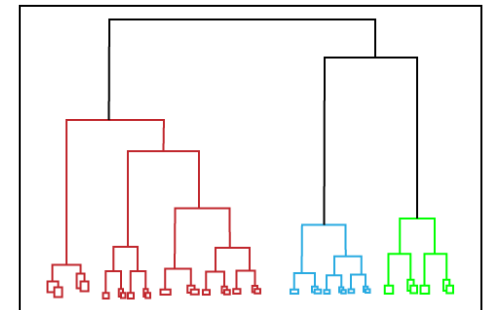
분리형 군집화
(Partitioning Clustering)



분포 기반 군집화
(Distribution-based Clustering)



계층적 군집화
(Hierarchical Clustering)



1. Machine Learning

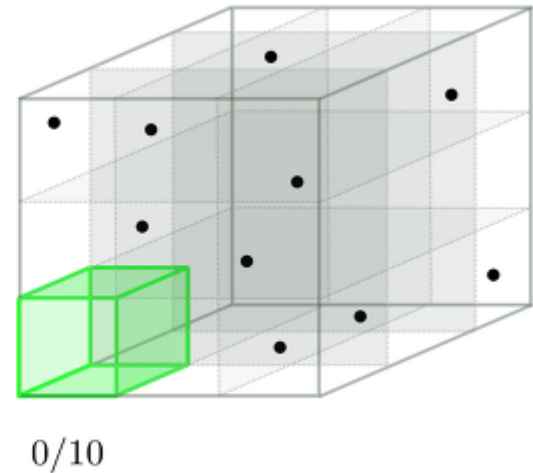
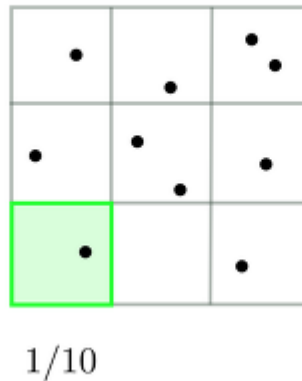
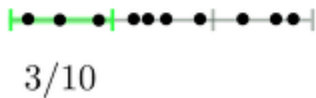
1-9. 차원 축소 (Dimension Reduction)

차원 축소란 주어진 데이터가 가지고 있는 특성을 줄이는 과정이다.

차원의 저주 (Curse of Dimensionality)는 학습 데이터의 개수(Row의 개수)가 특성의 수(Column의 개수) 보다 적어지면 과적합(Overfitting)으로 인해 모델의 성능 저하되는 것을 말한다.

차원이 증가할 수록 데이터의 밀도가 희소(Sparse)해진다.

* 참고 : PCA / LDA / SVD / NMF

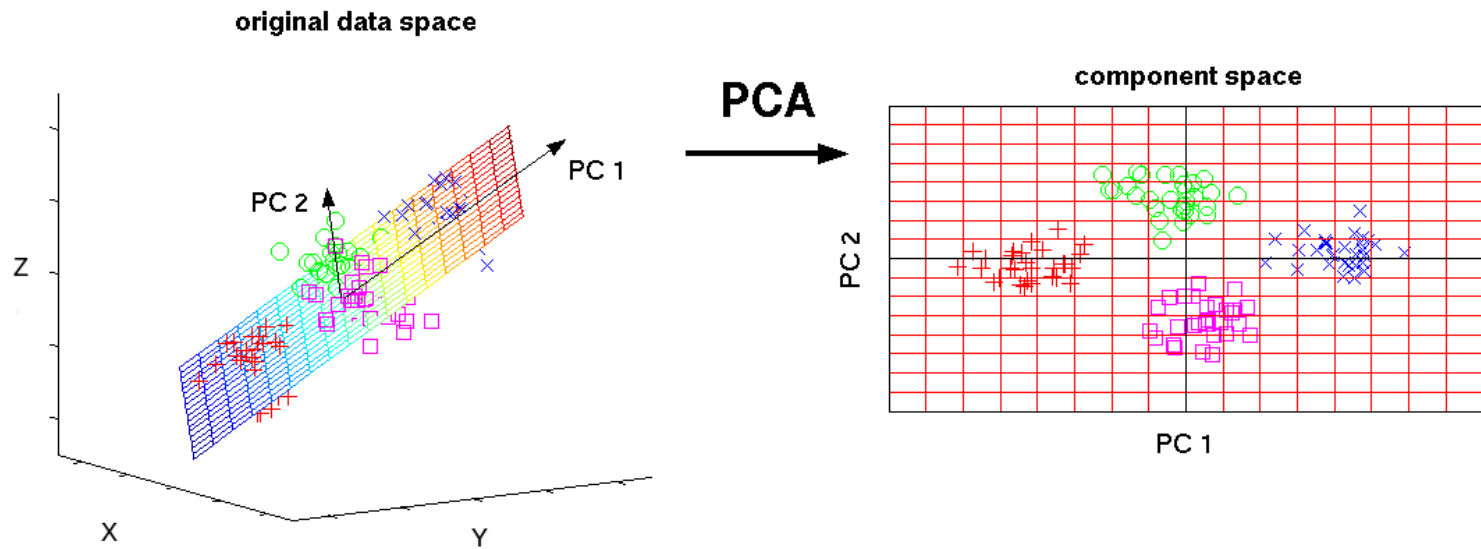


1. Machine Learning

1-9. 차원 축소 (Dimension Reduction)

PCA (Principal Components Analysis)

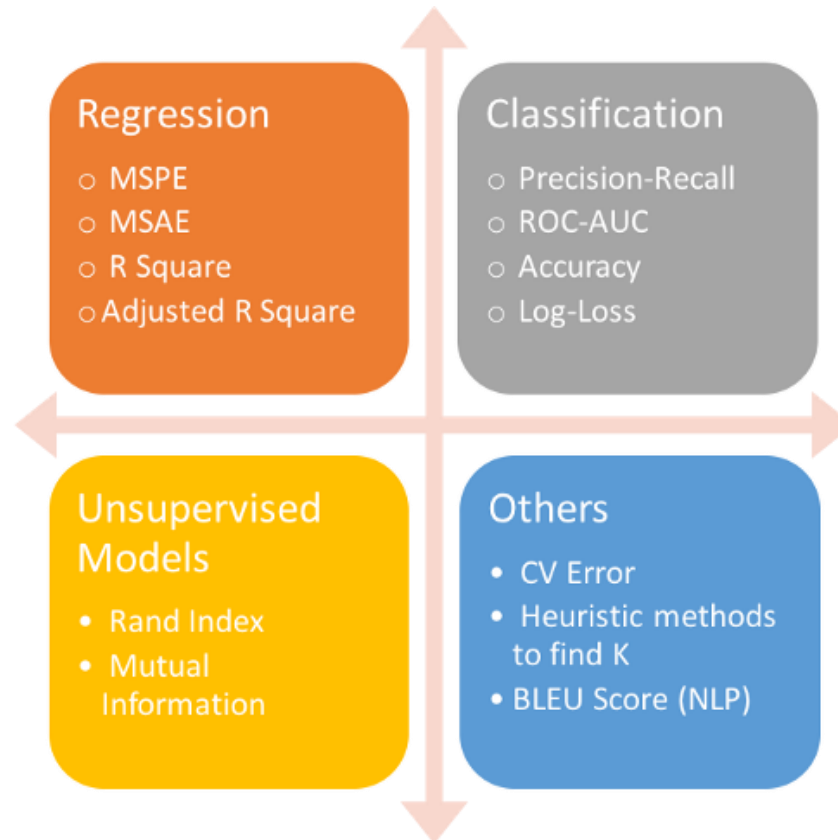
PCA는 특성 추출(Feature Extraction) 기법 중 하나이다. 단순히 특성을 선택 (Feature Selection)하는 것이 아니라 기존 특성의 조합으로 새로운 특성을 생성하는 방식이다.



1. Machine Learning

1-10. 모델 평가 (Model Evaluation)

기계 학습에는 모델의 성능을 평가하기 위한 다양한 지표들이 있다. 대표적인 지표로는 오차 행렬(Confusion Matrix), AUC/ROC 곡선 등이 있다.



1. Machine Learning

1-10. 모델 평가 (Model Evaluation)

회귀 모델 (Regression Model) 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	$\frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$

1. Machine Learning

1-10. 모델 평가 (Model Evaluation)

오차 행렬 (Confusion Matrix)

오차 행렬(또는 혼동 행렬)은 분류 모델의 성능을 평가하기 위해 실제 값(Actual Values)와 예측 값(Predictive Value)을 비교 하는 표이다.

		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=3, FP=2)

RECALL (red box around TP=3, FN=1)

정확도 $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

재현도 $Recall / Sensitivity = \frac{TP}{TP + FN}$

특이성 $Specificity = \frac{TN}{TN + FP}$

정밀도 $Precision = \frac{TP}{TP + FP}$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

1. Machine Learning

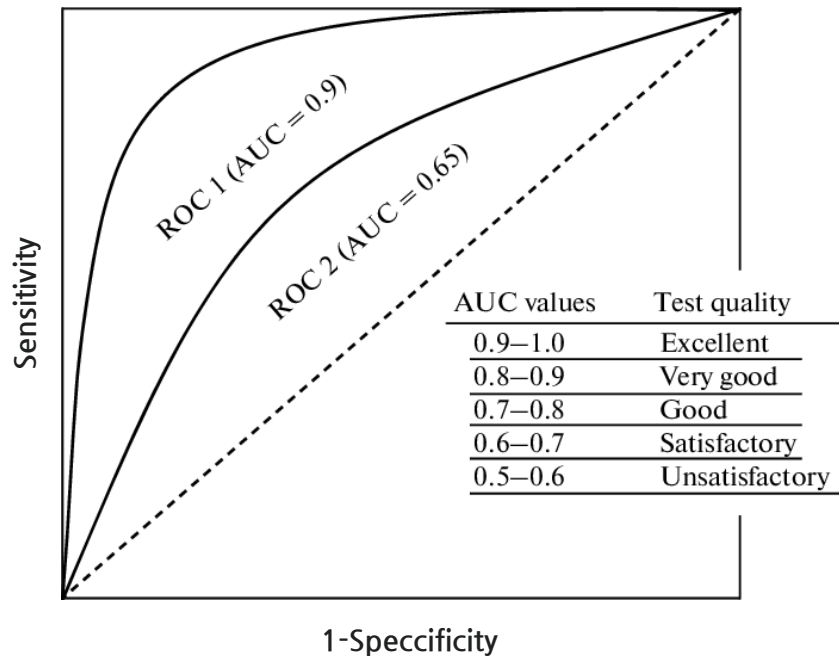
1-10. 모델 평가 (Model Evaluation)

ROC (Receiver Operating Characteristic) / AUC (Area Under Curve)

ROC/AUC는 분류 모델의 성능을 평가하기 위한 지표이다.

ROC/AUC는 오차 행렬의 민감도와 특이성을 그래프화 한 것으로 AUC 값이 1에 가까울 수록 분류를 잘 하는 모델이다.

ROC는 확률 곡선이고 AUC value는 ROC 곡선의 면적을 구한 값이다.



1. Machine Learning

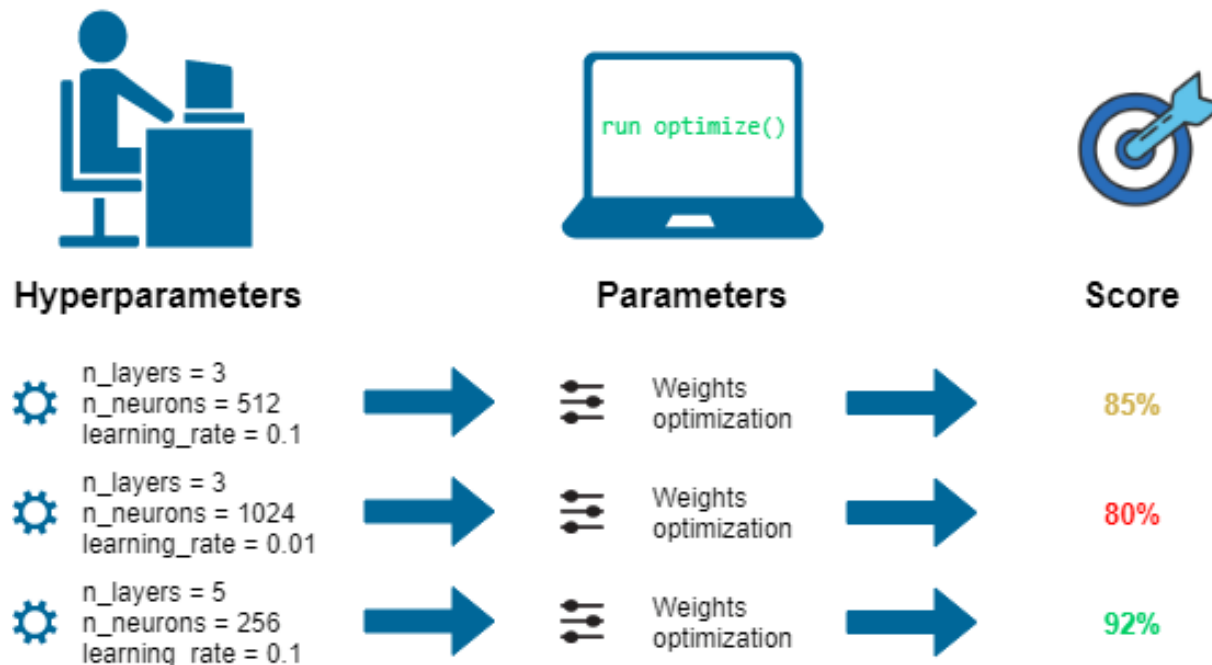
1-11. 하이퍼파라미터 튜닝 (Hyperparameter Tuning)

파라미터 (매개변수)는 모델의 학습 중에 정해지는 매개변수이고 (예: 가중치)

하이퍼파라미터 (초매개변수)는 학습 전에 사용자가 임의로 설정 가능한 매개변수를 말한다. (예: 학습률)

하이퍼파라미터 튜닝은 모델의 최대 성능을 위해 하이퍼파라미터의 올바른 조합을 결정하는 일이다.

* 참고 : Cross Validation (K-fold) / Grid Search



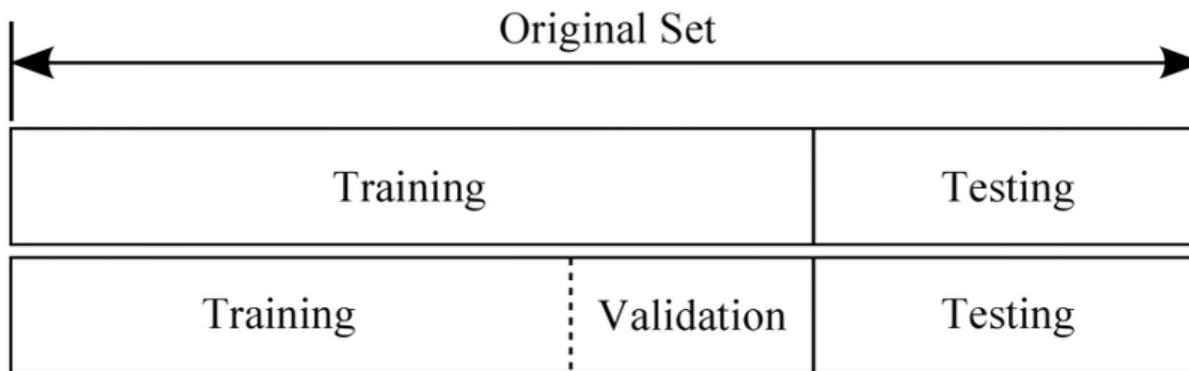
1. Machine Learning

1-11. 하이퍼파라미터 튜닝 (Hyperparameter Tuning)

데이터 분할

머신 러닝을 위한 데이터는 **Training / Validation / Test** 셋으로 나눈다.

- **Training Set** : 모델 생성 및 학습에 이용
- **Validation Set** :
 - 모델의 과적합 (Overfitting) 방지
 - 모델의 복잡도 축소
 - 모델의 파라미터(Parameter) 탐색
- **Test Set** : 모델의 예측 성능(Predictive Performance) 평가



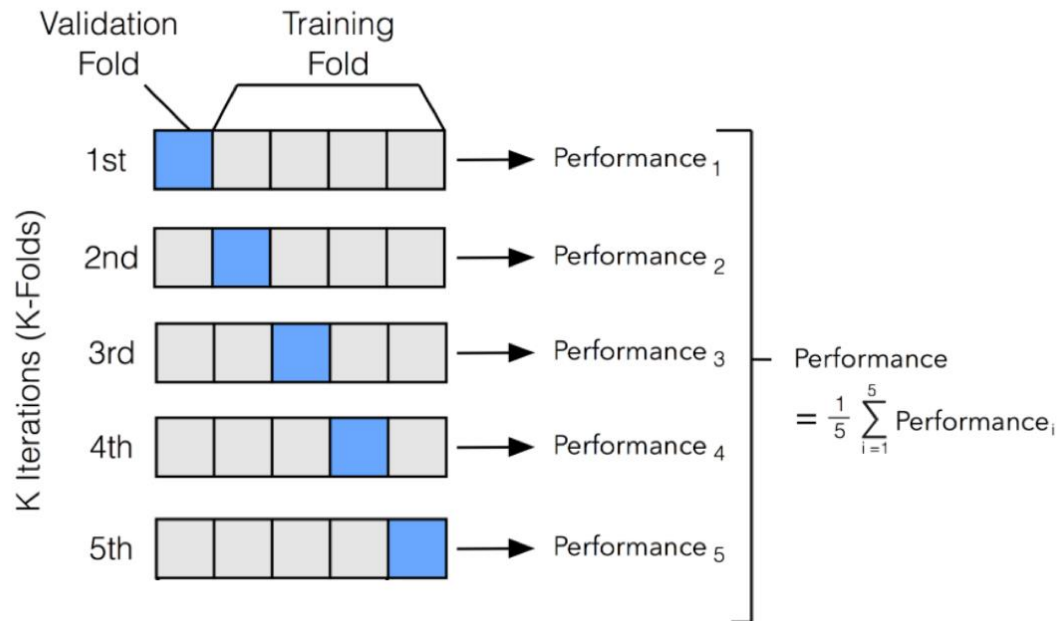
1. Machine Learning

1-11. 하이퍼파라미터 튜닝 (Hyperparameter Tuning)

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

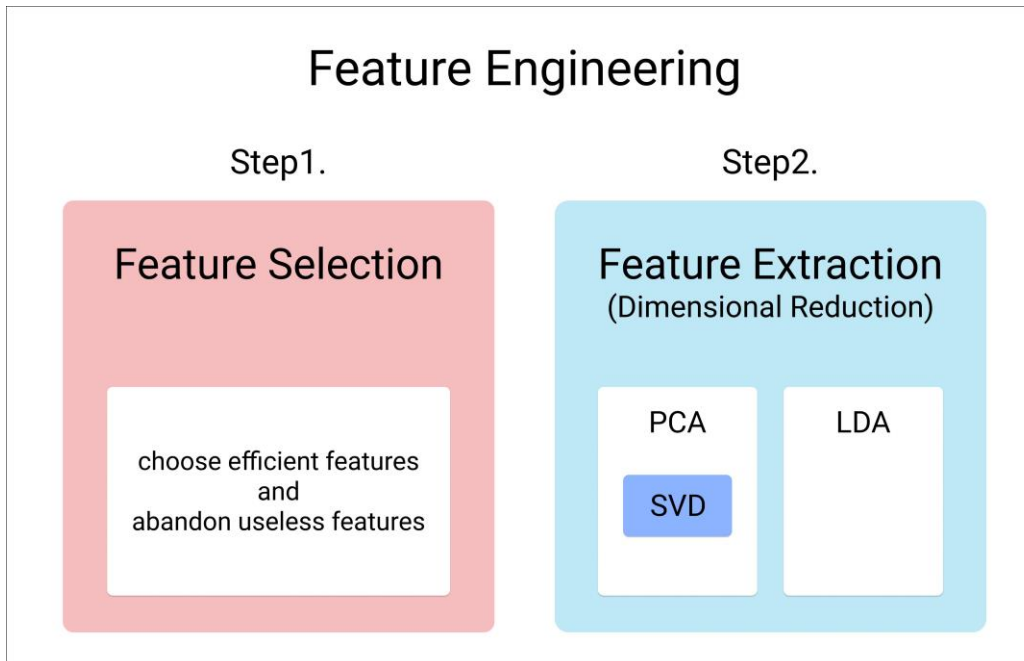


1. Machine Learning

1-12. 피처 엔지니어링 (Feature Engineering)

피처 엔지니어링은 모델의 성능 향상을 위해 Feature를 변환/생성 하는 과정이다. 탐색적 데이터 분석과 도메인 지식을 바탕으로 하며, 다음과 같은 측면으로 구분해볼 수 있다.

- Feature 결합/생성 : 2개 이상의 features 더하기, 빼기, 곱하기, 나누기 등
- Feature 선택 및 추출



End of Document