준비사항

<첫째 시간 파일 다운로드>

- 실습 데이터 셋 : https://imghub.insilicogen.com/media/files/heartdisease.csv
- 실습 자료 :2022_아주대학교_AI교육_ML_heartdisease_실습.ipynb

CoLab에서 열어두기!!

<둘째 시간 파일 다운로드>

- 실습 자료:
  - 주피터 노트북:deeplearning-classification-with-pneumonia-chest-x-ray-image-practice.ipynb
  - 실습 데이터:
    1. 다운로드(데이터 링크 이동→ chest_xray 클릭 → 다운로드 클릭 → 다운로드 완료되면 압축 해제)
    2. 구글 드라이브에 업로드(구글 드라이브로 이동 → 새로 만들기 → 폴더 업로드 → 압축해제한 chest_xray 폴더 업로드)
    3. 위 방식으로 업로드가 안되는 경우, 데이터 링크 → chest_xray 클릭 → [드라이브에 바로가기 추가] → 내 드라이브 선택 → 바로가기 추가

링크 : incodom.kr/ISG-Edu22-10

컴퓨터 pw : ajoumc1234

의료인공지능 부트캠프 2일차 첫째시간

머신러닝 이해와 의료데이터 기반 머신러닝 실습

2022. 10. 14. FLEX 권영인

INDEX

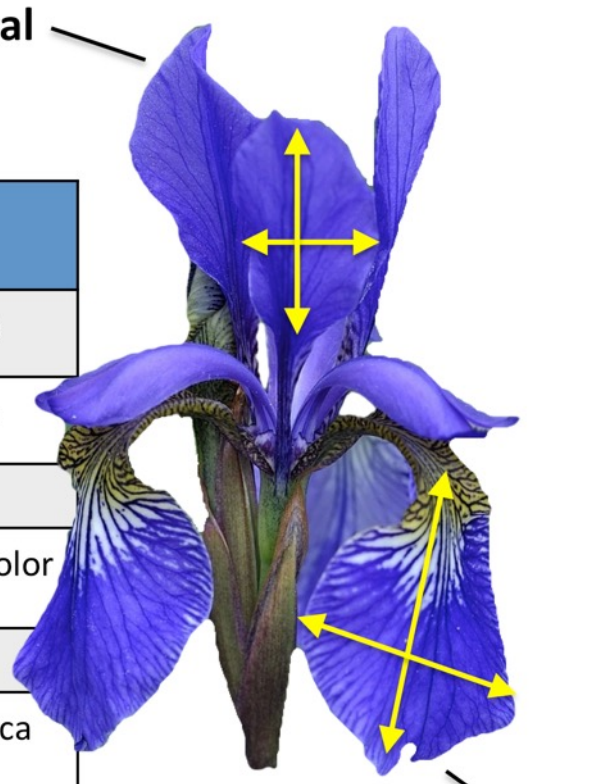
01

머신러닝 개요

01 iris dataset를 이용한 “지도학습-분류” 예시

Samples
(instances, observations)

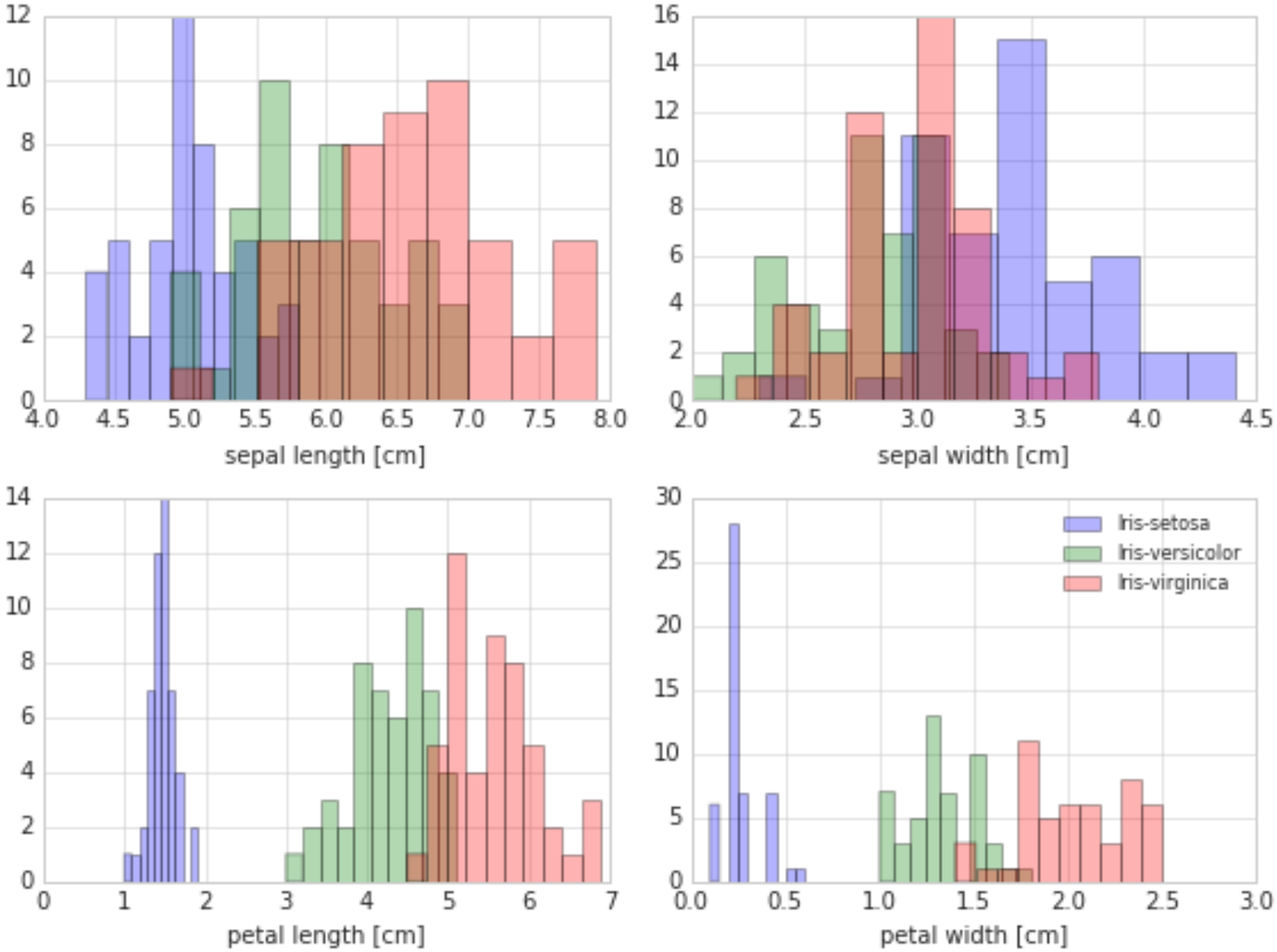
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

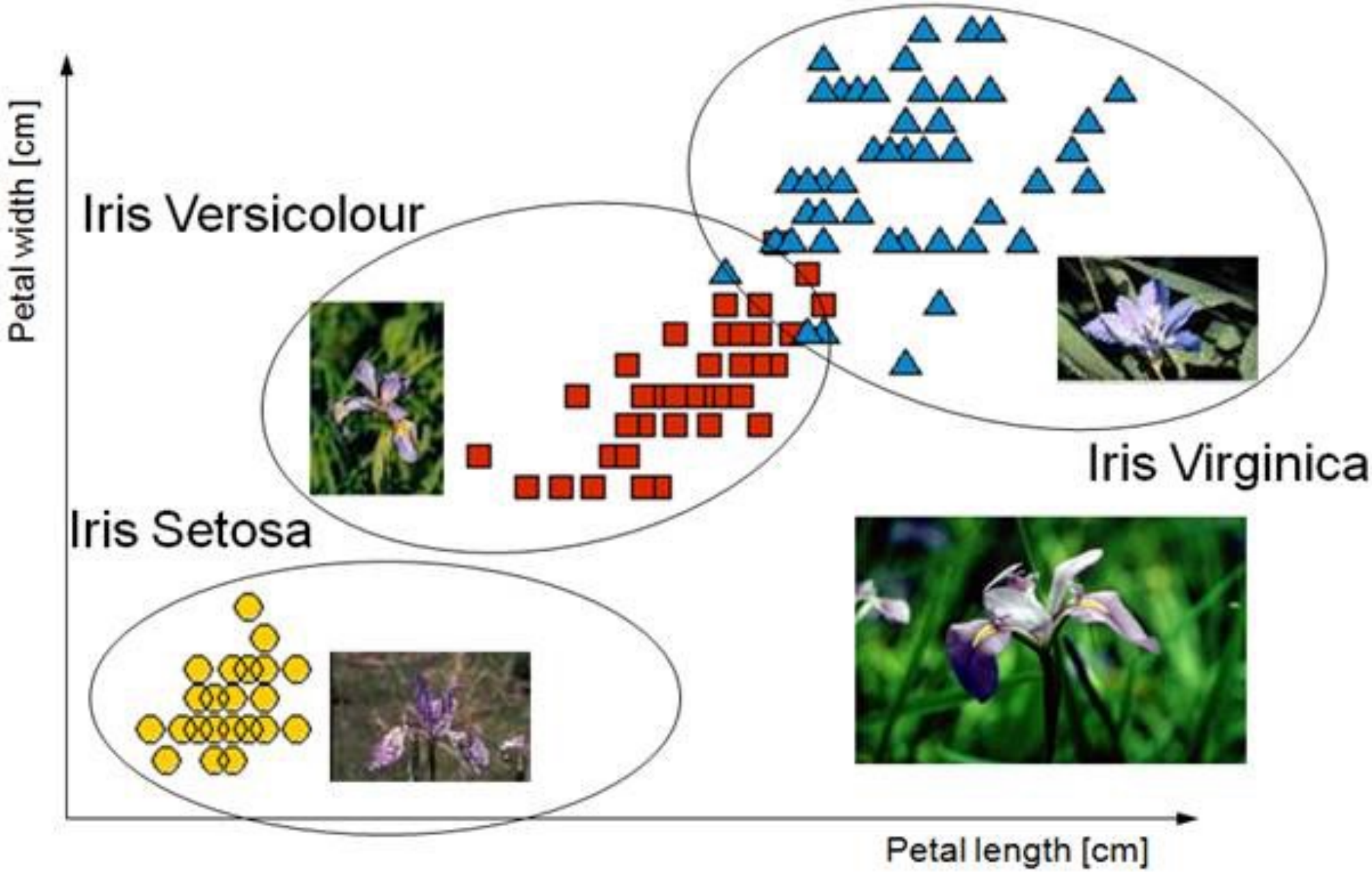


Features
(attributes, measurements, dimensions)

Class labels
(targets)

01 Feature 별 데이터 분포



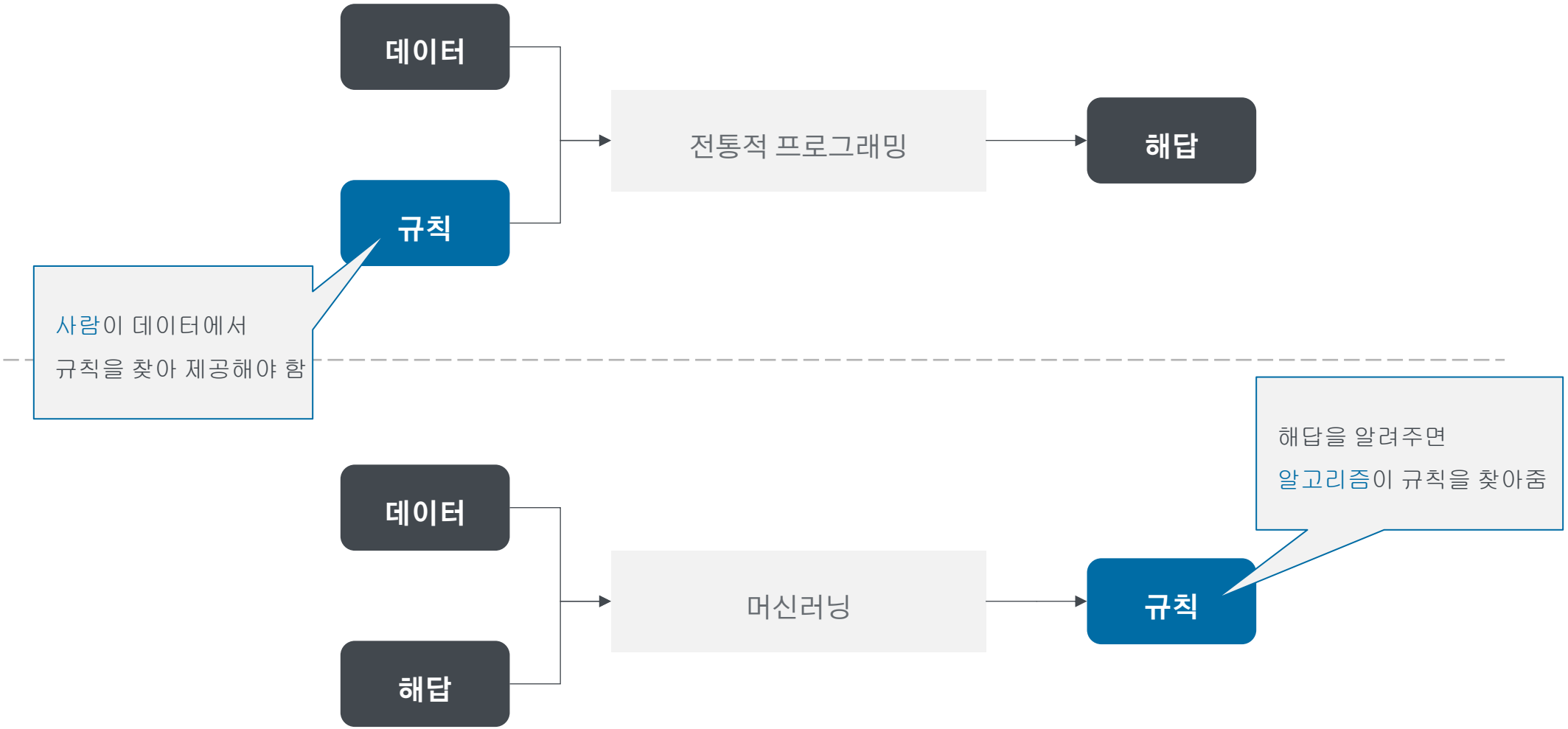

Petal width [cm]
Iris Versicolour
Iris Virginica
Iris Setosa
Petal length [cm]

머신러닝(Deep learning)이란?

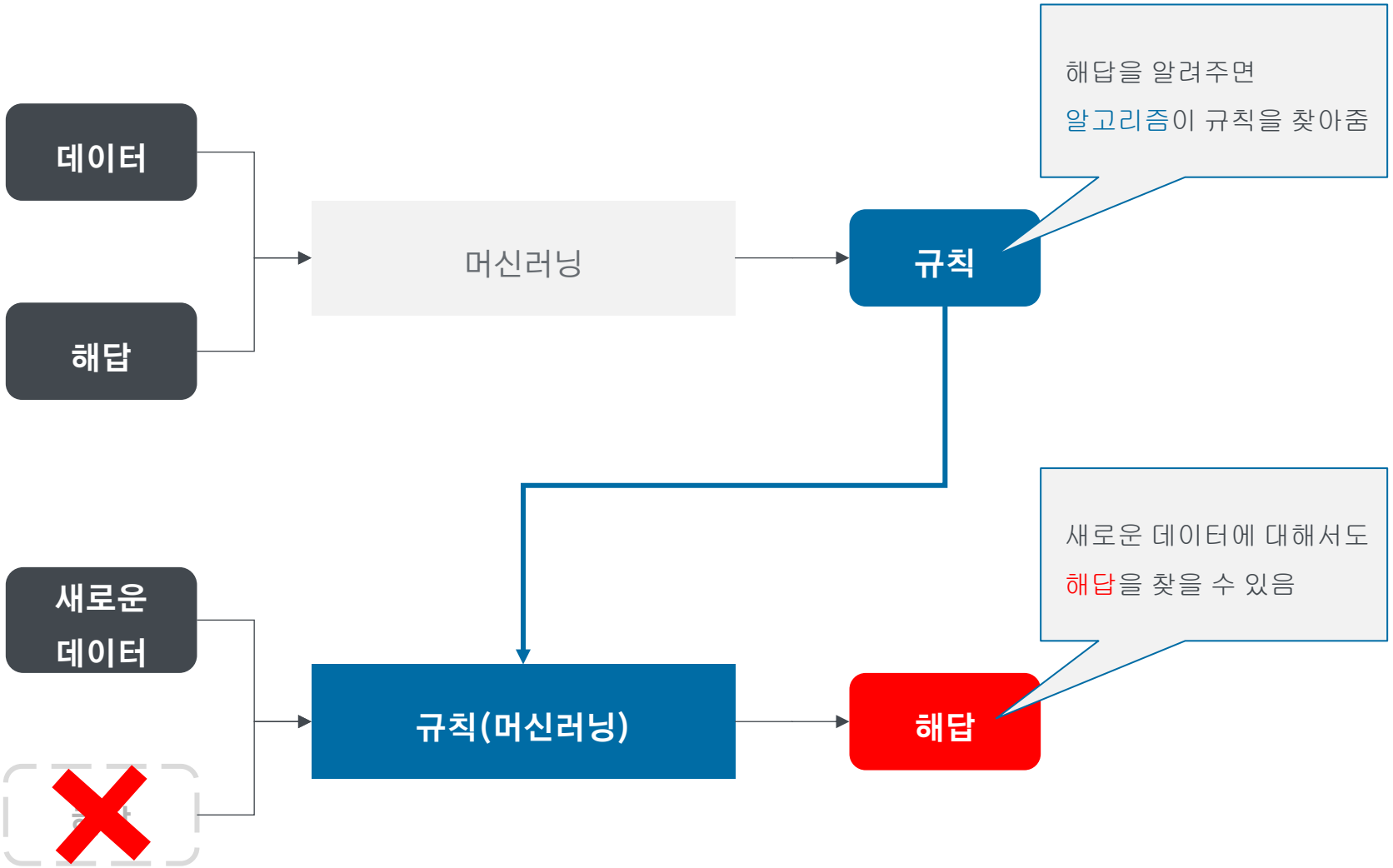
머신 러닝은 기존 데이터를 이용해
아직 일어나지 않은 **미지의 일을 예측**하기 위해 만들어진 기법

머신러닝은 명시적인 프로그래밍 없이
컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야

01 머신러닝 : 새로운 프로그래밍 패러다임



01 머신러닝 : 새로운 프로그래밍 패러다임



02

머신러닝 주요 개념

02 머신러닝 – 회귀 & 분류 차이

회귀 vs 분류

회귀란 ?

- 예측하고 싶은 타겟 변수가 숫자로 이루어졌을 때 회귀 (Regression)라는 머신러닝 방법을 사용

온도 X 2 = 판매량

원인 → 결과

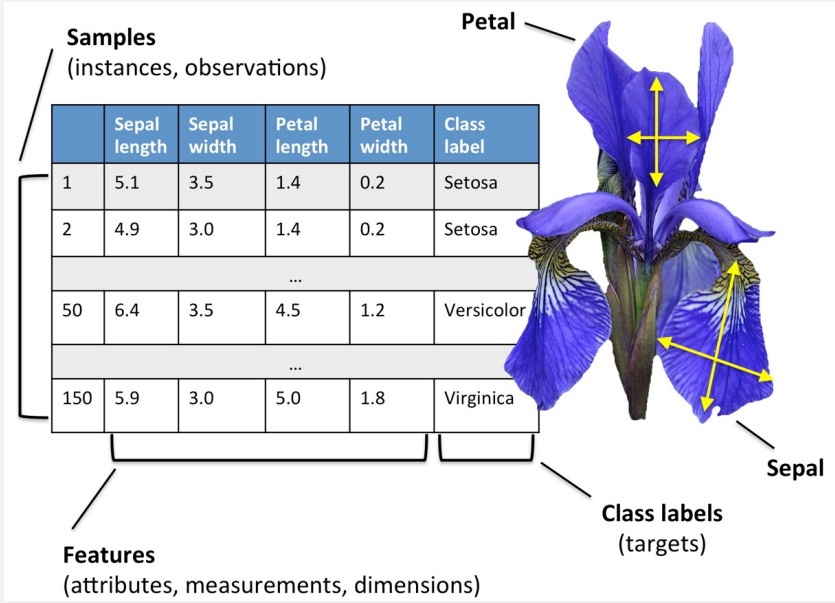
날짜	요일	온도	판매량
2020.1.3	금	20	40
2020.1.4	토	21	42
2020.1.5	일	22	44
2020.1.6	월	23	46
2020.1.7	화	24	48
2020.1.8	수	25	

과거의 데이터

미지의 데이터

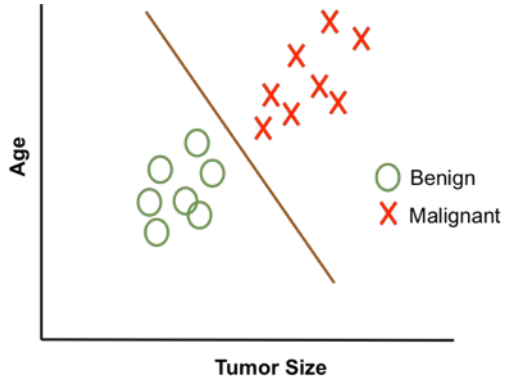
분류란 ?

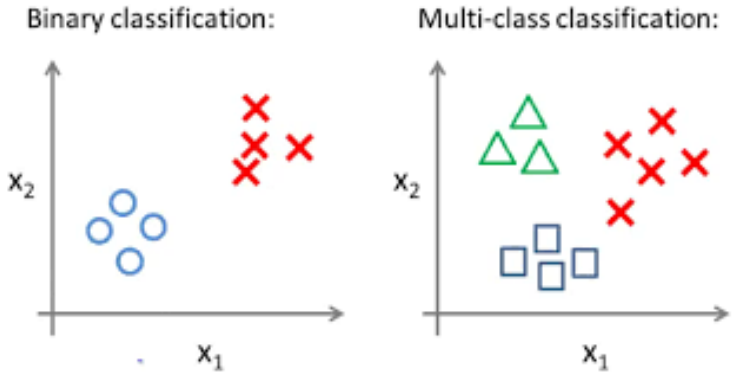
- 예측하고 싶은 타겟변수가 이름이나 카테고리화 되어 있을 때, 분류 (Classification)를 이용

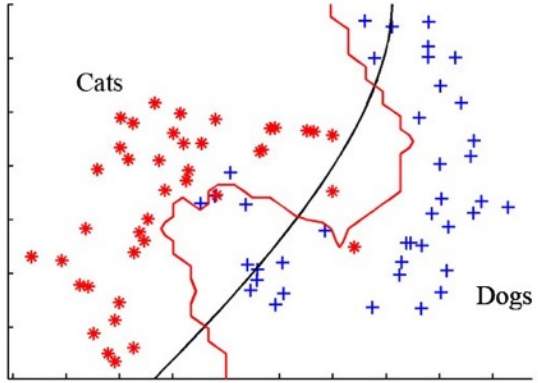


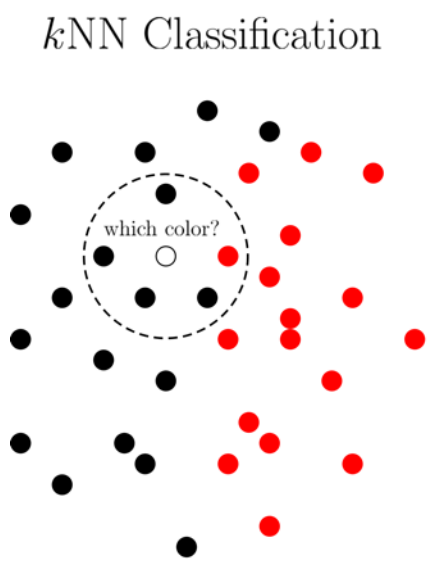
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

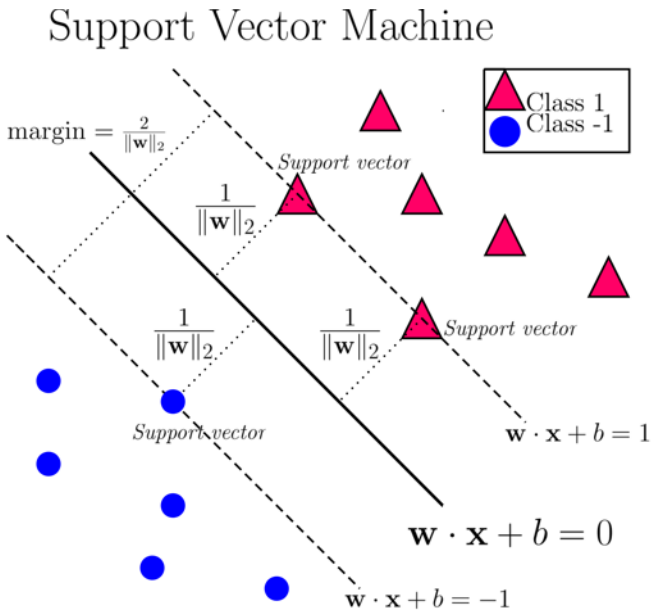
02 기계학습은 다차원 공간에 경계 선을 그리는 것 (classification)

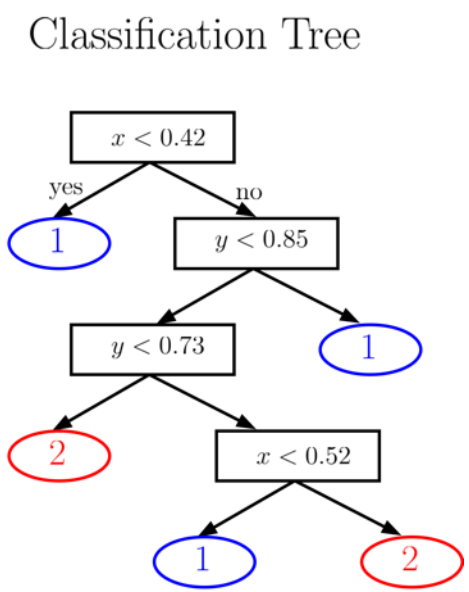






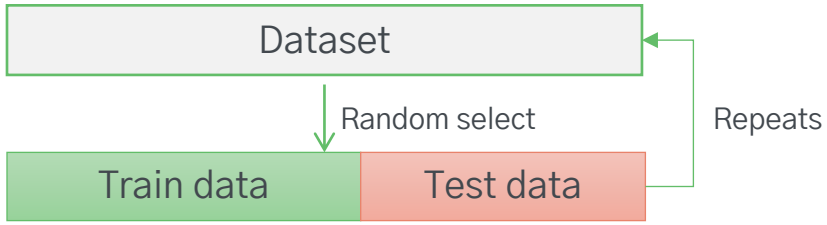




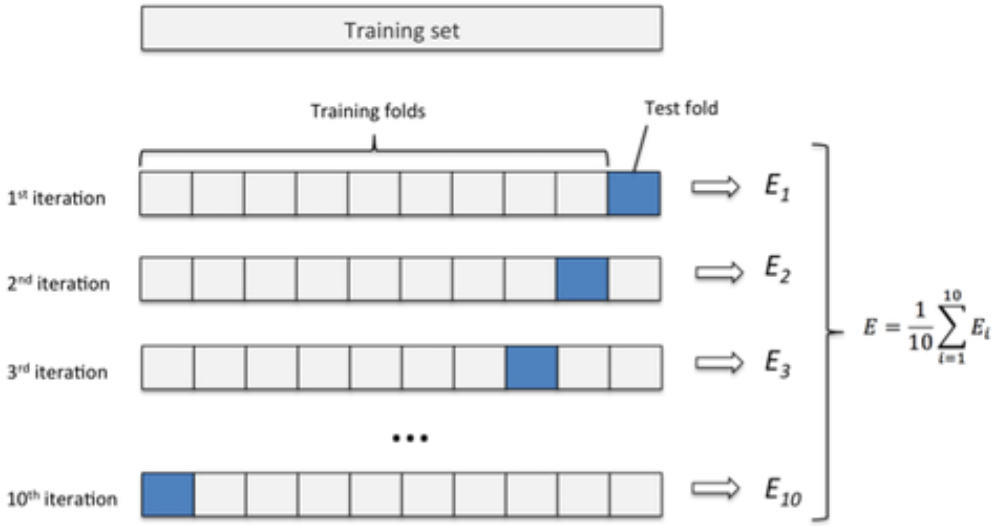


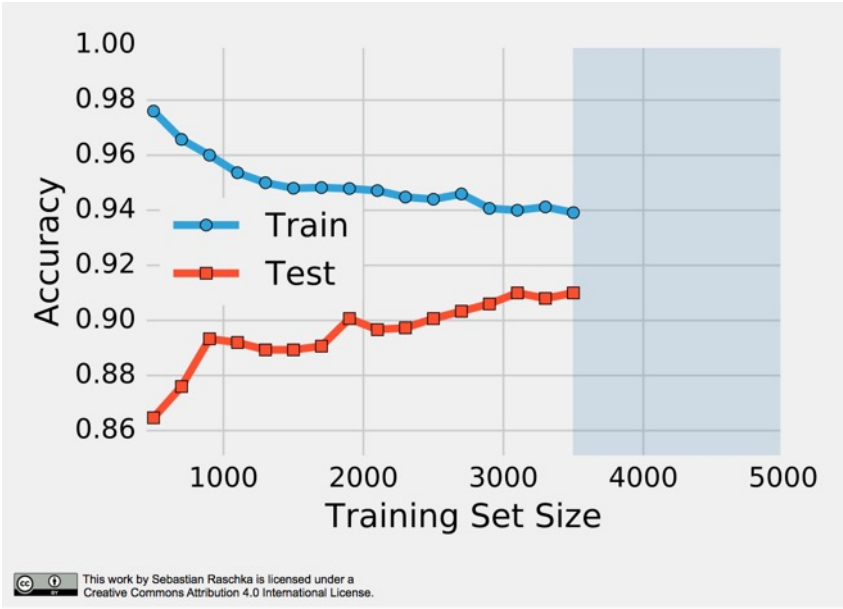
02 Training set and Test set

<Random selection>



<K-fold method>



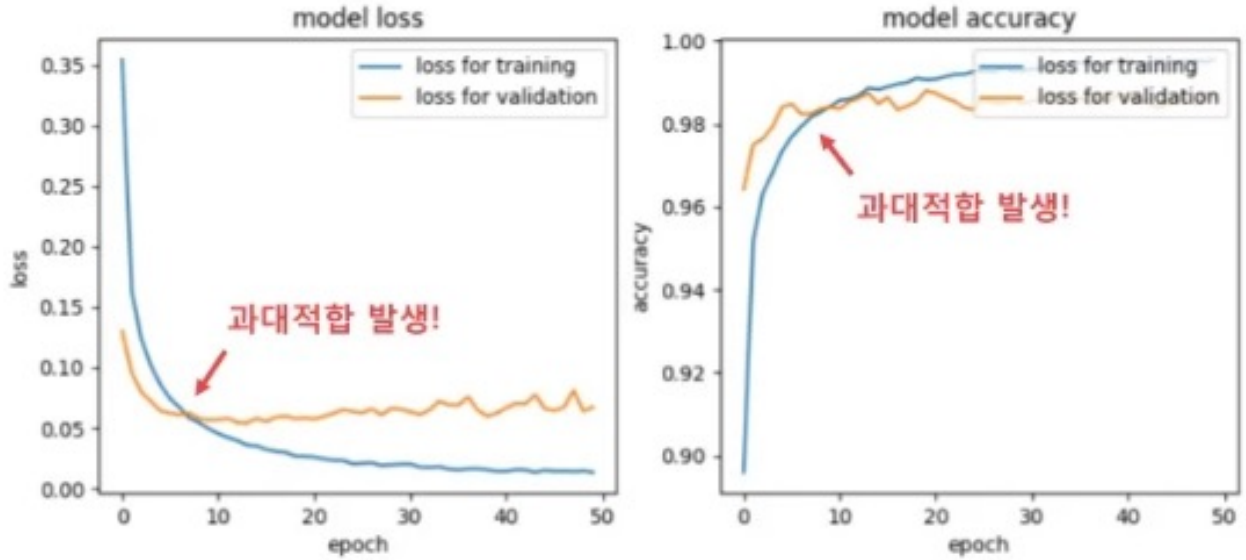


02 훈련데이터 과대적합 (Overfitting)

과대적합이란 모델이 훈련 데이터에만 너무 잘 학습 되어 일반성이 떨어짐을 의미
과대적합에 의해 학습하지 않은 새로운 데이터에 대한 모델의 예측 성능이 떨어짐

해결방법

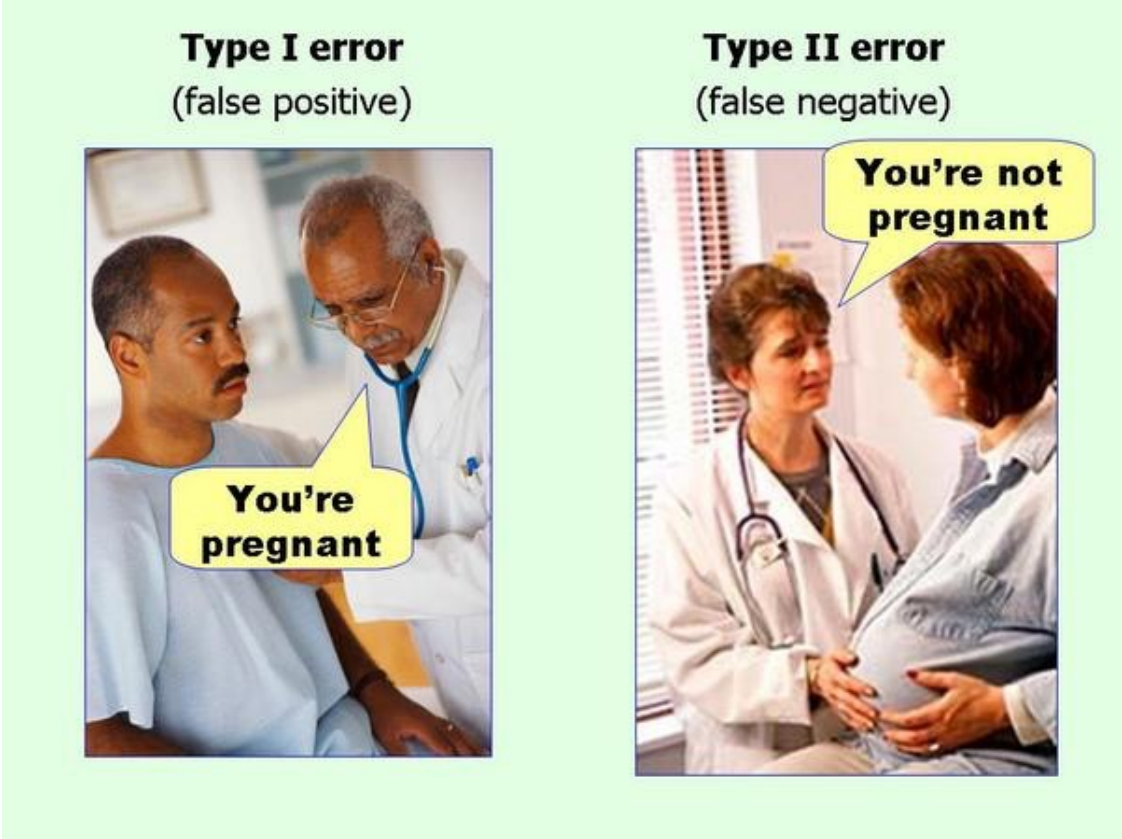
- 파라미터의 수가 적은 모델 or 훈련 데이터의 특성 수를 줄이기
- 훈련 데이터의 수를 늘리기



02 거짓 양성과 거짓 음성 (False positive & False negative)

틀리는 것에도 종류가 있다.

- 맞는 것을 아니라고! → False negative
- 아닌 것을 맞다고 → False positive



02 Confusion Matrix

(암환자) (암환자 아님)

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	Condition negative	
Fecal occult blood screen test outcome	Test outcome positive (암환자라 판정)	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test outcome negative (암환자가 아니라고 판정)	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

(민감도) (특이도)

맞는것을 맞다고 하는 정확도

아닌것을 아니라고 하는 정확도

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

false discovery rate (FDR)

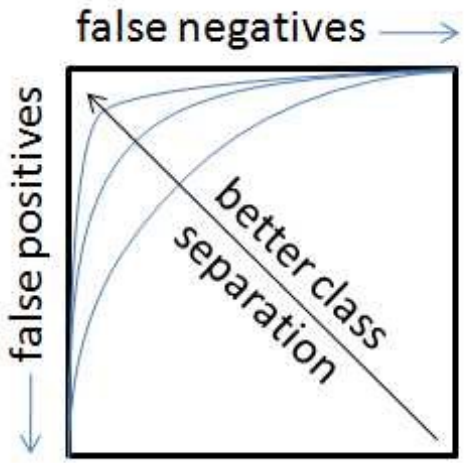
$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

false omission rate (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$



실습 데이터 셋

HeartDisease_Dataset.csv

From UCI Machine Learning Repository

303 rows x 14 columns

index	변수명	한글 변수명	변수 설명
1	age	나이	range : 29~77
2	sex	성별	M = male; F = female
3	cp	가슴통증	0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic
4	trestbps	혈압	In mm Hg / range : 94~200
5	chol	콜레스테롤	in mg/dl / range : 126~564
6	fbs	공복 혈당	1 = true; 0 = false (fasting blood sugar > 120 mg/dl)
7	restecg	심전도	0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
8	thalach	최대 심장박동 수	range : 71~202
9	exang	운동 유발 협심증	1 = yes; 0 = no
10	oldpeak	ST 우울증	range : 0~6.2
11	slope	ST segment 기울기	0 = upsloping; 1 = flat; 2 = downsloping
12	ca	혈관 수	0, 1, 2, 3, 4
13	thal	빈혈 여부	3 = normal; 6 = fixed defect; 7 = reversable defect
14	target	심장질환 여부	0 = normal; 1 = heartdisease

데이터 출처 :
https://archive.ics.uci.edu/ml/datasets/heart+disease

데이터 전처리 과정

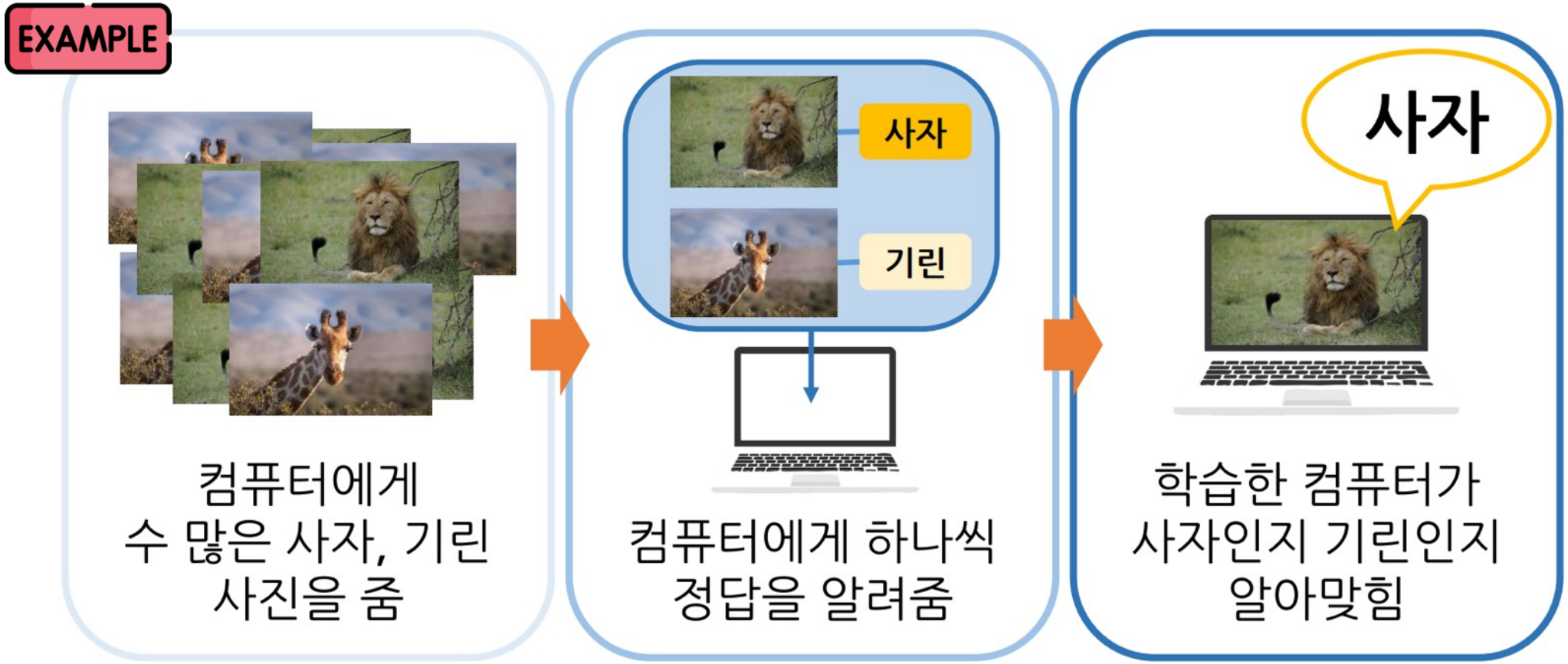
- 데이터 로드 및 확인
- Nan 값 제거
- Coefficient 확인
- Categorical & Continuous Column 나누기
- One-Hot Encoding
- StandardScaler 적용
- Train, Test dataset 분리

03

지도학습 – KNN

1) 지도학습

정답을 제공한 상태로 학습을 진행한 알고리즘

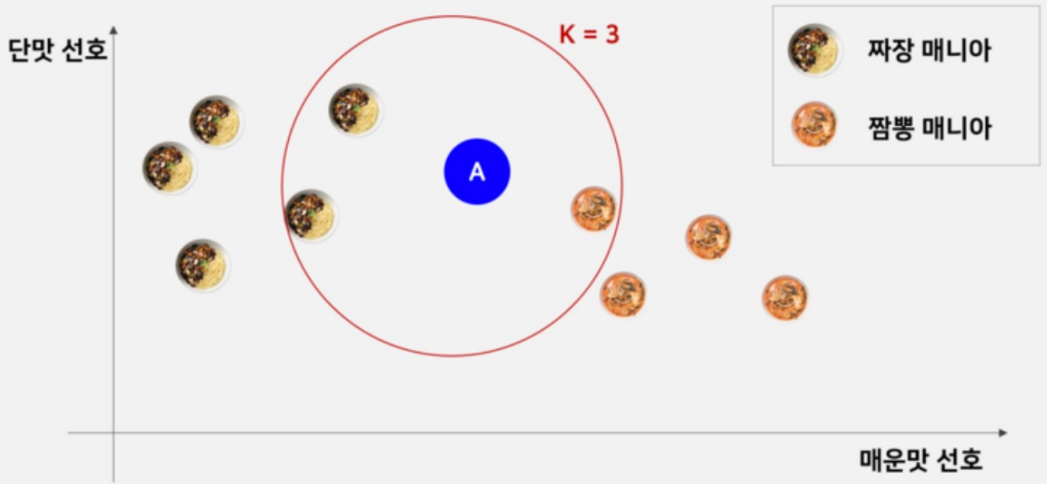


03 지도학습 – KNN(K-Nearest Neighbors)

KNN

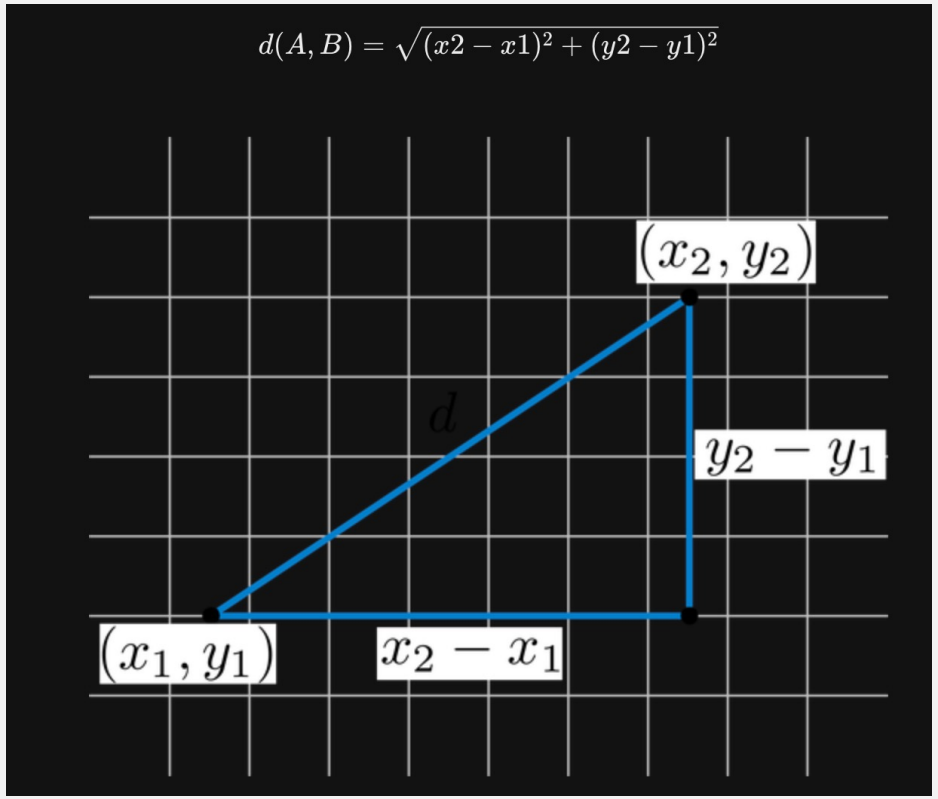
정의

- 데이터를 가장 가까운 속성에 따라 분류하여 레이블링 하는 기본적인 알고리즘
- **“일종의 다수결”**
- K개의 가까운 이웃의 속성에 따라 분류
- 거리기반 분류분석 모델로 거리를 기반으로 분류하는 알고리즘
- 유클리드 거리, 맨해튼 거리 등
- 이미지 처리, 글자 및 얼굴 인식, 영화나 상품 추천 등 추천 시스템, 유전자 데이터의 패턴 인식

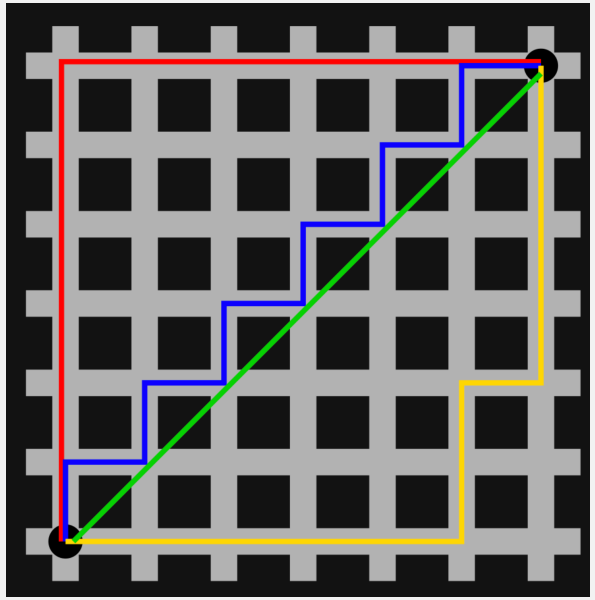


유클리드 거리 및 맨해튼 거리

유클리드 거리



맨해튼 거리



$$d(A,B) = \sum_{i=1}^{n} |a_i - b_i|$$

KNN

장점

- 단순하고 효율적임 (높은 정확도)
- 수치 기반 데이터 분류 작업에서는 매우 우수한 성능 보유
- 기저 데이터 분포에 대한 가정을 하지 않음

단점

- 데이터 양이 방대해 질수록 분류 단계가 매우 느려짐 -> 모든 데이터를 비교
- 명목 데이터 및 누락 데이터를 위한 추가 처리가 필수적

04

지도학습 - Linear Regression & Logistic Regression

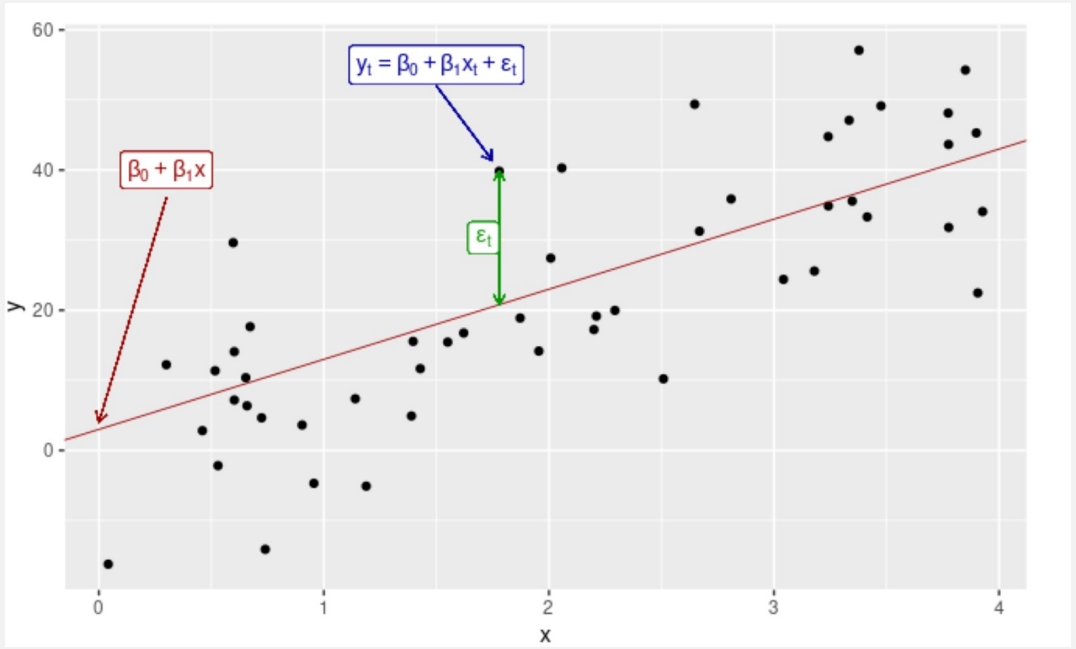
Linear Regression

정의

- 선형회귀 모델은 '회귀 계수'를 선형 결합으로 표현할 수 있는 모델
- 목표 예상변수 y와 하나의 예측변수 x 사이의 선형 관계
- 계수 β0와 β1는 각각 직선의 절편과 기울기

- 다중 회귀 모델 : 두개 이상의 예측 변수가 존재할 때 사용

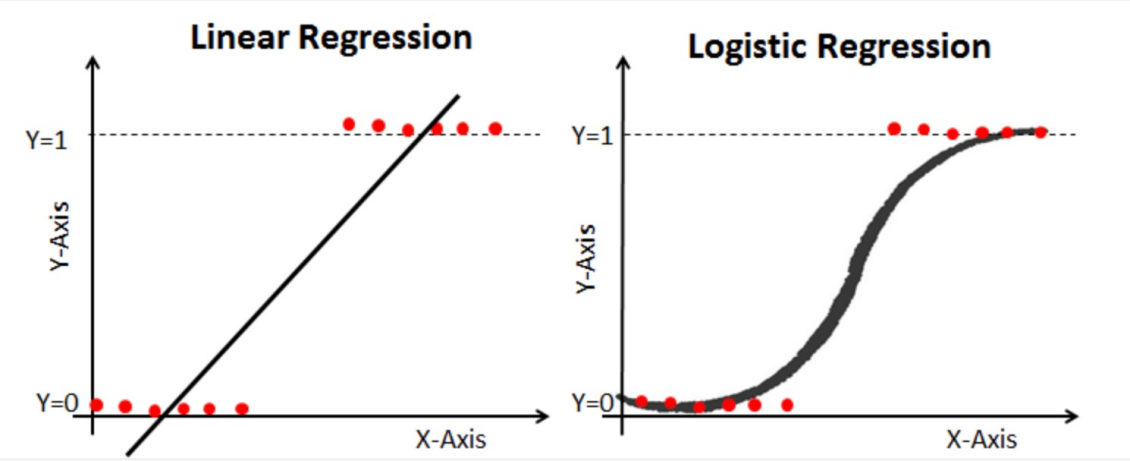
$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$



Logistic Regression

정의

- 로지스틱 회귀 : 데이터가 어떤 범주에 속할 확률을 0에서 1사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도 학습 모델
- 데이터의 종속변수가 범주형 (0 또는 1)일 때 사용
- 일반적으로 확률이 0.5보다 크면 사건이 일어나고, 0.5보다 작으면 사건이 일어나지 않는 것으로 예측
- EX). 성별, 합격/불합격, 양성/음성 등



Logistic Regression 모델의 장 & 단점

장점

- 분류 모델로서의 역할 뿐만 아니라 확률값 얻을 수 있음 (Ex). 위험도 분석)
- 분류 문제에서 베이스 모형으로 활용 가능
- 회귀 계수의 해석이 가능

단점

- 이상치에 민감함.
- 누락 데이터를 위한 추가 처리가 필수적

05

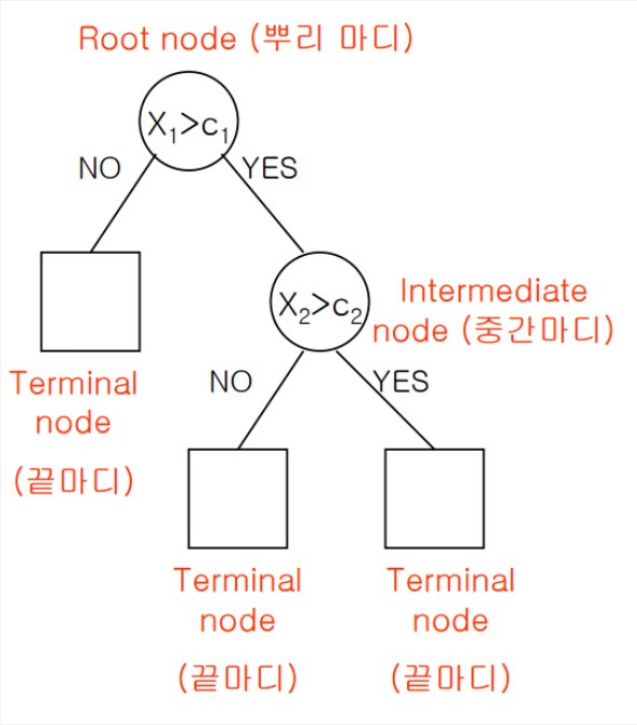
지도학습 – Decision Tree

05 지도학습 – Decision Tree (결정 트리)

결정 트리 (Decision Tree)

정의

- 분류(Classification)와 회귀(Regression) 모두 가능한 지도 학습 모델
- 스무고개
- 특정 기준(질문)에 따라 데이터를 구분하는 모델
- 한번의 분기 때마다 변수 영역을 두개로 구분
- 노드 : 질문이나 정답을 담은 네모 상자
- Root Node : 맨 처음 분류 기준
- Terminal Node : 맨 마지막 노드



Decision Tree 모델의 장 & 단점

장점

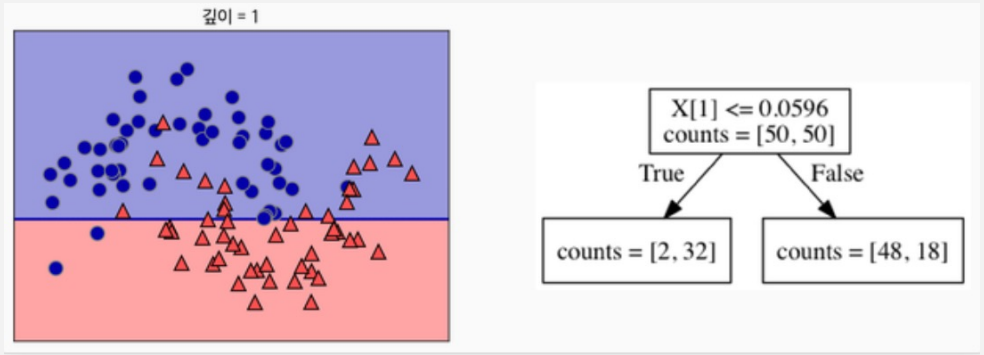
- 이해하기 쉬운 규칙에 의거한 분석이기에 결과 해석 용이함.
- 비모수적 모델 -> 통계 모델과 다르게 요구되는 가정이 없음
  -> 분류와 회귀에 널리 사용되는 모델
- 변수의 중요성 비교 가능

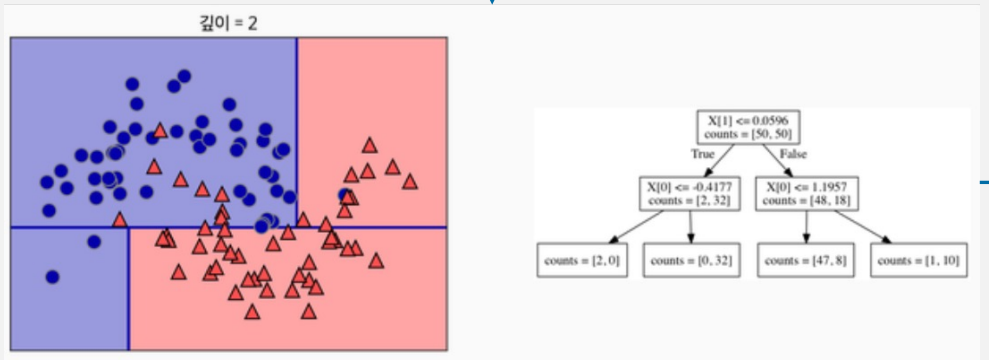
단점

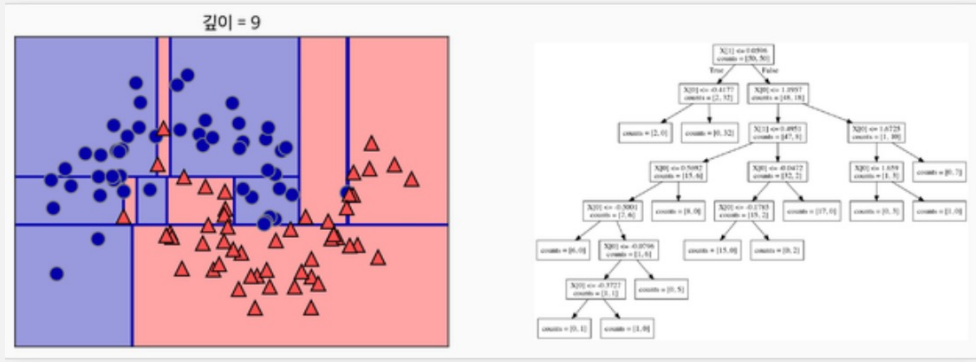
- 안정성이 떨어짐 (데이터 수, 과대적합)
- 연속형 변수 값을 예측할 때 적합하지 않음.
- 훈련 데이터 밖 새로운 데이터에 대한 예측할 능력이 떨어짐.

05 지도학습 – Decision Tree (결정 트리)

결정 트리 (Decision Tree) 프로세스







06

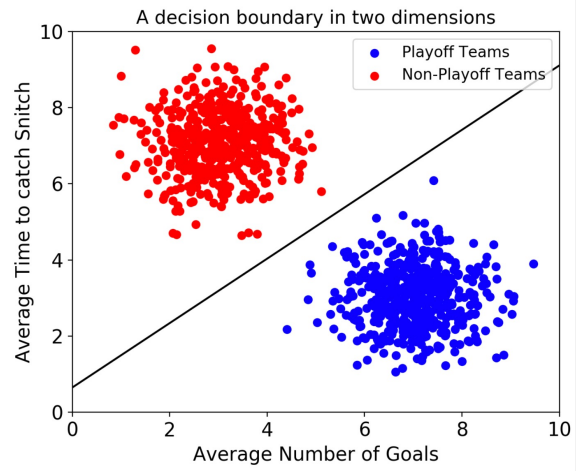
지도학습 – Support Vector Machine (SVM)

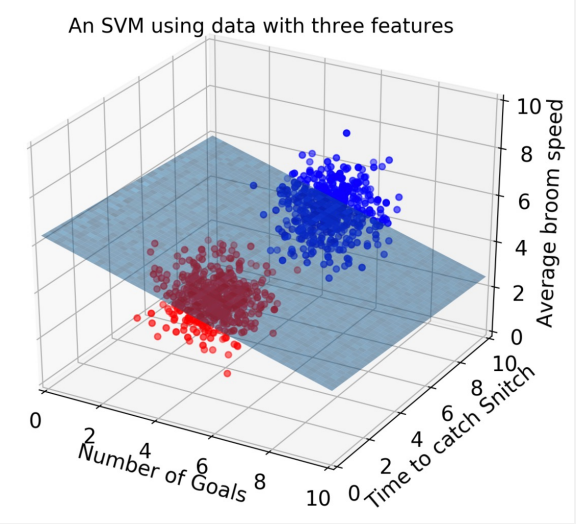
SVM(Supporting Vector Machine)

정의

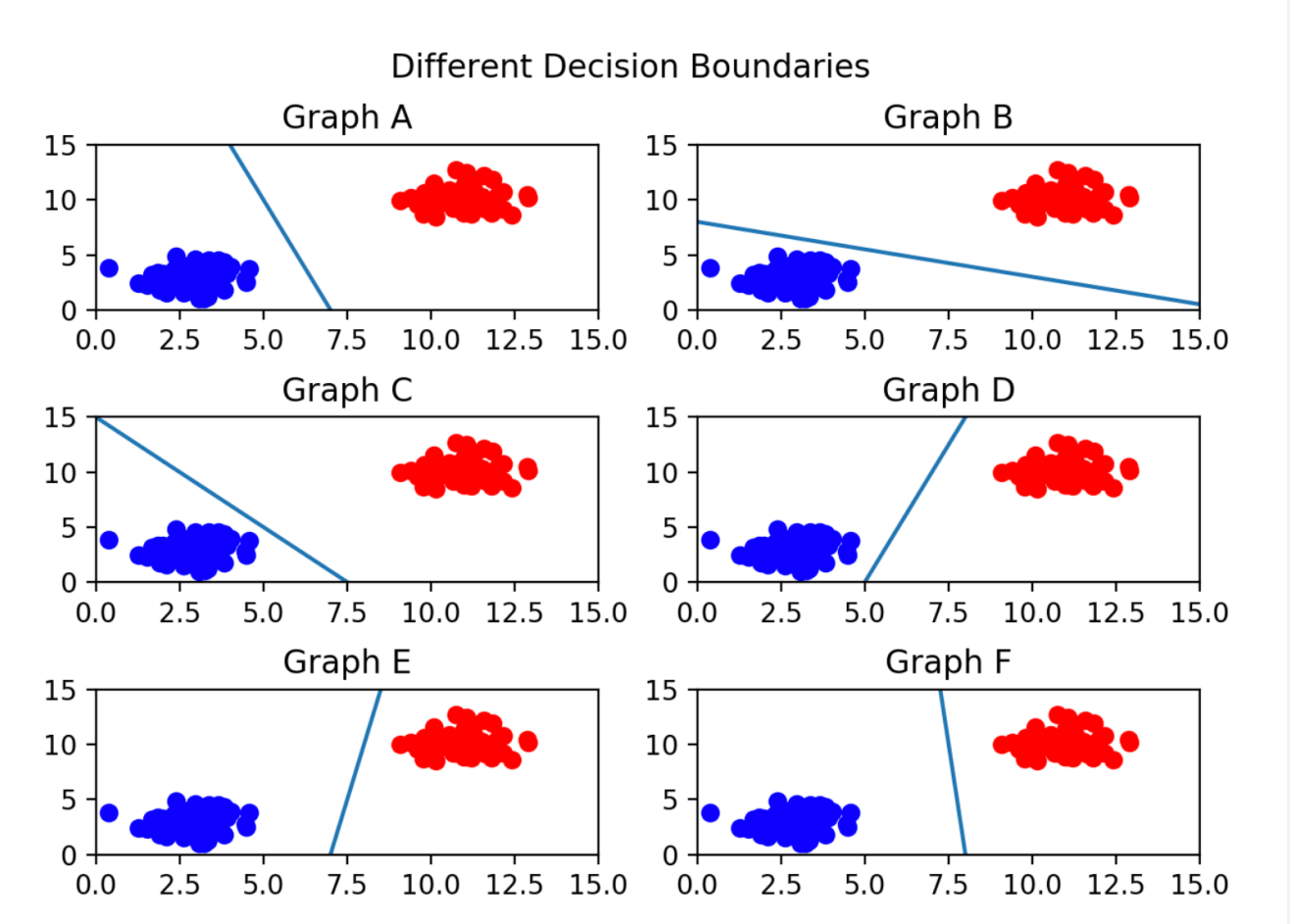
- 결정 경계 즉, 분류를 위한 기준 선을 정의하는 모델
- 데이터에 2개 속성이 있다면 결정 경계는 선 (속성 3개면 -> ??)
- 초평면 (Hyperplane) : 속성이 n개처럼 무수히 많을 때

  생기는 경절 경계의 고차원





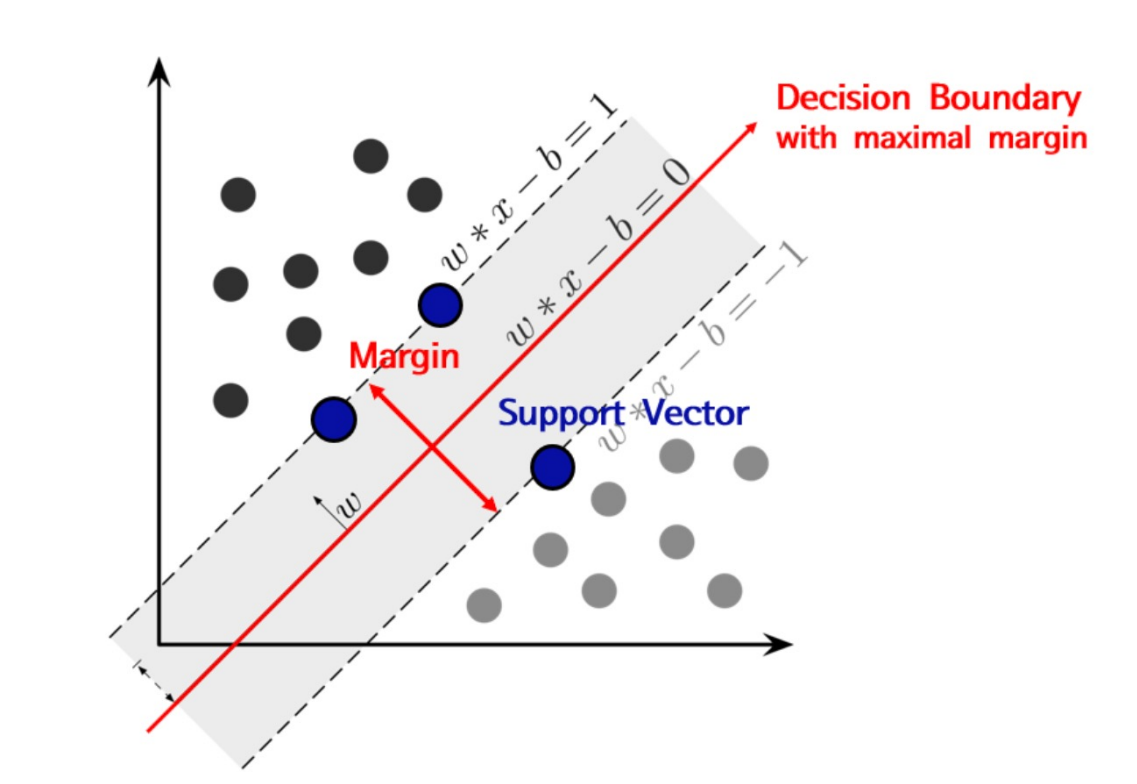
SVM(Supporting Vector Machine)



SVM(Supporting Vector Machine)

정의

- Margin : 결정 경계와 서포트 벡터 사이의 거리
- Support Vectors 정의 : 결정 경계(Decision Boundary)와 가까이 있는 데이터 포인트들을 의미
- **결론 : 최적의 결정 경계의 조건은 마진을 최대화**
- Why ?
- 마진이 넓으면 새로운 데이터가 입력 되었을 때, 통계적으로 잘 분류가 가능하여 오차가 적음



Support Vector Machine의 장 & 단점

장점

- 회귀 및 분류 예측 문제 모두 사용 가능
- 오류 데이터에 대한 영향이 다른 모델에 비해 현저히 적음 -> 일반화 능력이 좋음
- 과적합 되는 경우가 적음

단점

- 최적의 모델을 찾기 위해선 여러 개의 테스트 필요 (커널, hyper-parameter)
- 학습 속도가 느림 -> 고차원으로 갈수록 부담스러움
- 해석이 어렵고 복잡한 블랙박스 형태로 되어 있어 해석은 일부 가능
  -> 결과에 대한 설명력이 떨어짐.

07

Bagging – Random Forest

07 앙상블모델 – Bagging (Bootstrap Aggregating)

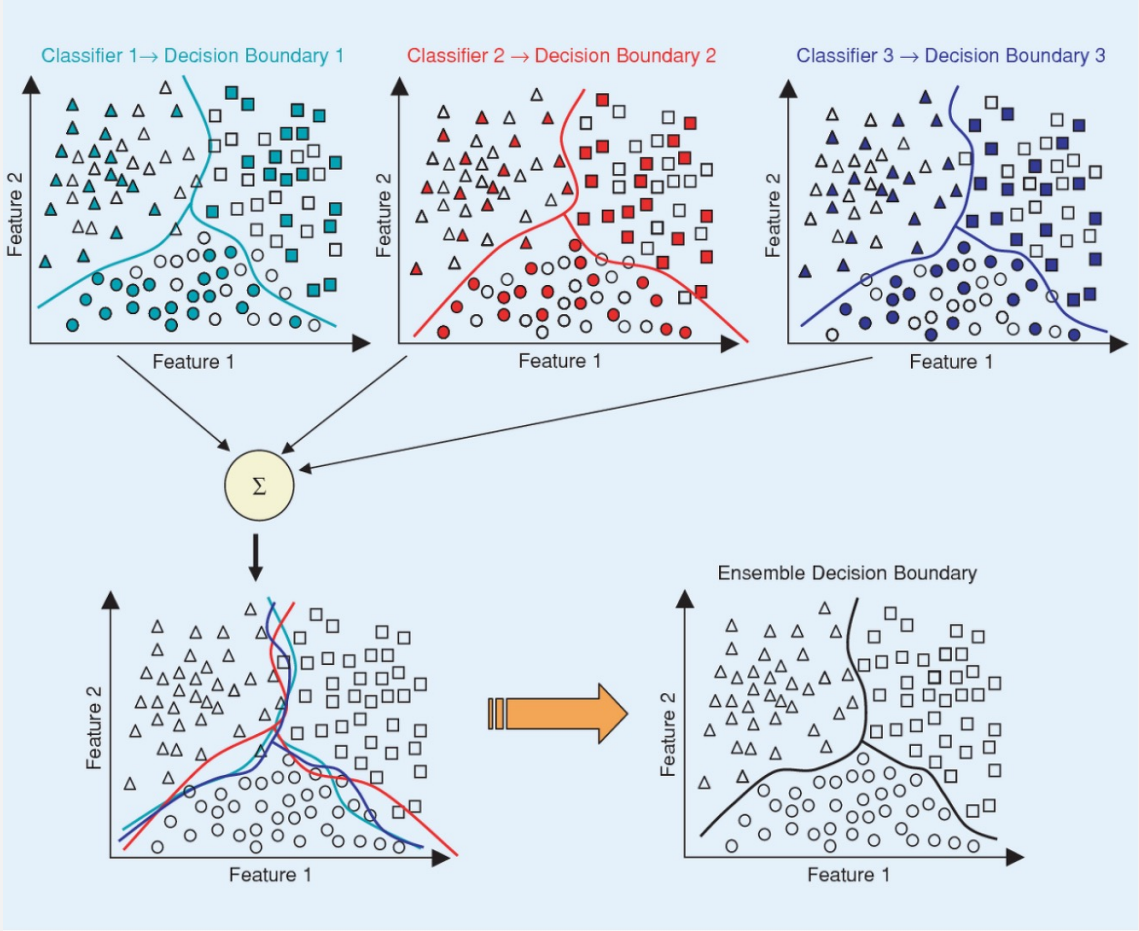
Bagging 기본 정의 및 용어 정리

정의

- Bagging : 기존 학습 데이터로부터 랜덤하게 '복원추출'하여 동일한 사이즈의 데이터셋을 여러 개 만들어 앙상블을 구성하는 여러 모델을 학습시키는 방법
- 반복적인 학습 모델 결합을 통해 노이즈에 대한 위험 방지

모델 결합 방법

- Majority Voting
- Weighted Voting
- Stacking

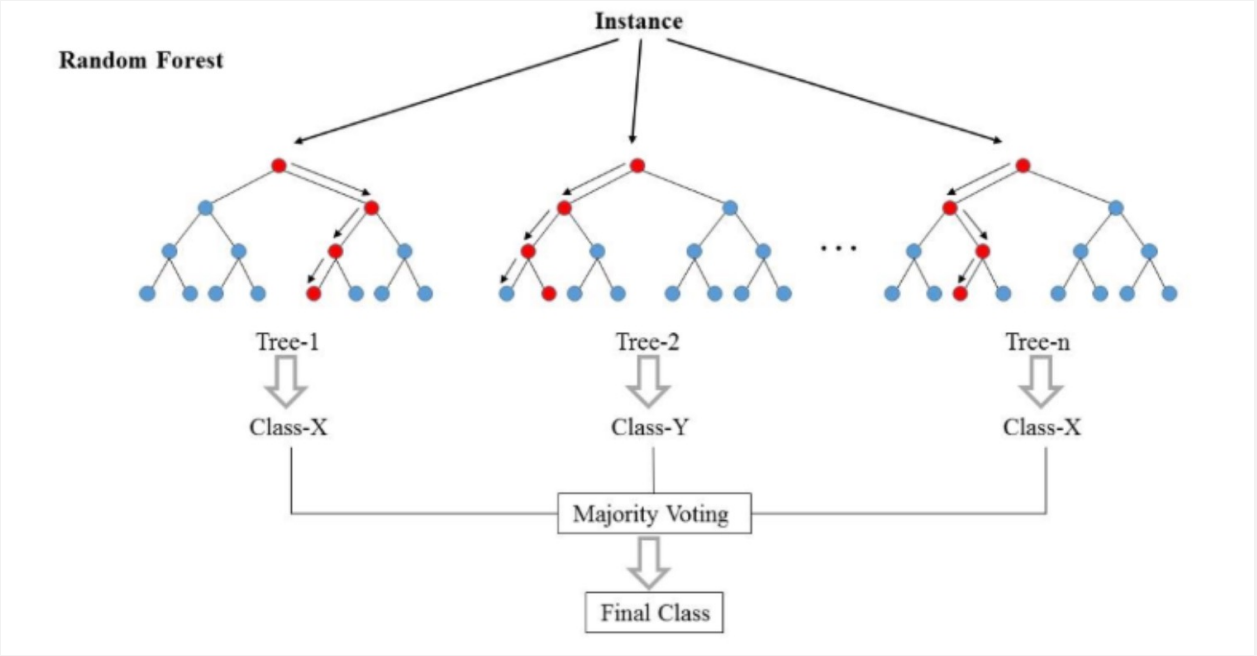


07 앙상블모델 – Random Forest

Random Forest 기본 정의 및 용어 정리

정의

- Random Forest : 의사결정나무 모델 여러 개를 훈련시켜 결과를 종합해 예측하는 앙상블 알고리즘
- 서로 모두 다른 트리 구조 구성
- 트리의 갯수로 정확도 개선 가능하나 elbow point 존재
- 모델 내부를 볼 수 없다는 단점 존재



Random Forest 모델의 장 & 단점

장점

- 회귀와 분류 문제 모두 사용 가능함
- 대용량 데이터 처리에 효과적임
- 앞선 모델에 비해 과대적합 문제를 최소화하여 모델의 정확도 향상시킴

단점

- 데이터 크기에 비례해서 수백개에서 수천개의 의사결정나무를 형성하기 때문에 예측에 오랜 시간이 걸림.
- 모델 특성상 모든 트리 모델을 다 확인이 불가해서 해석이 어려움.

08

Boosting – AdaBoost & GBM

08 앙상블모델 – Boosting

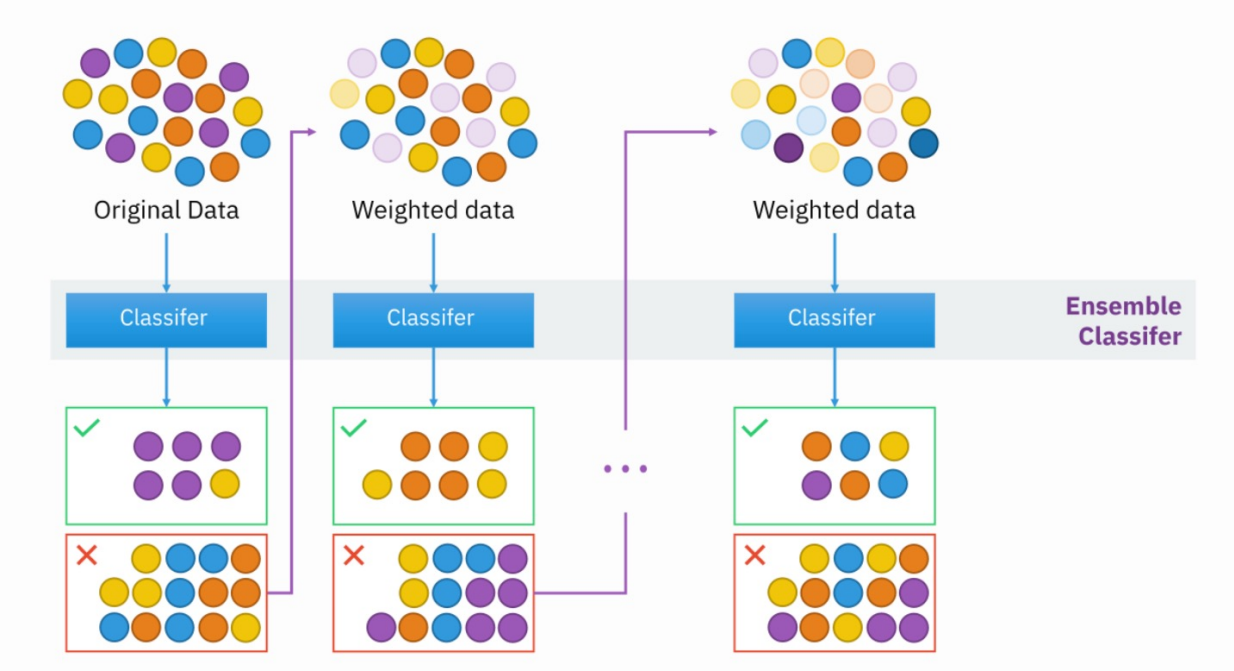
Boosting 기본 정의 및 용어 정리

정의

- Boosting : 머신러닝 앙상블 기법 중 하나로, 약한 학습기를 순차적으로 여러 개 결합하여 예측 혹은 분류 성능을 높이는 알고리즘

진행방식

- 순차적으로 학습, 예측 하면서 다음 학습 알고리즘에 가중치를 부여하여 학습과 예측을 진행하는 방식

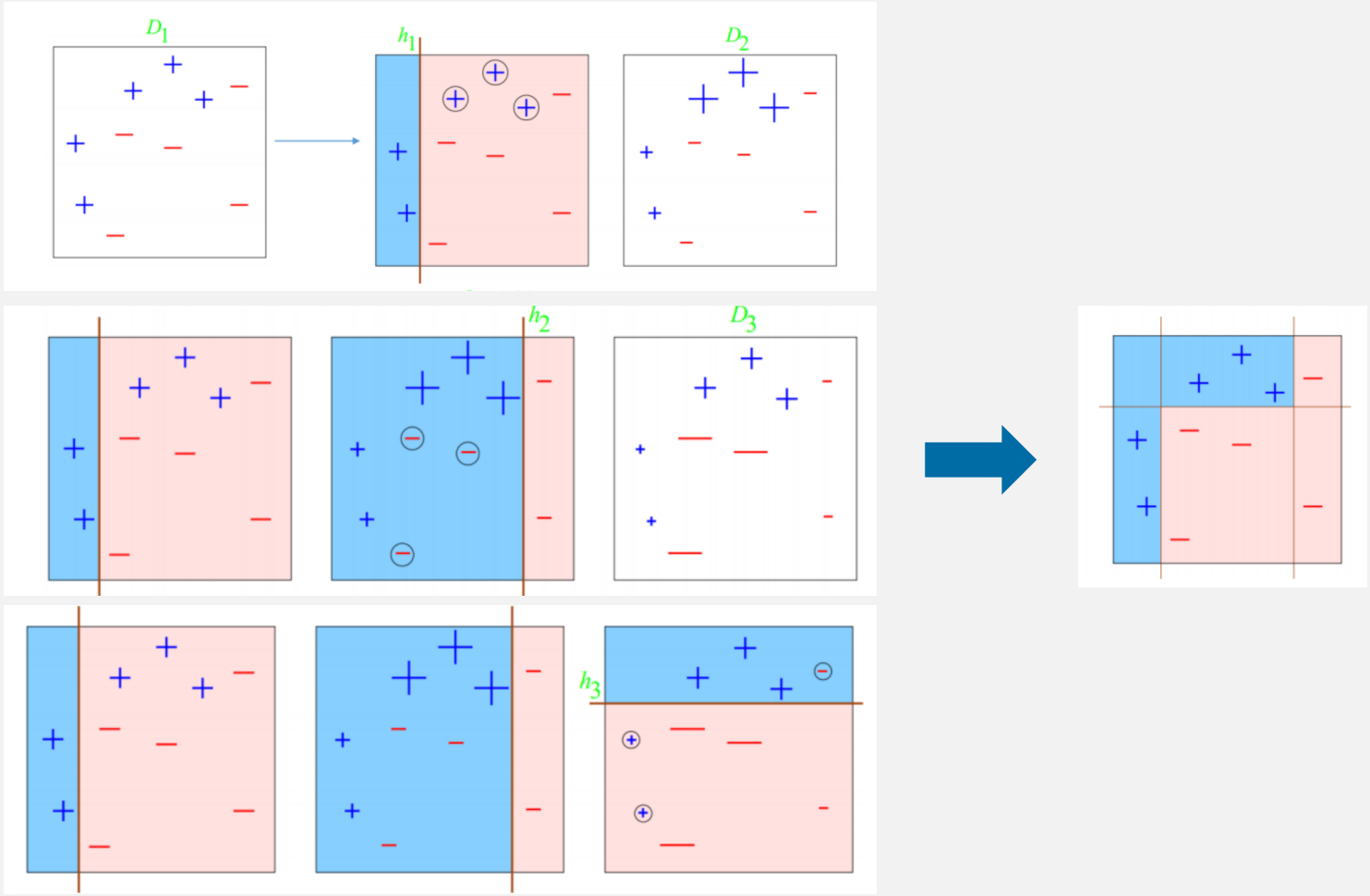


08 앙상블모델 – AdaBoost

AdaBoost 기본 정의 및 용어 정리

정의

- Adaboost : 약한 분류기를 상호 보완하여 순차적으로 학습 진행하고, 이들의 결과를 조합하여 최종적으로 강한 분류기의 성능을 향상



AdaBoost 모델의 장 & 단점

장점

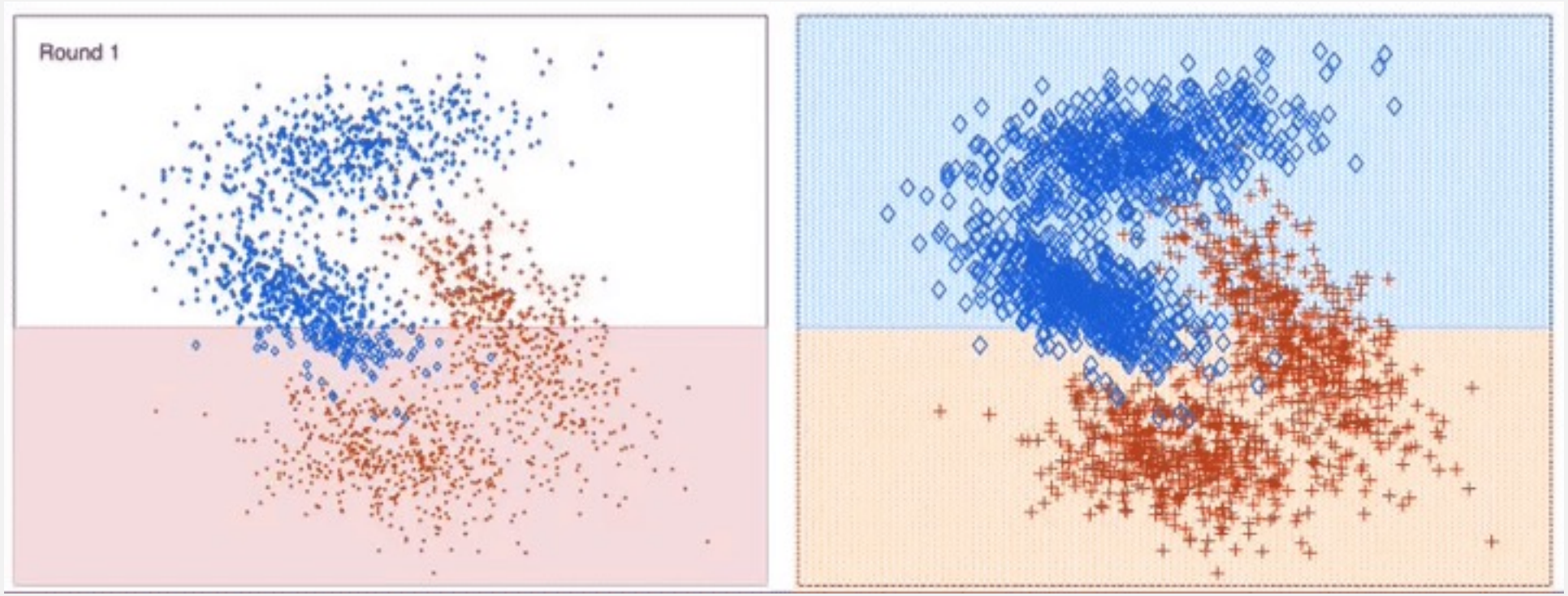
- 모델이 과적합 되는 부분을 학습 진행되면서 줄여줌
- Random Forest와 비교시, 이진분류에서 대체로 더 좋은 결과가 나왔음.

단점

- 노이즈 데이터 및 이상치에 민감하게 반응함.
- 모델 특성상 모든 트리 모델을 다 확인이 불가해서 해석이 어려움.

07 앙상블모델 – AdaBoost

AdaBoost 예시

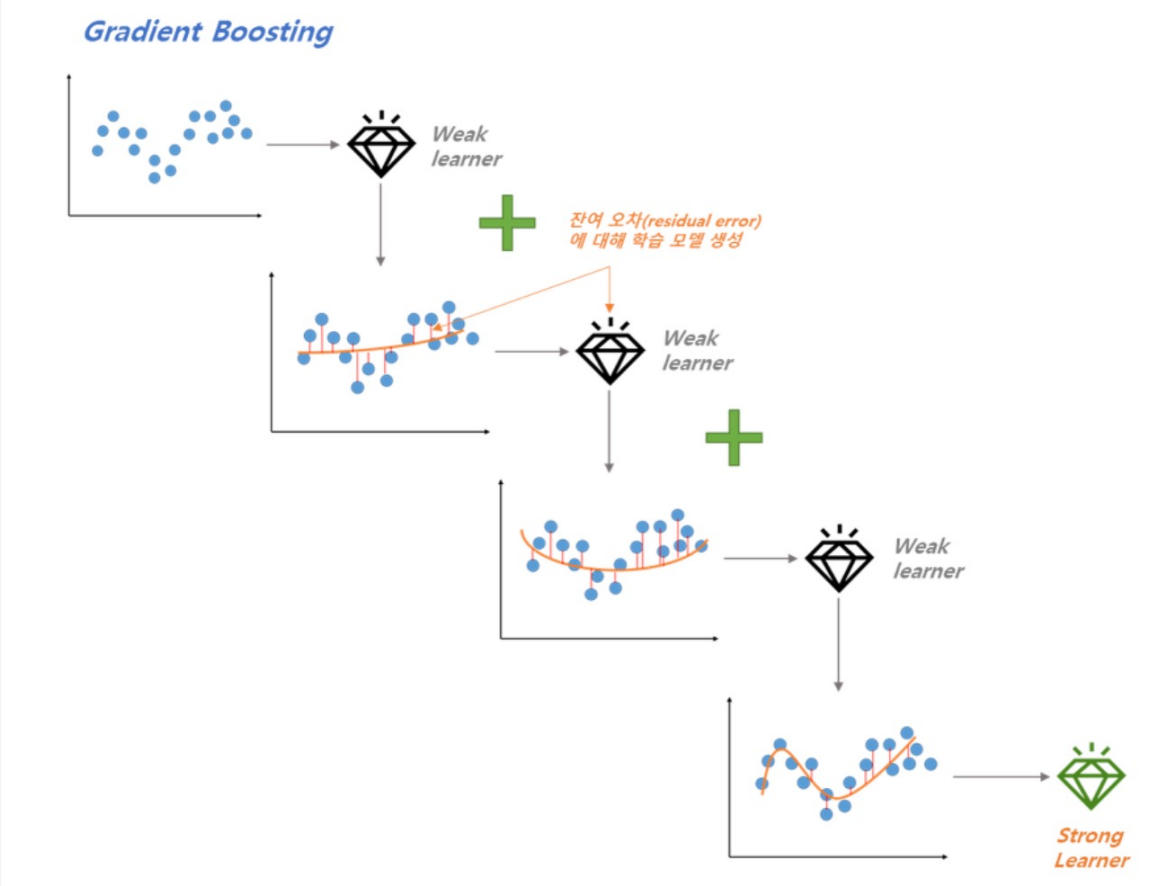


07 앙상블모델 – GBM (Gradient Boosting Machine)

GBM 기본 정의 및 용어 정리

정의

- GBM : 순차적으로 약한 학습기를 만들어가며 이전 학습기의 잔차(residual)을 보완하는 방식
- 가중치 업데이트에 경사하강법을 이용하여 오류를 개선해 나아가는 방법



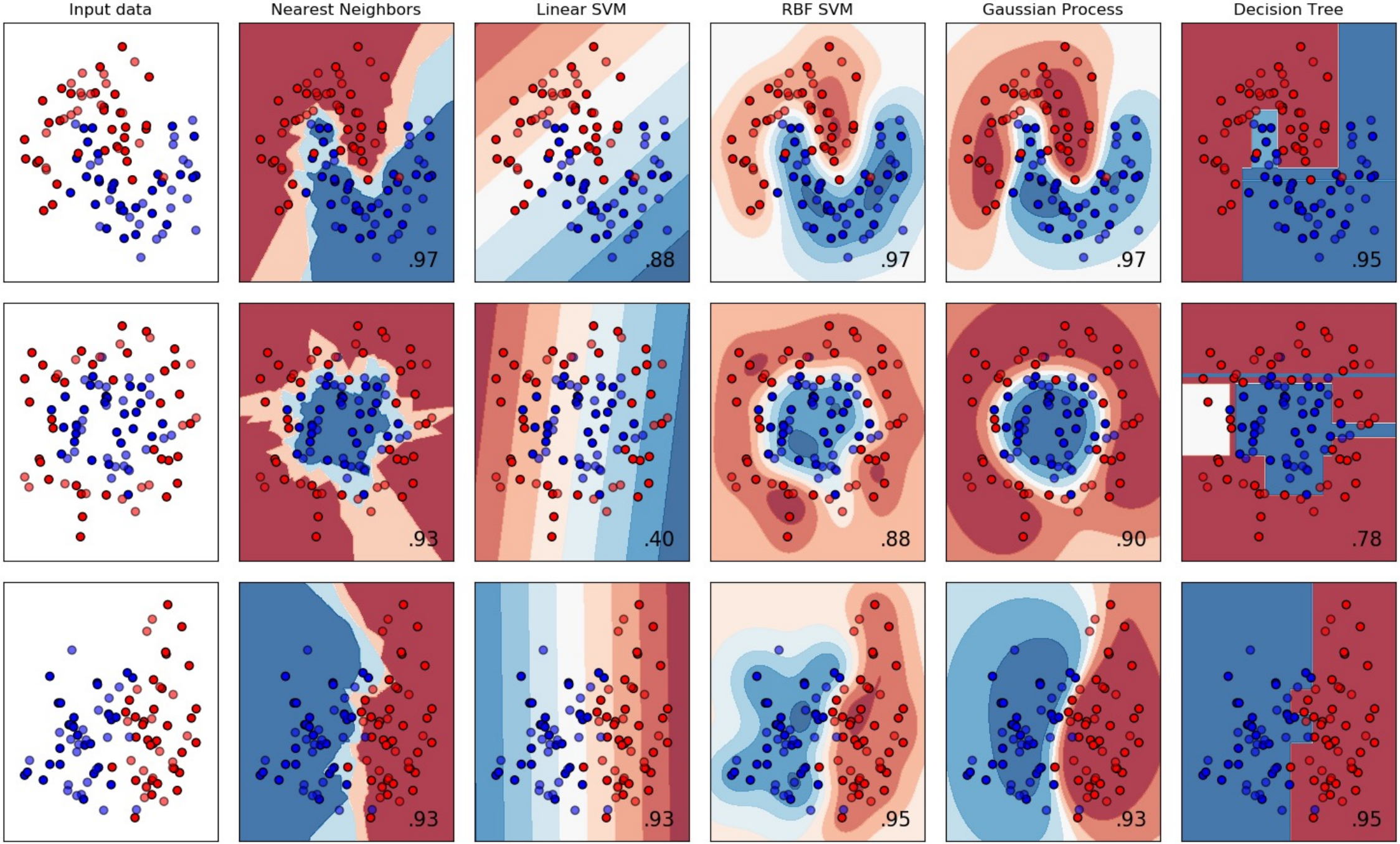
GBM 모델의 장 & 단점

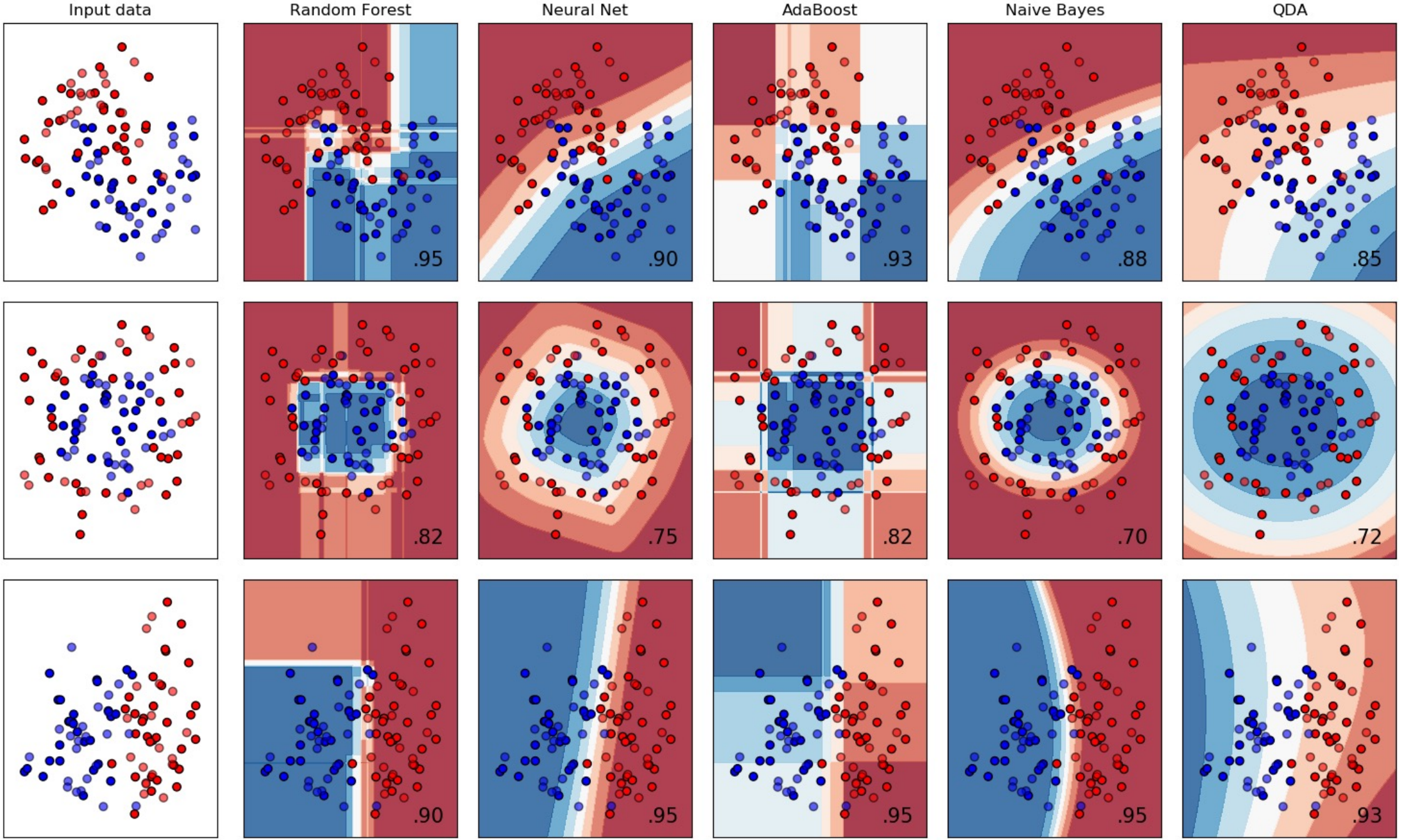
장점

- Random Forest보다 예측 성능이 대체로 뛰어남.
- 과적합 되더라도 테스트 데이터셋에 대해 강한 예측 성능을 가진 알고리즘
- GBM을 보완한 LightGBM, XGBoost 등의 모델이 있음.

단점

- 수행시간이 오래 걸리고, hyper-parameter tuning 노력이 필요.
  -> 약한 학습기의 순차적인 예측 오류 보정 과정으로 학습이 진행되기 때문에
  CPU 병렬 처리가 지원되지 않음


Input data
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
.97
.88
.97
.97
.95
.93
.40
.88
.90
.78
.93
.93
.95
.93
.95



08

비지도학습 – PCA

2) 비지도학습

데이터만 제공하고 정답을 제공하지 않은 상태로 학습을 진행한 알고리즘

PCA (Principal Component Analysis)

차원축소의 필요성

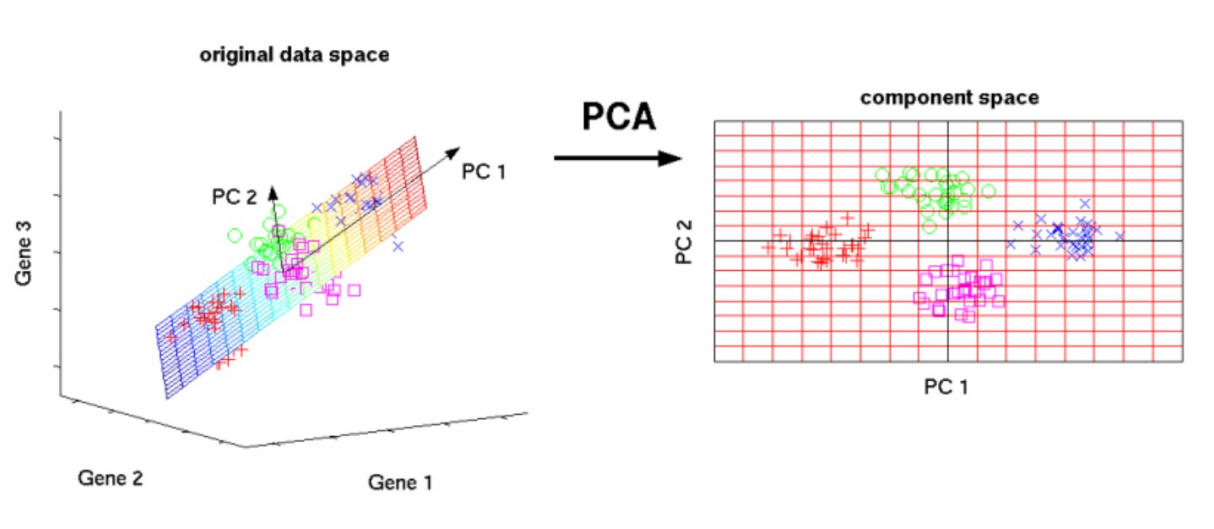
- 실제 데이터들의 대부분은 많은 설명 변수를 가지고 있어

  머신러닝 알고리즘을 적용해서 문제 해결하는 데 어려움 존재

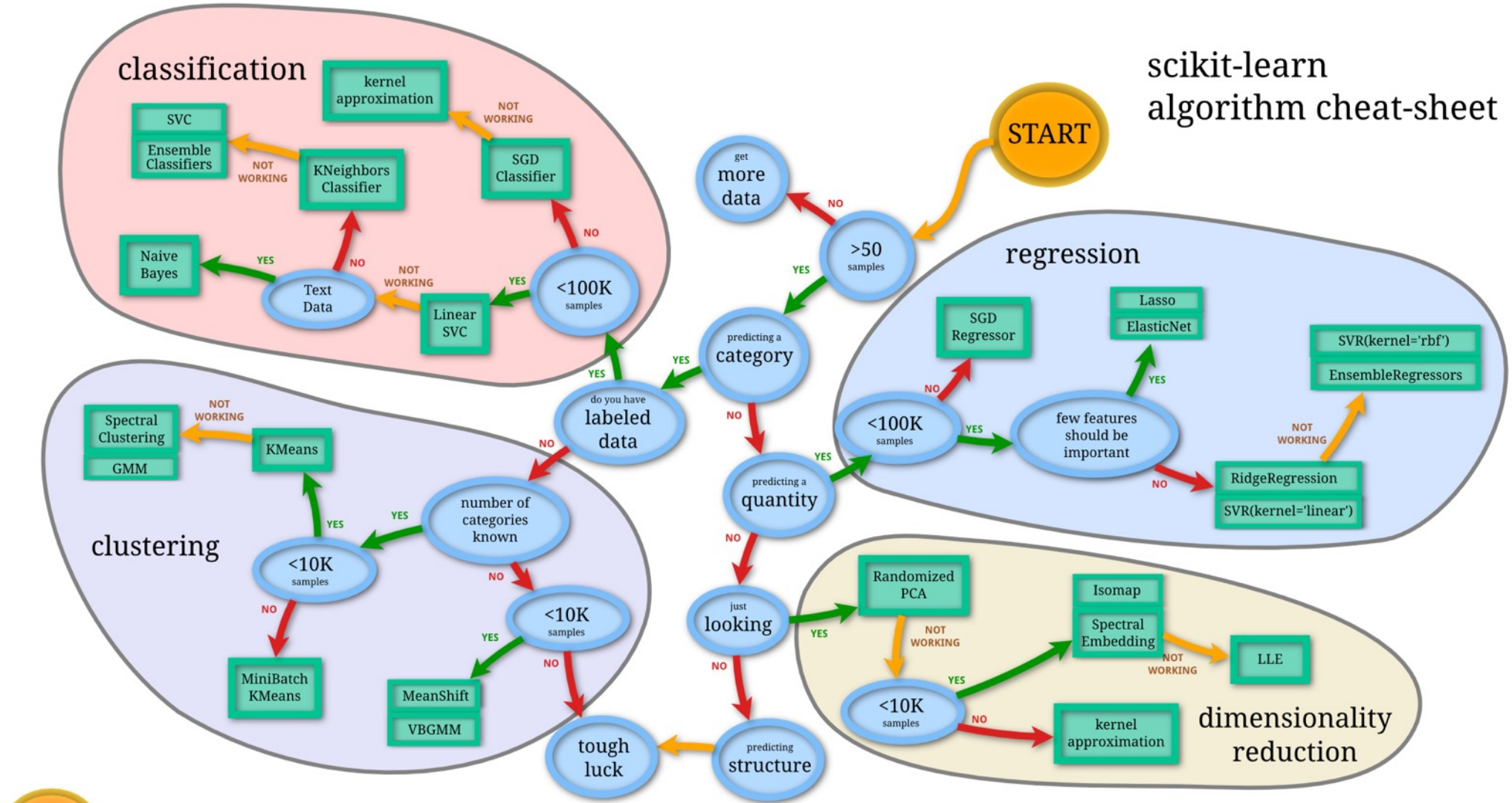
- => Feature Selection 또는 Dimensionality Reduction이 필수

PCA 정의

- 차원을 원데이터에서 줄이면서 정보의 손실을 최소화하는 방법
- 목표 : 더 적은 개수로 데이터를 충분히 잘 설명할 수 있는 축 설명



scikit-learn algorithm cheat-sheet



Thank you!

Contact Info.

OFFICE	경기도 용인시 기흥구 흥덕1로 13 흥덕IT밸리 타워A동 2904호
EMAIL	info@aidx.kr
PHONE #	031 8067 7767
FAX	031 8007 0761