

## 1 데이터분석 단계(Data Analysis Cycle)

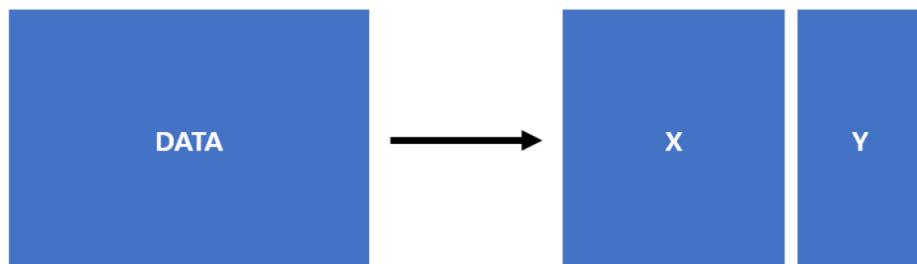
[Open in Colab](#)

✓ 데이터 전처리: (0) 쓸모 없을 뻔한 Raw를 쓸모 있는 Data로 변환

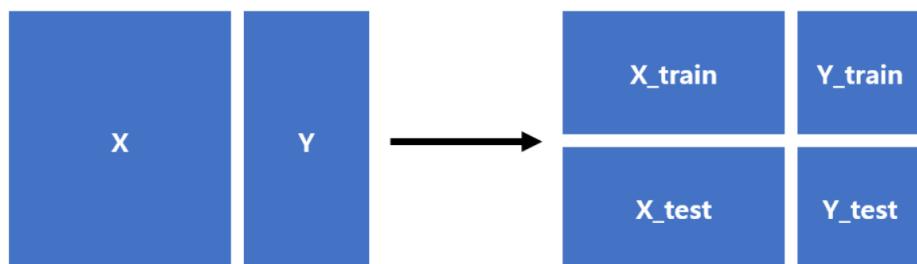
100	T50	횟수	111	TPU	...
few	Gds	Hvi	Rew	Fa	...
Fre	CT	QTP	D	합	...
'1'	1	23	22	NaN	43
76	NaN	43	32	1	8
'Hi'	NaN	NaN	NaN	NaN	87
23	98	NaN	64	46	NaN
c	90	'WW'	24	'KK'	4
t	NaN	2	NaN	NaN	6
64	NaN	90	'IU'	4	76

번호	시간	총량	기간	누적	...
1	1	23	22	21	43
76	33.3	43	32	1	8
5	33.3	52	35	21	87
23	98	52	64	46	61
90	33.3	2	24	33	4
55	2	52	35	21	6
64	33.3	90	11	4	76

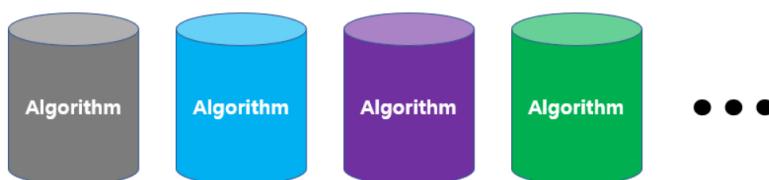
✓ 데이터 분할: (1) 목표/종속변수 Y와 설명/독립변수 X설정



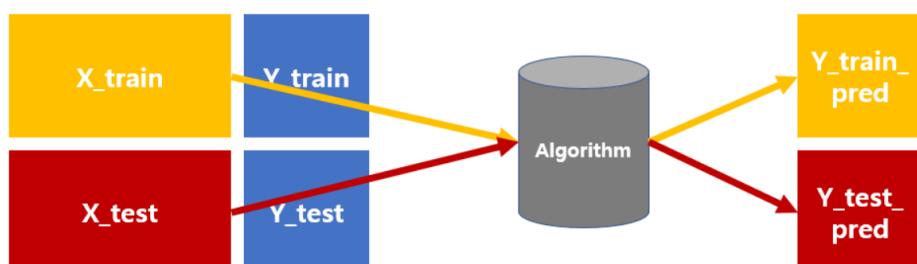
✓ 데이터 분할: (2) 학습데이터 Train과 예측 데이터 Test로 분할



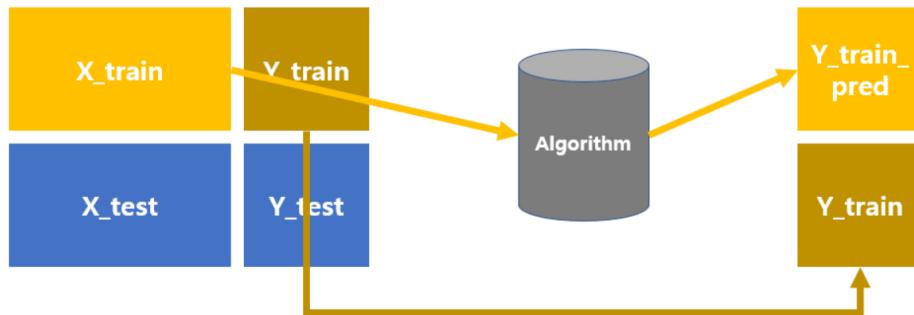
✓ 모델링: (3) 분석 목적에 맞는 알고리즘(Base & Advanced) 후보들 준비



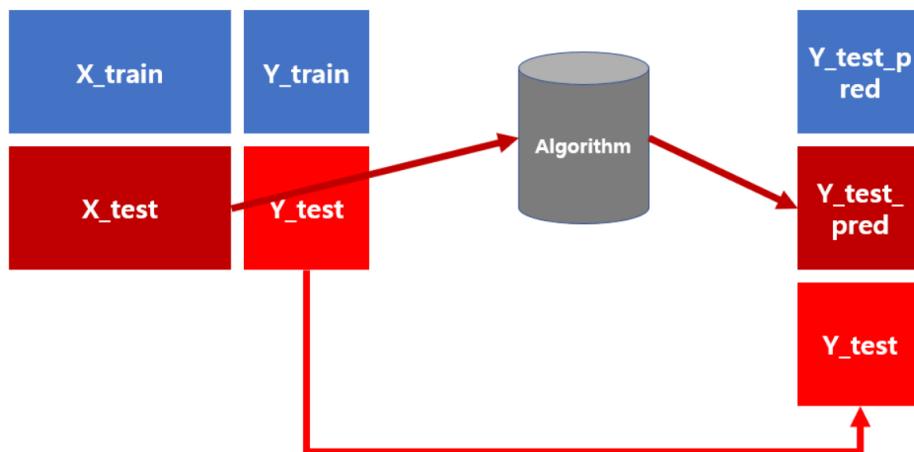
✓ 모델링 & 학습: (4) 알고리즘 평가를 위해 Train/Test의 예측값 추정



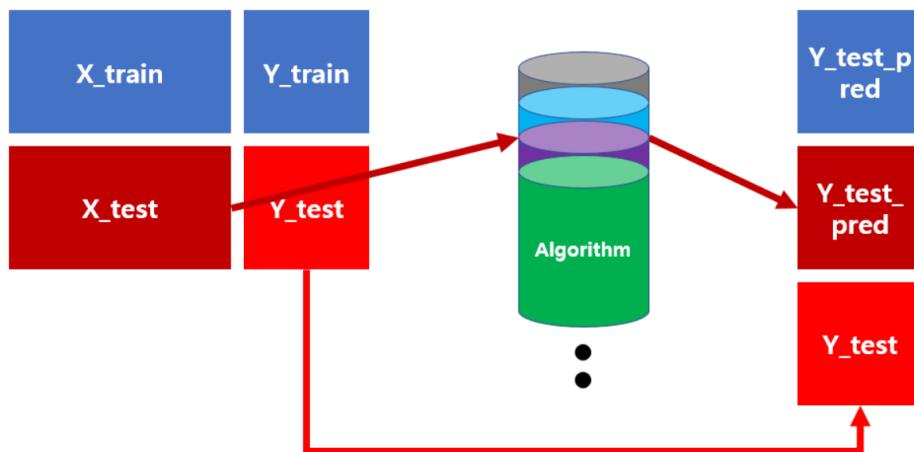
✓ 평가: (5) 학습(Train)이 잘 되었는지 알고리즘 성능검증



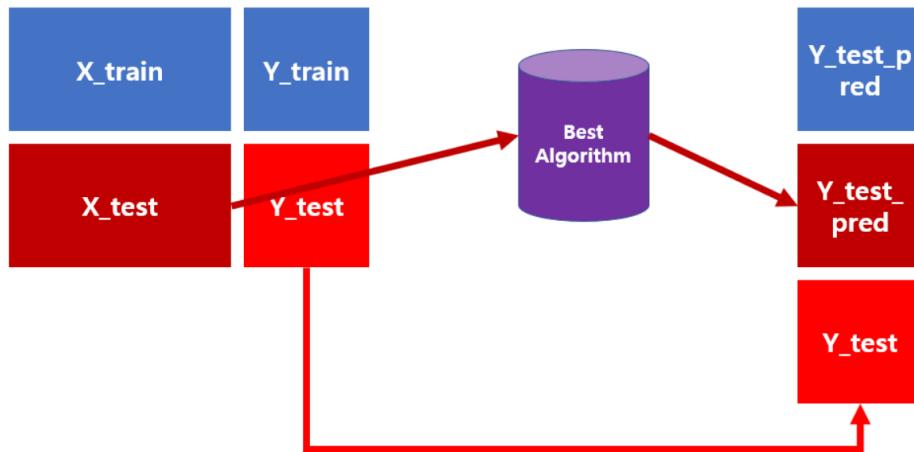
✓ 평가: (6) 예측(Test)이 잘 되었는지 알고리즘 성능검증



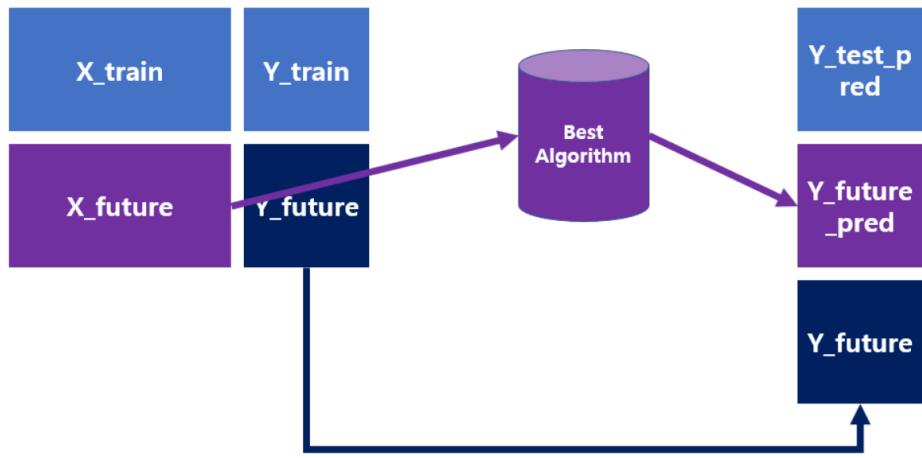
✓ 최적 알고리즘 선택: (7) 알고리즘을 변경하여 위 과정 반복 후



✓ 최적 알고리즘 선택: (7) 최고 성능의 알고리즘 선택

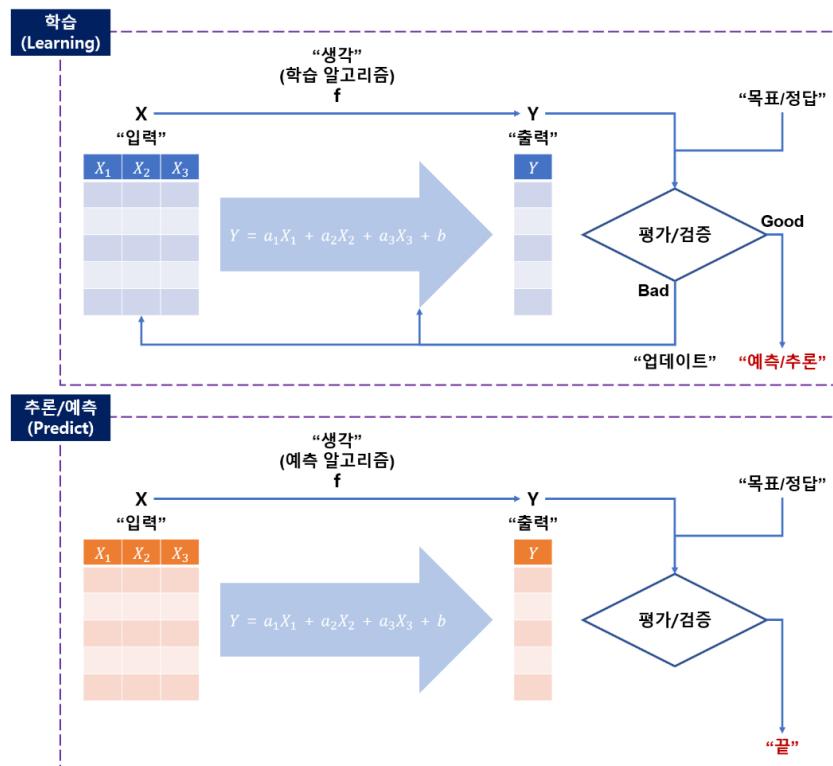


✓ 배포: (8) 실제 비즈니스 서비스 현업 적용 및 매출/수익/개선 정도 평가



## 2 지도학습(Supervised) 알고리즘: 분류분석

- 데이터분석 과정: 학습 + 추론/예측



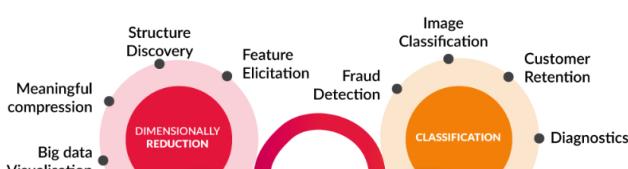
- 분류분석: 지도학습 알고리즘 중 분류를 위해 사용되는 가장 기본(Baseline) 알고리즘

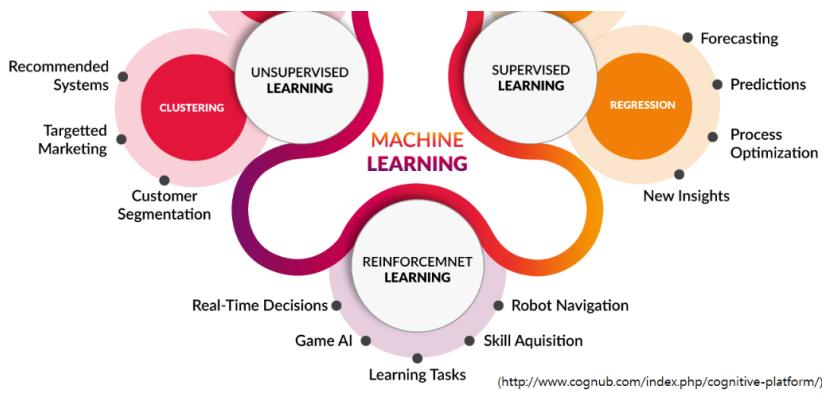
(비수학적) "일상 속 문제 중 여러개의 선택지 중에 정답을 고르는 문제"

- 주관식 시험문제의 숫자형 정답을 찾는 문제가 회귀문제
- 객관식 시험문제의 정답을 찾는 문제가 분류문제

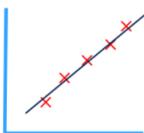
(수학적) "범주형 출력(  $Y$ , 종속변수 )에 영향을 주는 입력(  $X$ , 독립변수 )과의 관계를 정량적으로 추론/추정하여 미래 값을 분류 하는 알고리즘"

- 예측문제: 데이터 변수(Feature, Variable)들을 사용하여 연속적인(Continuous) 값을 예측
- 분류문제: 데이터 변수(Feature, Variable)들을 사용하여 특정 분류값을 예측



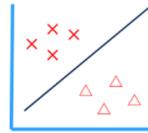


- How much is the stock of Samsung Electronics tomorrow?



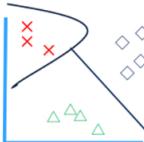
**Regression** – Looking for a statistical relationship across variables that may give us an estimate of a particular outcome. (Supervised)

- Will Samsung Electronics' stocks rise or fall tomorrow?



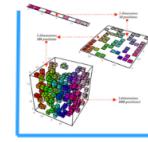
**Classification** – Similar to regression but looking for separations in the data given predefined classes. (Supervised)

- Are Samsung Electronics and Naver similar business companies?



**Clustering** – Do not have predefined classes but trying to find groups or sets based upon data at hand. (Unsupervised)

- What are the representatives among all stocks in the KOSPI?



**Dimensionality Reduction** – Transformation of data from high-dimensional into a low-dimensional space so that it retains some meaningful properties of the origin data. (Unsupervised)

Regression Algorithms	Instance-based Algorithms	Regularization Algorithms	Decision Tree Algorithms	Bayesian Algorithms	Artificial Neural Network Algorithms
Ordinary Least Squares Regression (OLSR)	k-Nearest Neighbor (kNN)	Ridge Regression	Classification and Regression Tree (CART)	Naive Bayes	Perceptron
Linear Regression	Learning Vector Quantization (LVQ)	Least Absolute Shrinkage and Selection Operator (LASSO)	Iterative Dichotomiser 3 (ID3)	Gaussian Naive Bayes	Back-Propagation
Logistic Regression	Self-Organizing Map (SOM)	Elastic Net	C4.5 and C5.0 (different versions of a powerful approach)	Multinomial Naive Bayes	Hopfield Network
Stepwise Regression	Locally Weighted Learning (LWL)	Least-Angle Regression (LARS)	Chi-squared Automatic Interaction Detection (CHAID)	Averaged One-Dependence Estimators (AODE)	Radial Basis Function Network (RBFN)
Multivariate Adaptive Regression Splines (MARS)	-	-	Decision Stump	Bayesian Belief Network (BBN)	-
Locally Estimated Scatterplot Smoothing (LOESS)	-	-	M5	Bayesian Network (BN)	-
-	-	-	Conditional Decision Trees	-	-

- Target Algorithm:

- Logistic Regression
- Ordinal Regression
- Cox Regression
- Naïve Bayes
- Stochastic Gradient Descent
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Support Vector Machine

### 3 예제 데이터셋(Dataset)

#### 3.1 statsmodels 모듈 사용 데이터셋

# 라이브러리 불러오기

```

import statsmodels.api as sm

• 대기중 CO2농도 데이터:
    data = sm.datasets.get_rdataset("CO2", package="datasets")

• 황체형성 호르몬(Luteinizing Hormone)의 수치 데이터:
    data = sm.datasets.get_rdataset("lh")

• 1974~1979년 사이의 영국의 호흡기 질환 사망자 수 데이터:
    data = sm.datasets.get_rdataset("deaths", "MASS")

• 1949~1960년 사이의 국제 항공 운송인원 데이터:
    data = sm.datasets.get_rdataset("AirPassengers")

• 미국의 강수량 데이터:
    data = sm.datasets.get_rdataset("precip")

• 타이타닉호의 탑승자들에 대한 데이터:
    data = sm.datasets.get_rdataset("Titanic", package="datasets")

```

- **data**가 포함하는 정보:

- **package**: 데이터를 제공하는 R 패키지 이름
- **title**: 데이터 이름
- **data**: 데이터를 담고 있는 데이터프레임
- **\_\_doc\_\_**: 데이터에 대한 설명 문자열(R 패키지의 내용 기준)

## 3.2 sklearn 모듈 사용 데이터셋

### 1) 패키지에 포함된 데이터( `load` 명령어 )

```

# 라이브러리 불러오기
from sklearn.datasets import load_boston

• load_boston: 회귀용 보스턴 집 값
    raw = load_boston()
    print(raw.DESCR)
    print(raw.keys())
    print(raw.data.shape, raw.target.shape)

• load_diabetes: 회귀용 당뇨병 자료
• load_linnerud: 회귀용 linnerud 자료
• load_iris: 분류용 붓꽃(iris) 자료
• load_digits: 분류용 숫자(digit) 필기 이미지 자료
• load_wine: 분류용 포도주(wine) 등급 자료
• load_breast_cancer: 분류용 유방암(breast cancer) 진단 자료

```

### 2) 인터넷에서 다운로드할 수 있는 데이터( `fetch` 명령어 )

```

# 라이브러리 불러오기
from sklearn.datasets import fetch_california_housing

• fetch_california_housing: 회귀용 캘리포니아 집 값
    raw = fetch_california_housing()
    print(raw.DESCR)
    print(raw.keys())
    print(raw.data.shape, raw.target.shape)

• fetch_covtype: 회귀용 토지 조사 자료
• fetch_20newsgroups: 뉴스 그룹 텍스트 자료
• fetch_olivetti_faces: 얼굴 이미지 자료
• fetch_lfw_people: 유명인 얼굴 이미지 자료
• fetch_lfw_pairs: 유명인 얼굴 이미지 자료
• fetch_rcv1: 로이터 뉴스 말뭉치
• fetch_kddcup99: Kddcup 99 Tcp dump 자료

```

### 3) 확률분포를 사용한 가상 데이터( `make` 명령어 )

```

# 라이브러리 불러오기
from sklearn.datasets import make_regression

• make_regression: 회귀용 가상 데이터
    X, y, c = make_regression(n_samples=100, n_features=10, n_targets=1, bias=0, noise=0, coef=True, random_state=0)

• make_classification: 분류용 가상 데이터 생성
• make_blobs: 클러스터링용 가상 데이터 생성

```

#### 4) load/fetch 명령어 데이터에서 raw가 포함하는 정보: Bunch라는 클래스 객체 형식으로 생성

- data:(필수) 독립 변수 ndarray 배열
- target:(필수) 종속 변수 ndarray 배열
- feature\_names :(옵션) 독립 변수 이름 리스트
- target\_names :(옵션) 종속 변수 이름 리스트
- DESCR:(옵션) 자료에 대한 설명

### 3.3 분류문제 예시

```
In [1]: # Classification
import statsmodels.api as sm
data = sm.datasets.get_rdataset("Titanic", package="datasets")
print(data.title)
print(data.__doc__)
print(data.keys())
display(data.data)
executed in 3.08s, finished 21:49:07 2022-06-14

Survival of passengers on the Titanic
.. container::

=====
Titanic R Documentation
=====

.. rubric:: Survival of passengers on the Titanic
:name: survival-of-passengers-on-the-titanic

.. rubric:: Description
:name: description

This data set provides information on the fate of passengers on the
fatal maiden voyage of the ocean liner 'Titanic', summarized
according to economic status (class), sex, age and survival.

.. rubric:: Usage
:name: usage

::

Titanic

.. rubric:: Format
:name: format

A 4-dimensional array resulting from cross-tabulating 2201
observations on 4 variables. The variables and their levels are as
follows:

== =====
No Name      Levels
1 Class      1st, 2nd, 3rd, Crew
2 Sex        Male, Female
3 Age        Child, Adult
4 Survived   No, Yes
== =====

.. rubric:: Details
:name: details

The sinking of the Titanic is a famous event, and new books are still
being published about it. Many well-known facts—from the proportions
of first-class passengers to the 'women and children first' policy,
and the fact that that policy was not entirely successful in saving
the women and children in the third class—are reflected in the
survival rates for various classes of passenger.

These data were originally collected by the British Board of Trade in
their investigation of the sinking. Note that there is not complete
agreement among primary sources as to the exact numbers on board,
rescued, or lost.

Due in particular to the very successful film 'Titanic', the last
years saw a rise in public interest in the Titanic. Very detailed
data about the passengers is now available on the Internet, at sites
such as *Encyclopedia Titanica*
(https://www.encyclopedia-titanica.org/).

.. rubric:: Source
:name: source

Dawson, Robert J. MacG. (1995). The 'Unusual Episode' Data Revisited.
*Journal of Statistics Education*, **3**.
doi:10.1080/10691898.1995.11910499 <https://doi.org/10.1080/10691898.1995.11910499>__.

The source provides a data set recording class, sex, age, and
survival status for each person on board of the Titanic, and is based
on data originally collected by the British Board of Trade and
reprinted in:

British Board of Trade (1990), *Report on the Loss of the 'Titanic'
```

(S.S.)\*. British Board of Trade Inquiry Report (reprint). Gloucester,  
UK: Allan Sutton Publishing.

```
.. rubric:: Examples
:name: examples

::

require(graphics)
mosaicplot(Titanic, main = "Survival on the Titanic")
## Higher survival rates in children?
apply(Titanic, c(3, 4), sum)
## Higher survival rates in females?
apply(Titanic, c(2, 4), sum)
## Use loglm() in package 'MASS' for further analysis ...
```

```
dict_keys(['data', '__doc__', 'package', 'title', 'from_cache'])
```

	Class	Sex	Age	Survived	Freq
0	1st	Male	Child	No	0
1	2nd	Male	Child	No	0
2	3rd	Male	Child	No	35
3	Crew	Male	Child	No	0
4	1st	Female	Child	No	0
5	2nd	Female	Child	No	0
6	3rd	Female	Child	No	17
7	Crew	Female	Child	No	0
8	1st	Male	Adult	No	118
9	2nd	Male	Adult	No	154
10	3rd	Male	Adult	No	387
11	Crew	Male	Adult	No	670
12	1st	Female	Adult	No	4
13	2nd	Female	Adult	No	13
14	3rd	Female	Adult	No	89
15	Crew	Female	Adult	No	3
16	1st	Male	Child	Yes	5
17	2nd	Male	Child	Yes	11
18	3rd	Male	Child	Yes	13
19	Crew	Male	Child	Yes	0
20	1st	Female	Child	Yes	1
21	2nd	Female	Child	Yes	13
22	3rd	Female	Child	Yes	14
23	Crew	Female	Child	Yes	0
24	1st	Male	Adult	Yes	57
25	2nd	Male	Adult	Yes	14
26	3rd	Male	Adult	Yes	75
27	Crew	Male	Adult	Yes	192
28	1st	Female	Adult	Yes	140
29	2nd	Female	Adult	Yes	80
30	3rd	Female	Adult	Yes	76
31	Crew	Female	Adult	Yes	20

```
In [2]: # Classification
from sklearn.datasets import load_breast_cancer
raw = load_breast_cancer()
print(raw.DESCR)
print(raw.keys())
print(raw.data.shape, raw.target.shape)
```

executed in 103ms, finished 21:49:07 2022-06-14

```
... _breast_cancer_dataset:
```

Breast cancer wisconsin (diagnostic) dataset

-----

\*\*Data Set Characteristics:\*\*

:Number of Instances: 569

:Number of Attributes: 30 numeric, predictive attributes and the class

:Attribute Information:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness ( $\text{perimeter}^2 / \text{area} - 1.0$ )
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry

- fractal dimension ("coastline approximation" = 1)

The mean, standard error, and "worst" or largest (mean of the three worst/largest values) of these features were computed for each image, resulting in 30 features. For instance, field 0 is Mean Radius, field 10 is Radius SE, field 20 is Worst Radius.

- class:

- WDBC-Malignant
- WDBC-Benign

:Summary Statistics:

	Min	Max
radius (mean):	6.981	28.11
texture (mean):	9.71	39.28
perimeter (mean):	43.79	188.5
area (mean):	143.5	2501.0
smoothness (mean):	0.053	0.163
compactness (mean):	0.019	0.345
concavity (mean):	0.0	0.427
concave points (mean):	0.0	0.201
symmetry (mean):	0.106	0.304
fractal dimension (mean):	0.05	0.097
radius (standard error):	0.112	2.873
texture (standard error):	0.36	4.885
perimeter (standard error):	0.757	21.98
area (standard error):	6.802	542.2
smoothness (standard error):	0.002	0.031
compactness (standard error):	0.002	0.135
concavity (standard error):	0.0	0.396
concave points (standard error):	0.0	0.053
symmetry (standard error):	0.008	0.079
fractal dimension (standard error):	0.001	0.03
radius (worst):	7.93	36.04
texture (worst):	12.02	49.54
perimeter (worst):	50.41	251.2
area (worst):	185.2	4254.0
smoothness (worst):	0.071	0.223
compactness (worst):	0.027	1.058
concavity (worst):	0.0	1.252
concave points (worst):	0.0	0.291
symmetry (worst):	0.156	0.664
fractal dimension (worst):	0.055	0.208

:Missing Attribute Values: None

:Class Distribution: 212 - Malignant, 357 - Benign

:Creator: Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian

:Donor: Nick Street

:Date: November, 1995

This is a copy of UCI ML Breast Cancer Wisconsin (Diagnostic) datasets.  
<https://goo.gl/U2Uwz2>

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming," Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:

```
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/
```

.. topic:: References

- W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
  - O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
  - W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.
- dict.keys(['data', 'target', 'frame', 'target\_names', 'DESCR', 'feature\_names', 'filename', 'data\_module'])  
(569, 30) (569,)

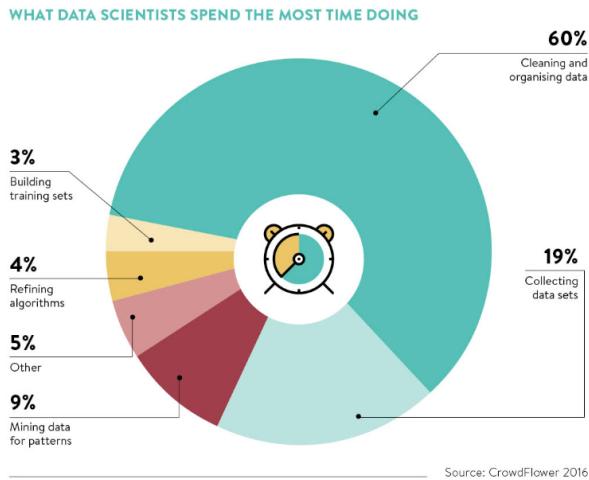
```
In [3]: # Classification
from sklearn.datasets import make_classification
X, y = make_classification(n_samples=100, n_features=10, n_classes=2,
                           n_informative=5, n_redundant=0, random_state=0)
print(X.shape, y.shape)

executed in 13ms, finished 21:49:07 2022-06-14
(100, 10) (100,)
```

## 4 전처리 방향(Preprocessing)

- 목표:

- 대량으로 수집된 데이터는 그대로 활용 어려움
- 잘못 수집/처리 된 데이터는 엉뚱한 결과를 발생
- 알고리즘이 학습이 가능한 형태로 데이터를 정리



Source: CrowdFlower 2016

### 일반적인 전처리 필요항목:

- 데이터 결합
- 결측값 처리
- 이상치 처리
- 자료형 변환
- 데이터 분리
- 데이터 변환
- 스케일 조정

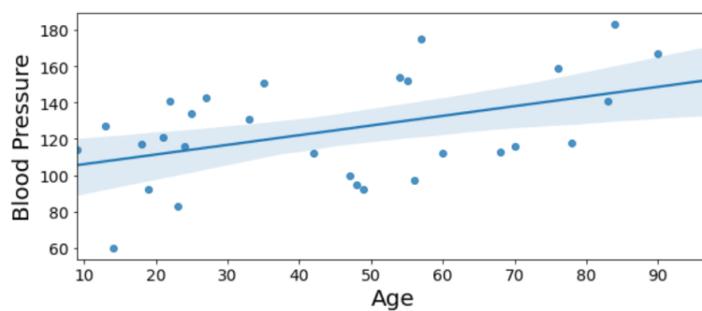
## 5 함수세팅 및 추정 방향(Modeling): Logistic Regression

### 5.1 분류문제에 회귀분석 사용시 한계 및 대응

#### 1) 회귀문제에 회귀분석은 적절

##### "연속형 종속변수 예시"

	Age	Blood Pressure
0	33	131
1	18	117
2	57	175
3	70	116
4	60	112

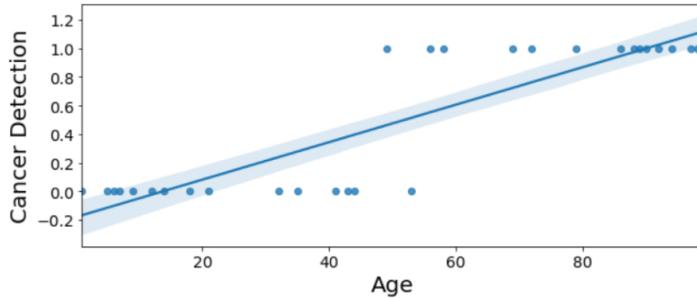


## 2) 눈듀눈세에 외귀눈식은 누석설

### "범주형 종속변수 예시"

- Outlier가 존재하면 Linear Regression의 추정은 왜곡을 발생시킴

Age	Cancer Detection
0	7
1	35
2	12
3	53
4	98



### 3) 분류문제 해결을 위한 대응:

**회귀분석:** 연속형 종속변수  $Y$ 의 값을 추론

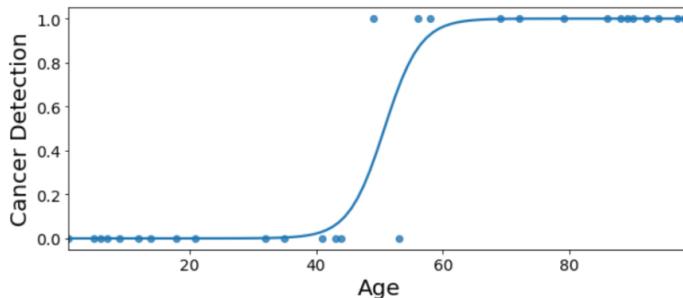
- 혈압의 경우 값 자체로 의미가 있지만 암발생은 발생(1)과 미발생(0) 사이의 중간값 무의미
  - 회귀분석으로 범주형 종속변수를 추론하면 범위가 맞지 않아 암발생 여부의 해석이 왜곡

- 나이가 많아지면 무조건 암이 걸리거나 나이가 어리면 무조건 암이 걸리지 않는 왜곡 발생
  - 범주형 Y일 경우 회귀분석의 한계 존재하여 이를 해결하기 위한 접근 필요

**분류분석:** 범주형(범주/카테고리/클래스/라벨) 종속변수  $Y$ 의 분류를 추론

- 시그모이드 함수(Sigmoid Function) 적용: 연속형 Y를 0과 1사이의 값으로 변환하면서 분류에 맞게 S자 형태로 적합(Fitting)하는 함수

Age	Cancer Detection
0	0
1	0
2	0
3	0
4	1



4) 분류분석 종류: Y 카테고리 갯수와 방향에 따라 Binary/Multi-class/Multi-label로 구분

- **Binary Classification:** 데이터가 2개의 카테고리 중 어떤 것인지 추론하는 문제 (ex. 성별 추론 / 스팸메일 추론)
  - **Multi-class Classification:** 데이터가 2개 이상의 카테고리 중 어떤 것인지 추론하는 문제 (ex. 사진으로 동물 이름 추론)
  - **Multi-label Classification:** 데이터가 2개 이상의 카테고리 중 어떠한 것들인지 추론하는 문제 (ex. 뉴스기사는 스포츠/사람/지역 관련임을 추론)

## 5.2 분류문제 해결을 위한 가설 및 비용함수

"선형회귀분석을 포함하여 머신러닝과 딥러닝 등의 모든 알고리즘은 큰 틀에서 작동방식이 동일"

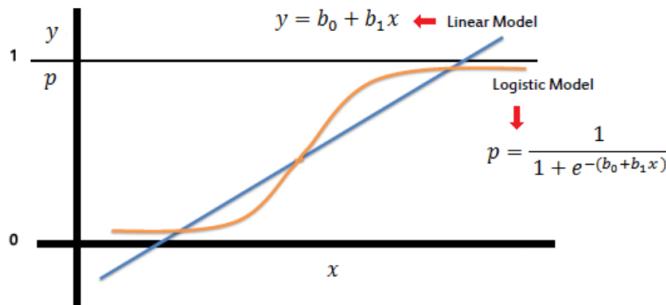
"머신러닝과 딥러닝의 작동방식을 이해하기 위해 가장 기초 예측 알고리즘인 선형회귀분석(Linear Regression) 작동방식부터"

"서현회귀분석을 포함한 대부분의 알고리즘은 큰 틀에서 3가지 도구를 사용하여 작동"

- 1) 방정식 (Equation) = 함수 (Function) = 가설 (Hypothesis)

## 1-1) 알고리즘 함수세팅: 분류문제를 푸는 대표적인 알고리즘 Logistic Regression

- 범주형 종속변수의 적합/추정하기 위한 변환과정 필요
- Logistic/Sigmoid Function를 사용하여 곡선(S-curve) 형태로 변환



## (1) 회귀분석 추정:

$$\begin{aligned} Y \approx \hat{Y} &= f(X_1, X_2, \dots, X_k) \\ &= w_0 + w_1 X_1 + w_2 X_2 + \dots + w_k X_k \\ &= XW \end{aligned}$$

## (2) 시그모이드 변환(Logistic/Sigmoid Transformation): Binary Classification 반영하는 곡선 형태로 변경

$$\begin{aligned} Pr(\hat{Y}) &= \frac{1}{1 + \exp(-\hat{Y})} \\ &= \frac{1}{1 + \exp(-XW)} \\ &= \frac{\exp(XW)}{1 + \exp(XW)} \end{aligned}$$

## (3) 로짓 변환(Logit Transformation): x의 선형관계 형태로 변환하여 변수들로 Y=1인 확률 추정

$$\begin{aligned} Pr(\hat{Y})(1 + \exp(XW)) &= \exp(XW) \\ Pr(\hat{Y}) &= (1 - Pr(\hat{Y})) \exp(XW) \\ \text{Odds(ratio)}: \left( \frac{Pr(\hat{Y})}{1 - Pr(\hat{Y})} \right) &= \exp(XW) \\ \text{Logit(log-odds)}: \log \left( \frac{Pr(\hat{Y})}{1 - Pr(\hat{Y})} \right) &= XW = w_0 + w_1 X_1 + w_2 X_2 + \dots + w_k X_k \end{aligned}$$

## 1-2) 추정 결과 해석:

(1) 해석 방향:  $\hat{Logit}$ 과  $\hat{Odds}$  변환으로 가능

$$\text{Logit: } \log \left( \frac{Pr(Y)}{1 - Pr(Y)} \right) = X\hat{W} = \hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2 + \dots + \hat{w}_k X_k$$

$$\text{Odds: } \left( \frac{Pr(Y)}{1 - Pr(Y)} \right) = \exp(X\hat{W}) = \exp(\hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2 + \dots + \hat{w}_k X_k)$$

(2) 회귀분석과 달리 Y의 로짓변환 값을 x의 선형관계로 추정하기 때문에, 해석시 Odds로 변환해서 해석해야 하므로 주의

$$\left( \frac{Pr(Y)}{1 - Pr(Y)} \right) = \exp(0.01 + 0.8X_1)$$

- 선형회귀분석:  $X_1$ 이 1만큼 증가하면  $Y$ 는  $w_1$ 만큼 증가

:  $X_1$ 이 1만큼 증가하면  $Y$ 는 0.8만큼 증가

- 로지스틱회귀분석:  $X_1$ 이 1만큼 증가하면  $\left( \frac{Pr(\hat{Y})}{1 - Pr(\hat{Y})} \right)$  범주변화는  $\exp(w_1)$ 만큼 증가

:  $X_1$ 이 1만큼 증가하면 암에 걸리지 않을 확률보다 암에 걸릴 확률이  $\exp(0.8) = 2.23$  배 더 높음

## (3) Y 확률 예측: 추정된 계수의 합수를 로지스틱 변환으로 출력

$$Pr(\hat{Y}) = \frac{1}{1 + \exp(-X\hat{W})} = \frac{\exp(X\hat{W})}{1 + \exp(X\hat{W})}$$

## (4) 분류 의사결정: 기본 임계값은 0.5로 Y 확률 예측 값이 0.5 이상이면 1, 0.5 미만이면 0으로 분류

$$\hat{v} = \int 1 \text{ if } Pr(\hat{Y}) \geq 0.5$$

$$t = \begin{cases} 0 & \text{if } Pr(\hat{Y}) < 0.5 \\ 1 & \text{otherwise} \end{cases}$$

2) 합수 추정을 위한 비용함수: 나의 주장 기반 알고리즘의 분류값 ( $Pr(\hat{Y})$ )과 실제 데이터 ( $Y$ )의 차이를 평가하는 함수

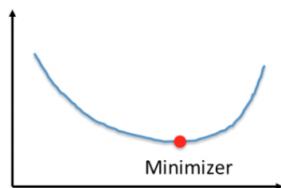
- 이슈: 잔차를 사용하는 Linear Regression 비용함수 적용 어려움

(1) 분류문제에서는  $\hat{Y}$ 를 사용한 잔차(에러)계산이 무의미

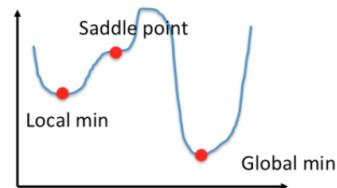
(2) 잔차( $Y - \hat{Y}$ )를 시그모이드 및 로짓 변환을 하면 Non-convex 형태가 되서 최소값(Global Minimum) 추정 어려움

(3) 정확한 수학적 방정식 기반 계수추정 어렵기에 확률론적 접근 필요

### Convex



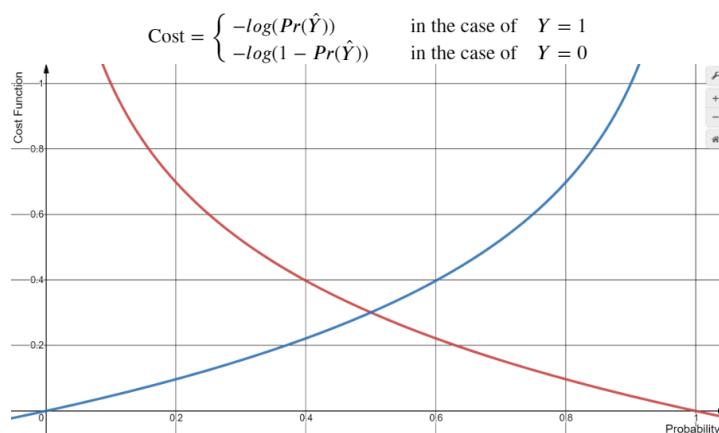
### Non-Convex



- 방향: 회귀문제와 달리 새로운 비용함수 가 필요

•  $Y$ 를 잘 분류하면  $cost=0$  으로 그렇지 않으면  $cost=\infty$  가 되는 방향

- (빨간선) 실제값이 1 일 때 예측값이 1 이면 Cost는 0
- (빨간선) 실제값이 1 일 때 예측값이 0 이면 Cost는 무한대



- Cross Entropy 등장:  $Y$ 가 0과 1인 경우의 Cost를 결합하여 하나의 식으로 표현

- 로지스틱 알고리즘은 비용함수로 Cross Entropy를 사용하고 최소로 하는 계수/가중치 추정
- $Y=0$  인 경우 파란부분만 남고  $Y=1$  인 경우 빨간부분만 남아 Class별로 독립적으로 작동
- 분류문제의 Cost 함수는 다양하고 많지만 통계학적으로 Cross Entropy는 계수 추정에 효율적 인 편
- Convex 형태이기 때문에 Global Minimum을 찾기가 용이함
- 추정된 계수/가중치( $\hat{w}$ )로 방정식을 만들어  $Y=1$ 인 분류확률을 계산 가능

$$\text{Cost} = \sum_{i=1}^m [-\hat{Y}_i \log(Pr(\hat{Y}_i)) - (1 - \hat{Y}_i) \log(1 - Pr(\hat{Y}_i))]$$

$$\hat{W} = \arg \min_W \sum_{i=1}^m [\text{Cost}]$$

## 5.3 확률론적 모형(Probabilistic Model): 통계적 모형

"증속변수의 발생가능성을 최대(최소)로하는  $W$ 를 추정"

- 범주형 분류문제를 확률로 반영하였기 때문에 확률론적 방식으로 접근

- 1) 실제  $Y$ 값의 추정가능성(Likelihood):

$$Pr(Y_i | X_i) = \prod_{i=1}^m Pr(Y_i)^{Y_i} [1 - Pr(Y_i)]^{1-Y_i}$$

- 2) 추정가능성의 더하기 표시 변환을 위한 Log함수 적용(Log-Likelihood):

$$\begin{aligned} LL &= \log Pr(Y_i | X_i) \\ &= \sum_{i=1}^m [Y_i \log(Pr(Y_i)) + (1 - Y_i) \log(1 - Pr(Y_i))] \end{aligned}$$

3) Log-Likelihood의 음수화 및 그레디언트가 최소가 되는  $\hat{\theta}$ :

$$-\frac{d}{dW} LL = \text{minimum}$$

4) 수치해석 방법론으로 초기값  $W$ 의 반복적 업데이트를 통한 최종  $\hat{W}$  추정:

$$\hat{W}^{new} = \hat{W}^{old} - (\frac{d^2}{dW dW^T} LL)^{-1} \frac{d}{dW} LL$$

(1) 계수에 임의의 초기값을 반영하여 Logit 추정

$$\text{Logit: } \log \left( \frac{Pr(Y)}{1 - Pr(Y)} \right) = X\hat{W} = \hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2 + \cdots + \hat{w}_k X_k$$

(2) 추정된 Logit으로 Likelihood기반 비용함수 계산

$$\begin{aligned} \text{LL} &= \log Pr(Y_i \mid X_i) \\ &= \sum_{i=1}^m [Y_i \log(Pr(Y_i)) + (1 - Y_i) \log(1 - Pr(Y_i))] \end{aligned}$$

(3) 비용함수를 감소시키는 방향으로  $\hat{W}$ 를 업데이트하여 최적화

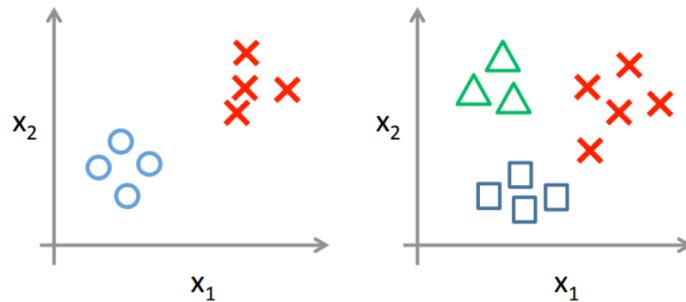
$$\hat{W}^{new} = \hat{W}^{old} - (\frac{d^2}{dWdW^T} LL)^{-1} \frac{d}{dW} LL$$

## 5.4 Multi-class 분류문제

- Binary vs. Multi-class:

## Binary classification:

## Multi-class classification:

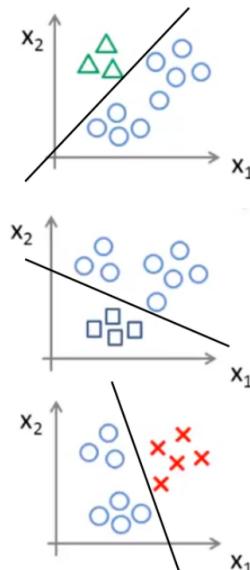
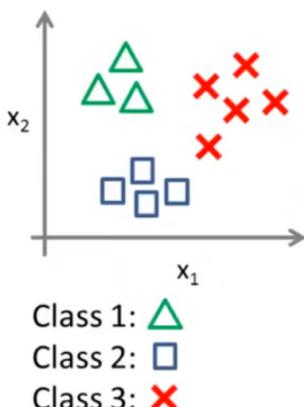


- **방향:** N개의 Class(Category)가 있는 문제는 N개의 Binary Classification으로 바꾸어 해결

- (1) 세모가 Positive일때 Y가 세모에 속할 확률
  - (2) 네모가 Positive일때 Y가 네모에 속할 확률
  - (3) 엑스가 Positive일때 Y가 엑스에 속할 확률

"데이터가 주어지면 3개의 경우를 모두 적용하여 최대 확률을 갖는 Class로 추정"

## One-vs-all (one-vs-rest):



## 6 검증지표 방향(Evaluation Metrics)

### < 1단계 >



### < 2단계 >



### < 3단계 >

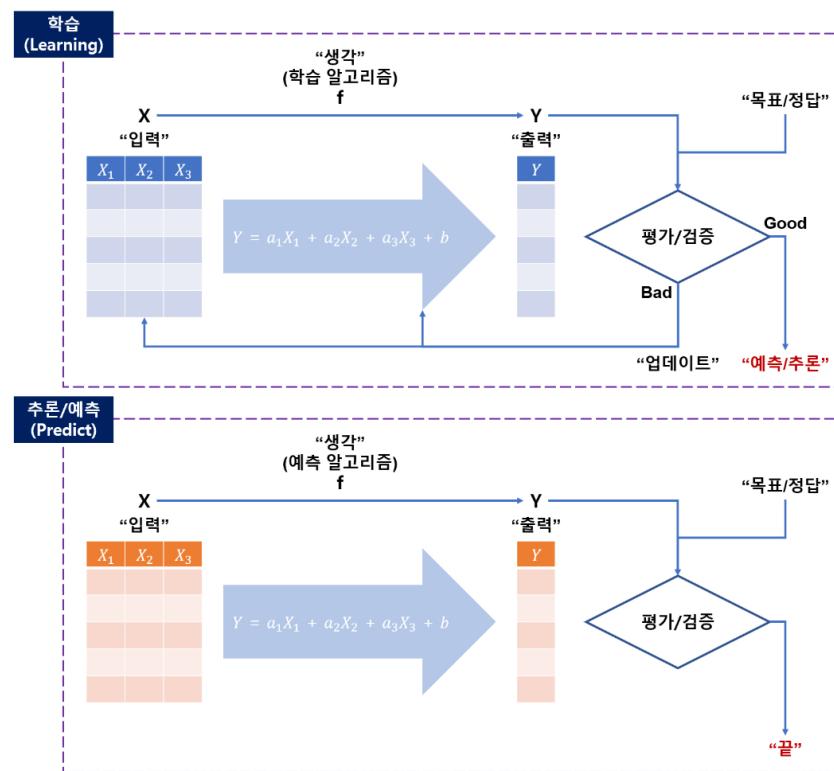
“문제해결 검증지표” 와 “알고리즘 검증지표” 는 같을 수 있으나 대부분은 다른 편”

(1) 문제해결 검증지표: 실제 문제를 잘 해결하는지 평가 (3단계)

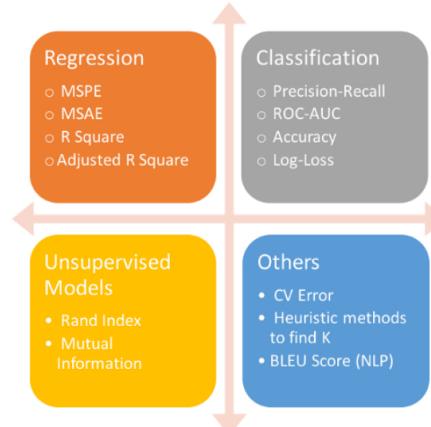
(2) 알고리즘 검증지표: 데이터의 패턴이 잘 추출되고 예측의 정확성을 평가 (2단계)

- 알고리즘 성능이 좋은 것과 문제해결이 가능한 것은 다르기 때문에 문제해결 지표와 알고리즘 지표는 대부분은 다른 편
- 알고리즘 검증지표는 없어도 되지만 문제해결 검증지표는 반드시 필요
- (이론적) 알고리즘들은 일반적으로 특정 알고리즘 검증지표를 향상시키는 방향으로 개발됨

## 6.1 대표적인 검증지표



### 1) 문제별 종류:



#### • Statistical Metrics: Correlation

- 입력(Input): -무한대 ~ 무한대 범위의 연속형 값
- 출력(Output): 이론적으로  $-1 \sim 1$  범위의 연속형 값

- **Regression Metrics:** MSE, MSPE, RMSE, RMSLE, MAE, MAPE, MPE, R^2, Adjusted R^2, ... ( $Y$ 의 범위가 무한대가 가능한 연속형일 때)

- 입력(Input): 무한대 ~ 무한대 범위의 연속형 값
- 출력(Output): 이론적으로 0 ~ 무한대 범위의 연속형 값

- **Classification Metrics:** Log Loss, Cross-entropy, ROC, AUC, Gini, Confusion Matrix, Accuracy, Precision, Recall, F1-score, Classification Report, KS Statistic, Concordant-Discordant Ratio, (ARI, NMI, AMI), ... ( $Y$ 가 2개 또는 그 이상개 수의 이산형일때)

- 입력(Input): 무한대 ~ 무한대 범위의 연속형 값
- 출력(Output): 알고리즘 종류에 따라 출력이 달라질 수 있음

- 확률(Probability): 0 ~ 1 범위의 연속형 값 (Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, ...)
- 집단(Class): 0 또는 1의 이산형 값 (SVM, KNN, ...)

- **Clustering:** Dunn Index, Silhouette, ...
- **Ranking Metrics:** Gain, Lift, MRR, DCG, NDCG, ...
- **Computer Vision Metrics:** PSNR, SSIM, IoU, ...
- **NLP Metrics:** Perplexity, BLEU score, ...
- **Deep Learning Related Metrics:** Inception score, Frechet Inception distance, ...
- **Real Problem:** ???

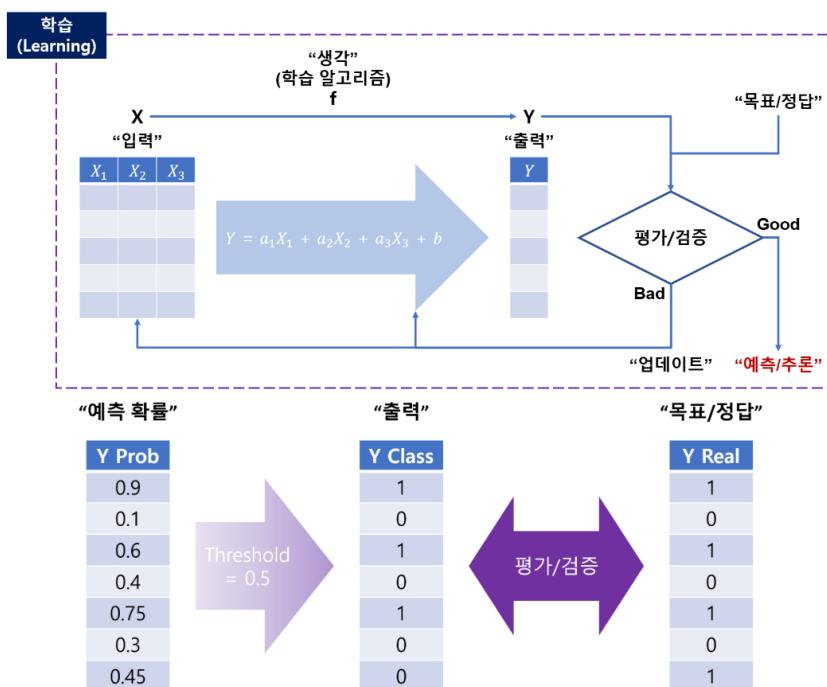
2) 검증지표 성능의 종류: 데이터/분석은 높은 정확도를 낳거나 높은 에러를 발생시킴

- **높은정확도(High Accuracy):** 과거 패턴이 미래에도 그대로 유지가 된다면 예측 정확도가 높아짐
- **높은에러(High Error):** 패턴이 점차적으로 또는 갑자기 변경되면 예측값은 실제값에서 크게 벗어날 수 있음

- **Black Swan:** 일어날 것 같지 않은 일이 일어나는 현상
- **White Swan:** 과거 경험들로 충분히 예상되는 위기지만 대응책이 없고 반복될 현상
- **Gray Swan:** 과거 경험들로 충분히 예상되지만 발생되면 충격이 지속되는 현상

## 6.2 분류분석 검증지표 및 해석하기

- **Structure:**



1) 오차행렬(Confusion Matrix): 정답 클래스와 알고리즘 예측 클래스의 일치 갯수 정리

- **Binary Classification**

	예측 0	예측 1
정답 0	정답이 0, 예측이 0인 데이터 수 (3)	정답이 0, 예측이 1인 데이터 수 (0)

- Multi-class Classification

예측 0	예측 1	...	예측 K
정답 0 정답 0, 예측 0인 데이터 수	정답 0, 예측 1인 데이터 수	...	정답 0, 예측 K인 데이터 수
정답 1 정답 1, 예측 0인 데이터 수	정답 1, 예측 1인 데이터 수	...	정답 1, 예측 K인 데이터 수
...	...	...	...
정답 K 정답 K, 예측 0인 데이터 수	정답 K, 예측 1인 데이터 수	...	정답 K, 예측 K인 데이터 수

2) 정확도(Accuracy): 전체 데이터 중 정확하게 예측한 클래스의 비율(0과 1을 모두 포함)

- 예측이 정답과 얼마나 정확한가?

예측 0	예측 1
정답 0 True Negative (TN) (3)	False Positive (FP) (0)
정답 1 False Negative (FN) (1)	True Positive (TP) (3)

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

3) 정밀도(Precision): 클래스 1로 예측한 값들 중 실제 클래스 1의 비율

- 예측한 것 중 정답의 비율은?
- 잘못 예측한 클래스 1의 비중을 파악하고 줄이는데 목적
- 암환자 가 아닌데 암에 걸릴거라고 예측하여 과도한 사람들의 검진 증가 우려

예측 1
정답 0 False Positive (FP) (0)
정답 1 True Positive (TP) (3)

$$\text{Precision} = \frac{TP}{TP + FP}$$

4) 재현율(Recall/Sensitivity/True Positive Rate): 실제 클래스 1 값들 중 예측 클래스 1의 비율

- 정답 클래스 1중 예측으로 맞춘 비율은?
- 잘못 예측한 클래스 0의 비중을 파악하고 줄이는데 목적
- 암환자 인데 암이 아니라 예측하여 과도한 사망률 증가 우려

예측 0	예측 1
정답 1 False Negative (FN) (1)	True Positive (TP) (3)

$$\text{Recall} = \text{TPR} = \frac{TP}{TP + FN}$$

5) F1점수(F1-score): 정밀도와 재현율의 Trade Off 관계 반영위해 (가중)평균으로 모두 잘 맞추었는지 평가

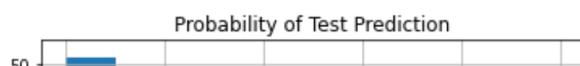
- 정밀도과 재현율이 모두 중요한 문제의 경우 중요
- 정밀도와 재현율을 따로 보면 한쪽으로 편중된(Bias) 의사결정이 될 수 있는 위험
- 다양한 평균의 종류 중 조화평균을 사용하여 계산
- 정밀도와 재현율 중 한쪽에 치우치지 않았을 때 높은 값

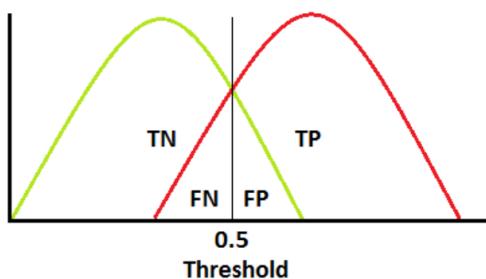
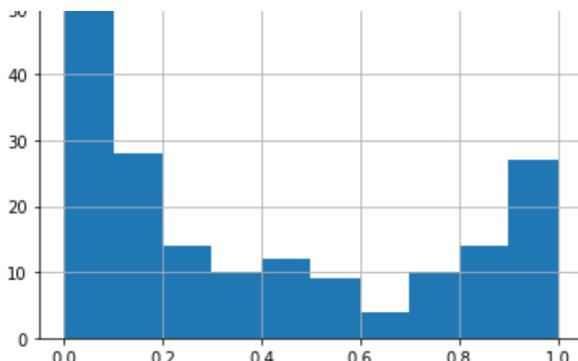
$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

6) ROC커브(Receiver Operator Characteristic Curve): 분류 기준값(Threshold)에 따라 재현율과 거짓율의 겸증지표의 변화를 확인하기 위한 시각화 지표



- 실제 예측결과 분포(히스토그램):





	예측 0	예측 1
정답 0	True Negative (TN) (3)	False Positive (FP) (0)
정답 1	False Negative (FN) (1)	True Positive (TP) (3)

- 재현율(Recall/Sensitivity/True Positive Rate): 실제 클래스 1 값들 중 예측 클래스 1의 비율

- 정답 클래스 1중 예측으로 맞춘 비율은?
- 잘못 예측한 클래스 0의 비중을 파악하고 줄이는데 목적
- 암환자 인데 암이 아니라 예측하여 과도한 사망률 증가 우려

	예측 0	예측 1
정답 1	False Negative (FN) (1)	True Positive (TP) (3)

$$TPR = \frac{TP}{TP + FN}$$

- 거짓율(Fall-out/False Positive Rate): 실제 클래스 0 값들 중 예측 클래스 1의 비율

- 정답 클래스 0중 예측으로 틀린 비율은?
- 다른 Metrics와 달리 낮을 수록 좋음

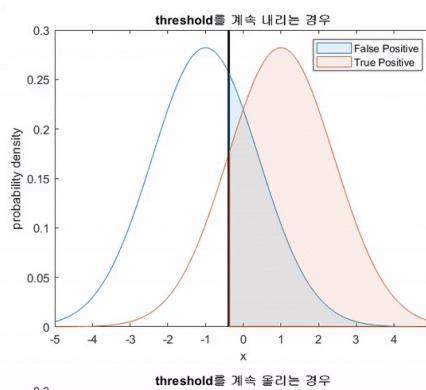
	예측 0	예측 1
정답 0	True Negative (TN) (3)	False Positive (FP) (0)

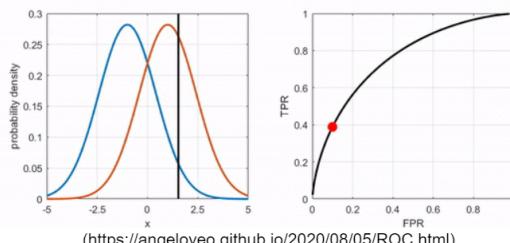
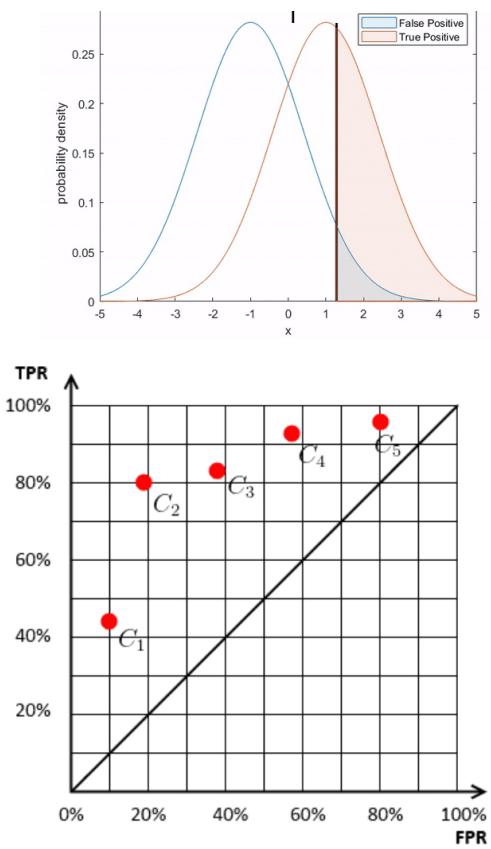
$$FPR = \frac{FP}{FP + TN}$$

- RUC Curve: 재현율과 거짓율의 변화를 시각화

기준값 변화에 따른 변화:

- 재현율(TPR)과 거짓율(FPR)은 양의 상관관계 존재
- Threshold가 낮아지면 1예측 갯수가 많아지고 TP와 FP모두 증가 → TPR & FPR 증가
- Threshold가 높아지면 1예측 갯수가 줄어들고 TP와 FP모두 감소 → TPR & FPR 감소





(<https://angeloyeo.github.io/2020/08/05/ROC.html>)

예시:

