빅데이터 표현

2023년도 1학기 – Week 01

조상구

과목명 : 빅데이터 표현

- 빅데이터 분석 프로세스를 이해하고 분석 수행에 필요한 Python의 기본적인 사용법과 주요 기능들을 익혀 데이터 분석 및 예측에 활용할 수 있다.
- 수업은 데이터 변환, 시각화 등을 포함한 인공지능 시스템을 구현하는데 필요한 이론적, 기술적 사항들을 배운다. 구체적인 주제는 다음과 같다.
- Python syntax와 자료 형태, Numpy, Pandas와 시각화
- 데이터 종류와 분석방법, 상관관계, 카이제곱검증, ANOVA, 회귀분석, 로지스틱회귀분석
- 데이터 특성공학(Feature Enginnering) - 실수형과 범주형 변환, survival ship bias
- 데이터 특성공학(Feature Enginnering) - Target mean, 이동평균법 등
- Matplotlib, seaborn - 1, Matplotlib, seaborn - 2

강의 계획

주차	요일	주제	강의 내용
1	3.09(목)	- 데이터 변환과 시각화, 데이터의 종류와 분석방법 이론 예시	범주형과 실수형 자료, 자료의 형태
2	3.16(목)	- Python syntax와 자료 형태	
3	3.23(목)	- Numpy	
4	3.30(목)	- Pandas와 시각화	
5	4.06(목)	- 데이터 종류와 분석방법	상관관계, 카이제곱검증, ANOVA, 회귀분석, 로지스틱 회귀분석
6	4.13(목)	- 데이터 특성공학(Feature Engineering)	실수형과 범주형 변환, survival ship bias
7	4.20(목)	- 데이터 특성공학(Feature Engineering)	Target mean, WOE, 이동평균법 등
8	5.04(목)	- Matplotlib, seaborn - 1	
9	5.11(목)	- Matplotlib, seaborn - 2	
10	5.18(목)	- 차원축소와 시각화(선형 - PCA)	Principal component analysis 와 시각화
11	5.25(목)	- 차원축소와 시각화(비선형, t-sne)	
12	6.01(목)	- 예측모형과 결과 시각화 - I	
13	6.08(목)	- 예측모형과 결과 시각화(Pycaret) - 2	
14	6.15(목)	- 딥러닝 representation - I	
15	6.22(목)	- 딥러닝 representation - II	

수업 내용

- **주요 수업 내용**
  - Python의 Syntax와 기본 기능을 기본 기능을 익히고 데이터 처리/변환과 예측 모형 개발을 실습한다.
    - 데이터 분석에 필요한 pandas와 시각화
    - 예측 모형에 투입될 데이터의 변환/표현 방법
    - 고차원 변수를 저차원 변수로 변환하여 시각화 표현하는 방법
    - 심층학습모델의 자료 representation 알고리즘 이해 및 실습

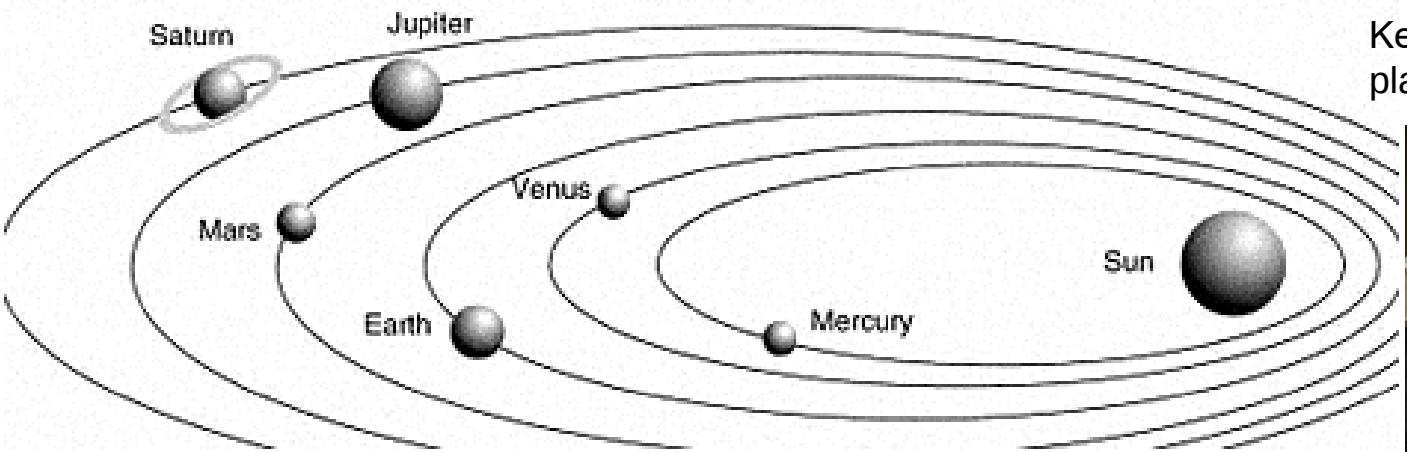
Data driven 의사결정 유형

데이터에 존재하는 규칙성(regularities)을 발견하기 위한 데이터분석(Data analytic)의 유형은 Descriptive, Diagnostic, Predictive, Prescriptive 등 4가지

https://www.analyticsinsight.net/four-types-of-business-analytics-to-know/

Pattern recognition - 1

패턴인식(Pattern recognition)은 데이터에 존재하는 규칙성(regularities)을 발견하여 새로운 데이터를 예측을 하는 것을 목적으로 한다.



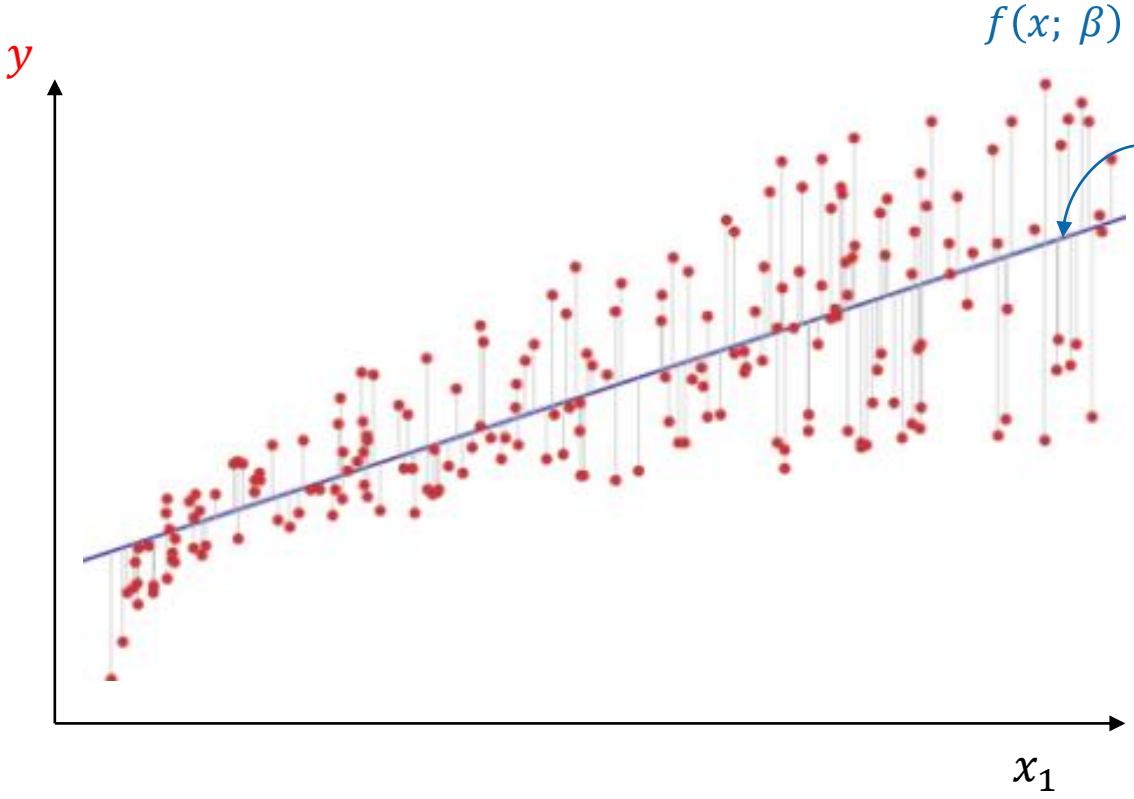
Kepler's empirical laws of planetary motion

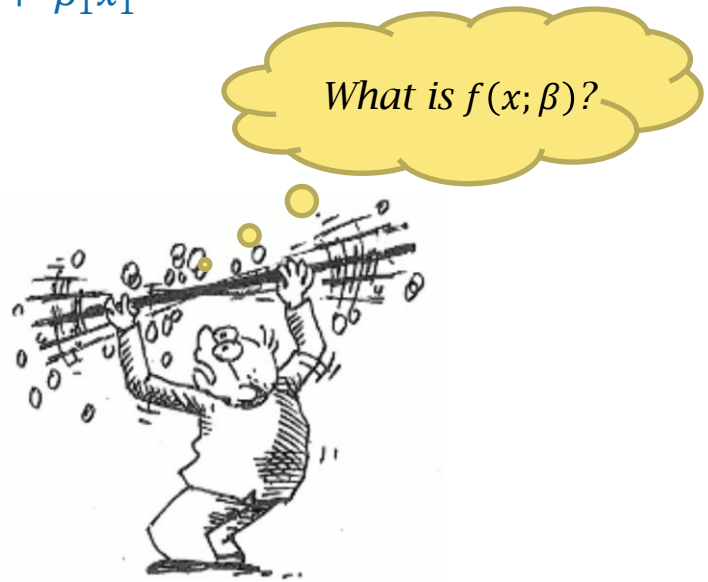
Rule 1, Rule 2, Rule 3

- 케플러 제1법칙(타원궤도 법칙),
- 케플러 제2법칙(면적속도 일정의 법칙),
- 케플러 제3법칙(조화의 법칙)

Pattern recognition - 2

선형함수(단순회귀분석 알고리즘) f 를 가정하여 손실함수(실제와 예측의 차이의 제곱의 합)을 최소화하는 절편과 기울기(β)로 구성된 예측 모델 $f(x;\beta)$을 개발하는 것이 머신러닝의 목적



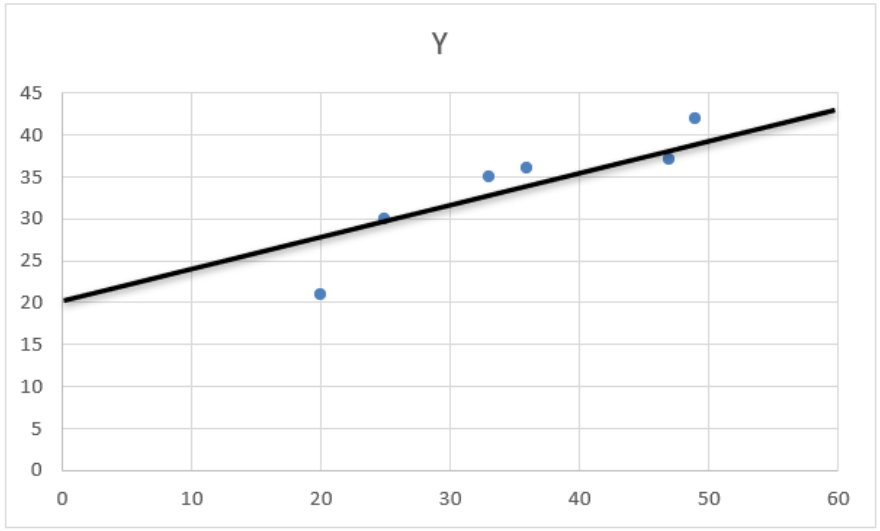


Minimize $(y - f(x;\beta))^2$
손실함수(loss function)

Pattern recognition - 3

F2 | =B$10+B$11*A2

	A	B	C	D	E	F
1	**X**	**Y**	Y hat	잔차	잔차 제곱	f(X)
2	20	21	27.8951	-6.895121951	47.5427	27.8951
3	25	30	29.7183	0.281707317	0.07936	29.7183
4	36	36	33.7293	2.270731707	5.15622	33.7293
5	33	35	32.6354	2.364634146	5.59149	32.6354
6	47	37	37.7402	-0.740243902	0.54796	37.7402
7	49	42	38.4695	3.530487805	12.4643	38.4695
8		201	200.188	0.812195122	71.3821	
9						
10	**Beta 0 hat**	**20.60244**				
11	**Beta 1 hat**	**0.364634**				



$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

머신러닝

컴퓨터에게 데이터에 존재하는 규칙성(regularities)을 발견하게 하여 새로운 데이터의 예측을 컴퓨터가 스스로 하게 하는 것 (사람이 규칙을 찾아서 문제를 해결하는 방법은 이제는 너무 어려움)

- Rule based approach

Data (Input) →
Program →
Traditional Programming → **Output**

- ✓ ***Big data***
- ✓ ***High dimensionality***
- ✓ ***Computing power***
- ✓ ***Algorithms***

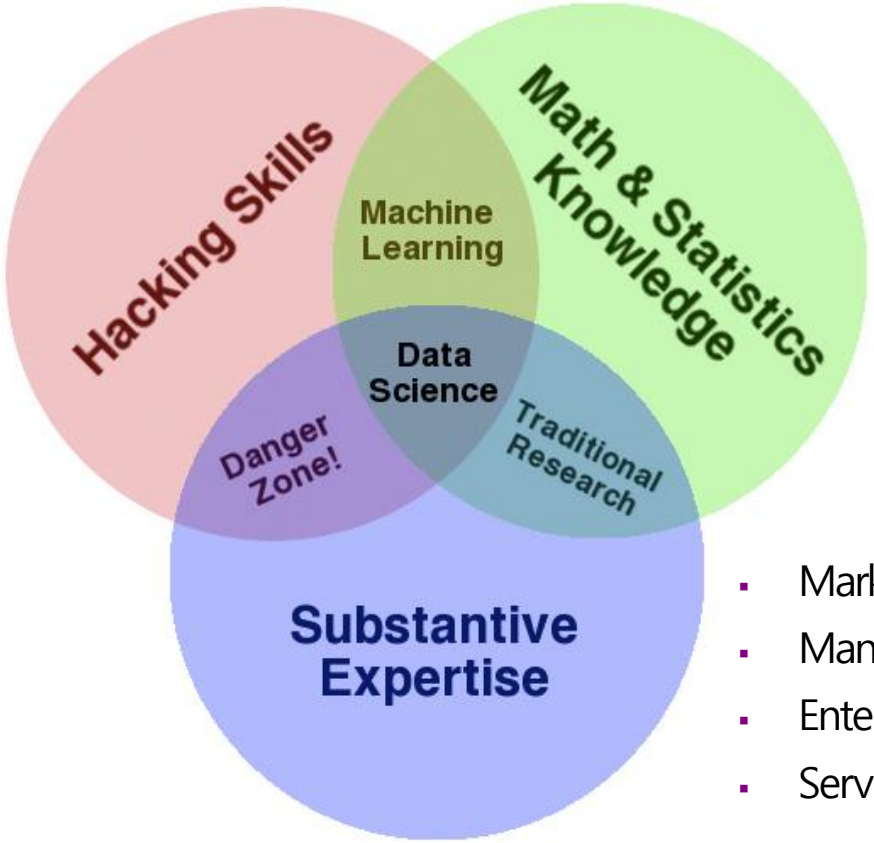
- Data driven approach

Data (Input) →
Output →
Machine Learning → **Program**

AI와 빅데이터 핵심역량

프로그램 코딩능력(Hacking Skills), 통계 추론 및 기초 수학 능력(Math & Statistics), 산업분야 Domain 전문 지식 등을 모두 갖추어야 데이터 사이언스가 완성

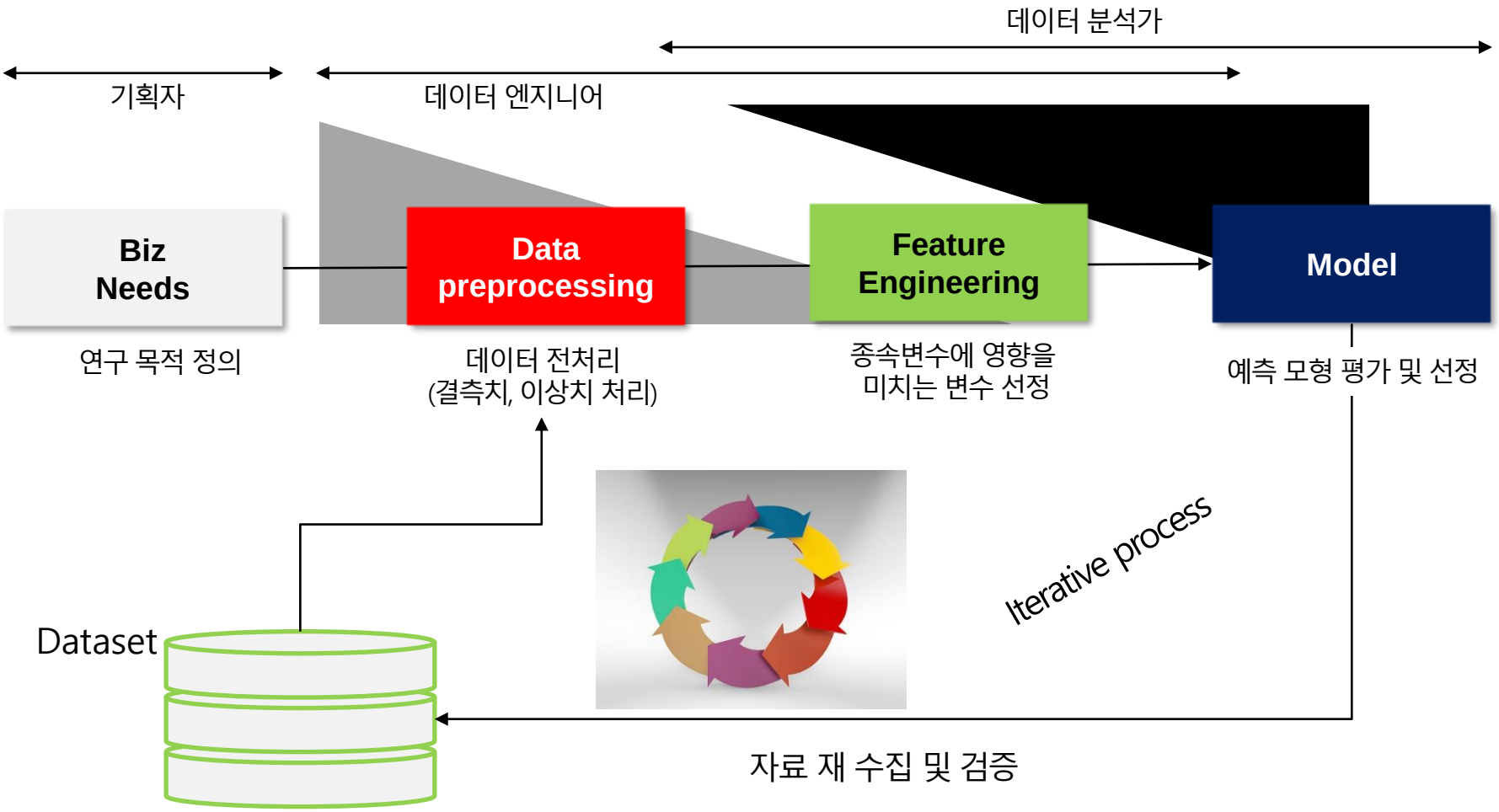
- Python
- Java, Java Script
- R
- Julia



- Linear Algebra, Calculus
- Probability theory
- Mathematical statistics
- SAS, STATA, SPSS, R

- Marketing
- Manufacturing
- Entertainment
- Services, etc

머신러닝 수행 절차


데이터 분석가
기획자
데이터 엔지니어
Biz Needs
Data preprocessing
Feature Engineering
Model
연구 목적 정의
데이터 전처리
(결측치, 이상치 처리)
종속변수에 영향을
미치는 변수 선정
예측 모형 평가 및 선정
Iterative process
Dataset
자료 재 수집 및 검증

머신러닝 분야

scikit-learn

Machine Learning in Python

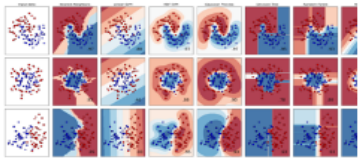
Getting Started | Release Highlights for 1.2 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

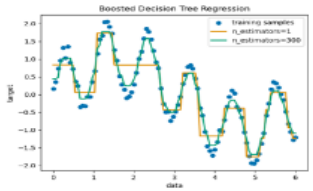


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

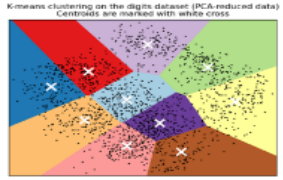


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

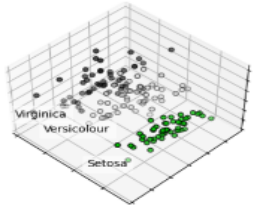


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization, and more...

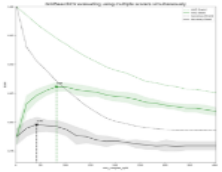


Examples

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...

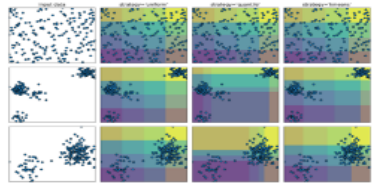


Examples

Preprocessing

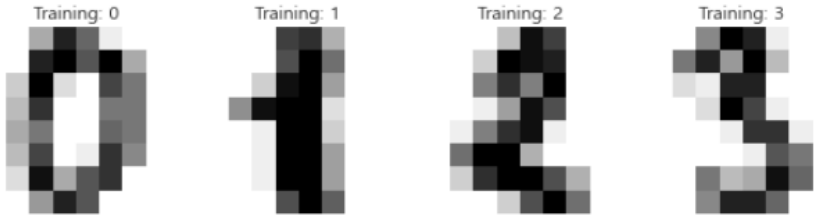
Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...

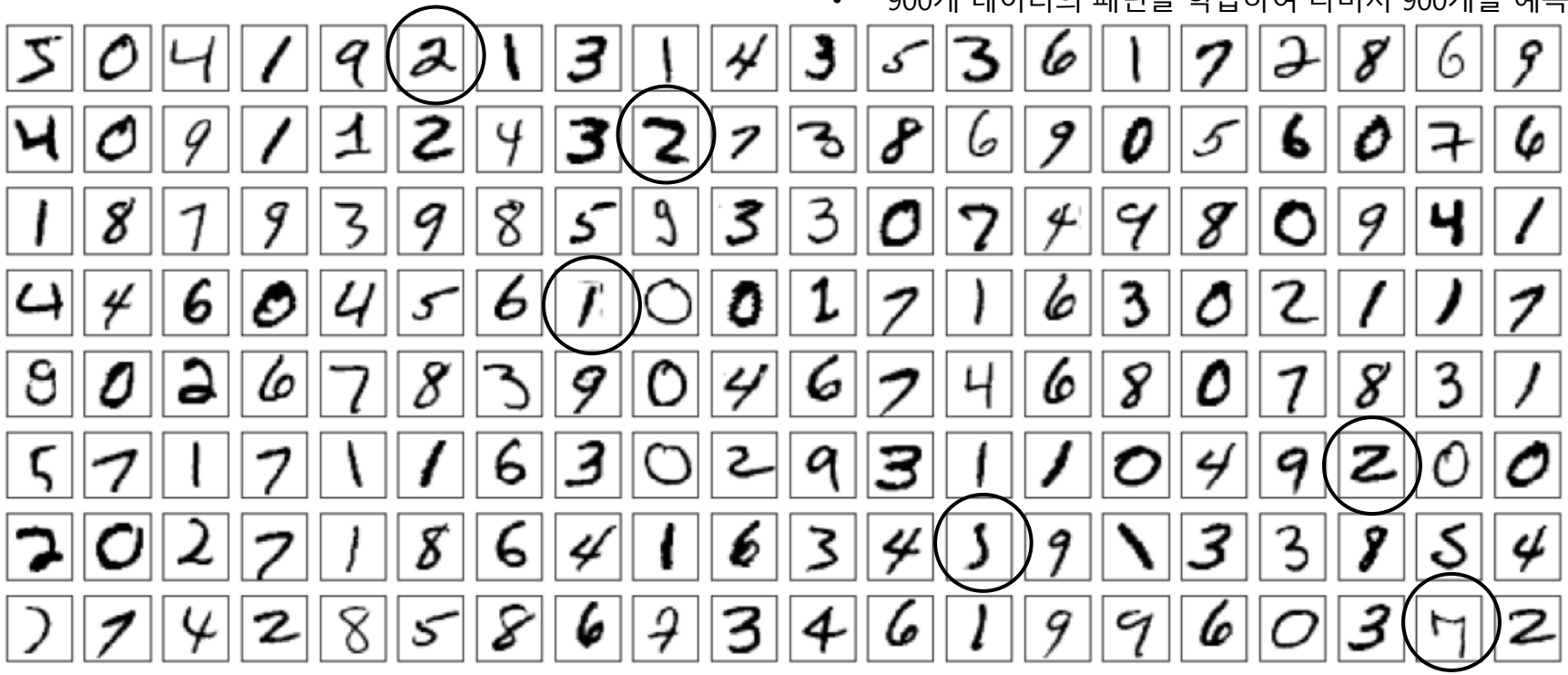


Examples

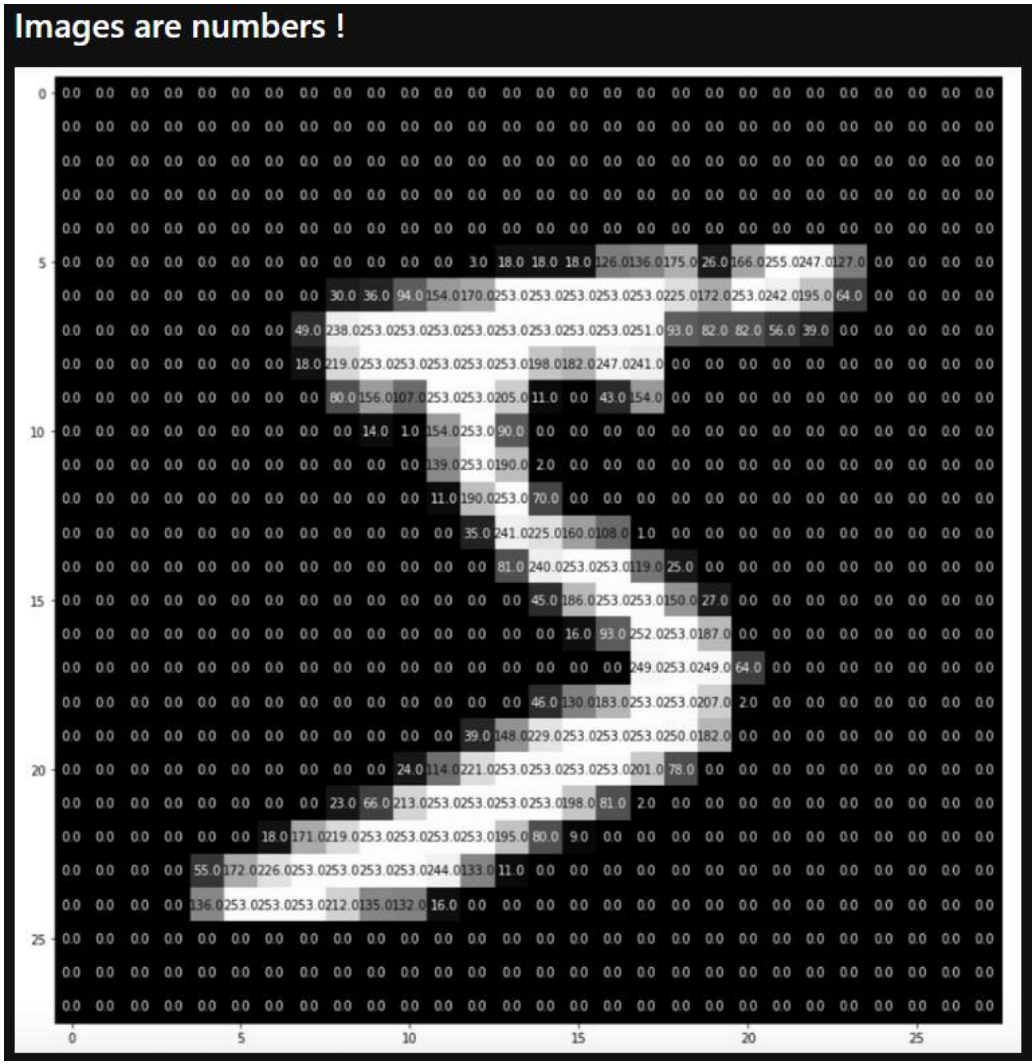
패턴인식_MNIST



- 1,797개의 다양한 필기체 아라비아 숫자
- 사람이 Rule을 찾기가 거의 불가능함
- 64개 특성변수(8 *8 pixel)
- 밝기는 각 픽셀에 0~255 실수 부여
- 900개 데이터의 패턴을 학습하여 나머지 900개를 예측



패턴인식_MNIST



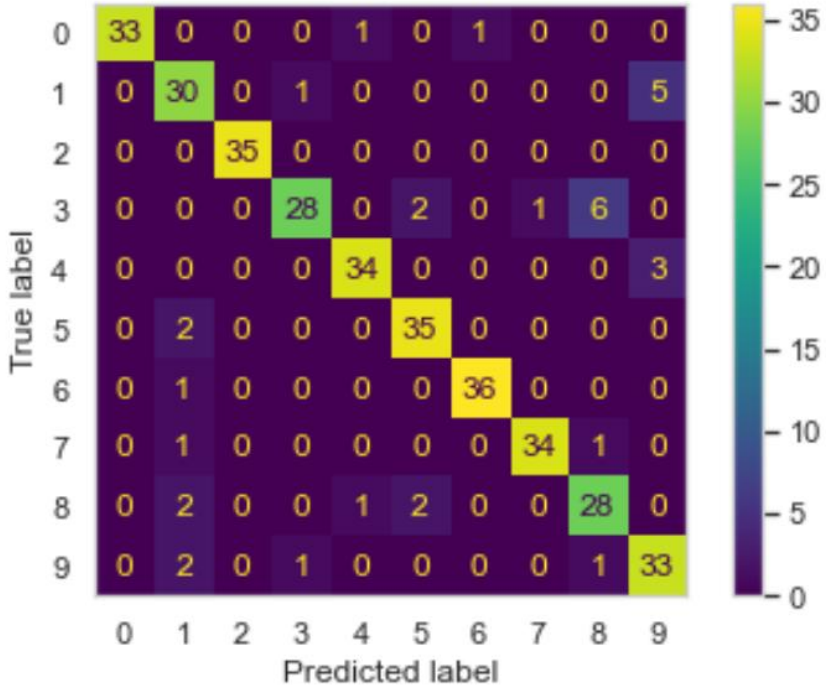
패턴인식_MNIST

간단한 프로그램으로 데이터의 패턴을 인식하여 숫자를 0 ~ 9 구분하는 정확도는 90.55%

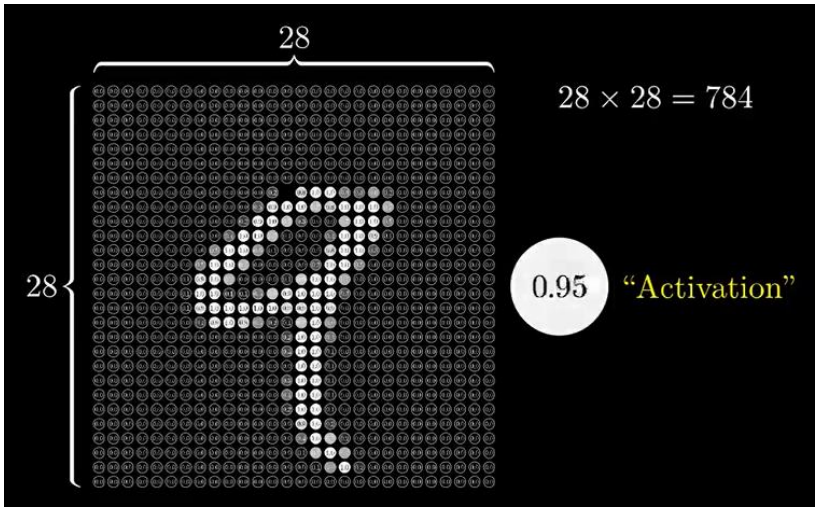
- '8', '9'번을 구분 못함
- '8', '9'번 판정의 정답률이 낮음

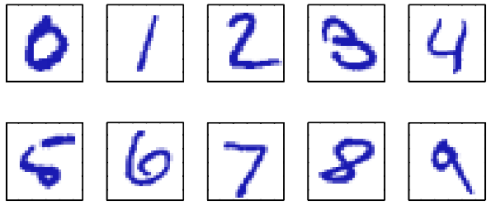
	pixel_0_0	pixel_0_1	pixel_0_2	pixel_0_3	pixel_7_5	pixel_7_6	pixel_7_7
0	0.0	0.0	5.0	13.0	0.0	0.0	0.0
1	0.0	0.0	0.0	12.0	10.0	0.0	0.0
2		0.0	0.0	4.0	16.0	9.0	0.0
3	0.0		7.0	15.0	9.0	0.0	0.0
4	0.0	0.0	0.0	1.0	4.0	0.0	0.0
5	0.0	0.0	12.0	10.0	10.0	0.0	0.0
6	0.0	0.0	0.0	12.0	11.0	3.0	0.0
7	0.0	0.0	7.0	8.0	0.0	0.0	0.0
8	0.0	0.0	9.0	14.0	11.0	1.0	0.0
9	0.0	0.0	11.0	12.0	3.0	0.0	0.0

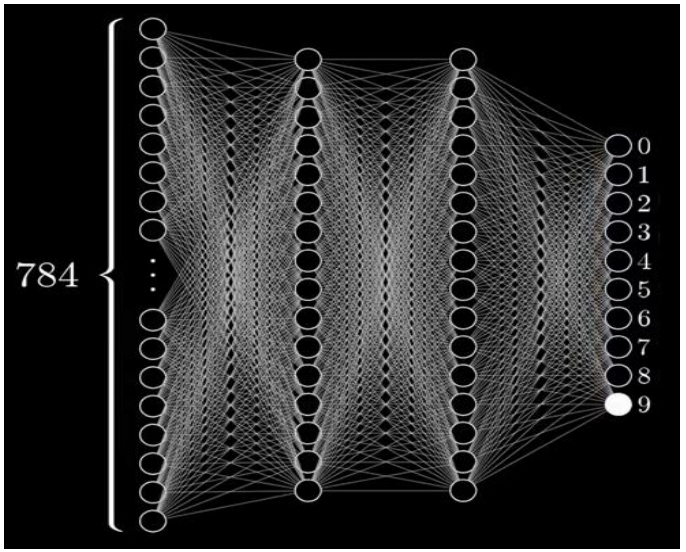

```
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
```



딥러닝





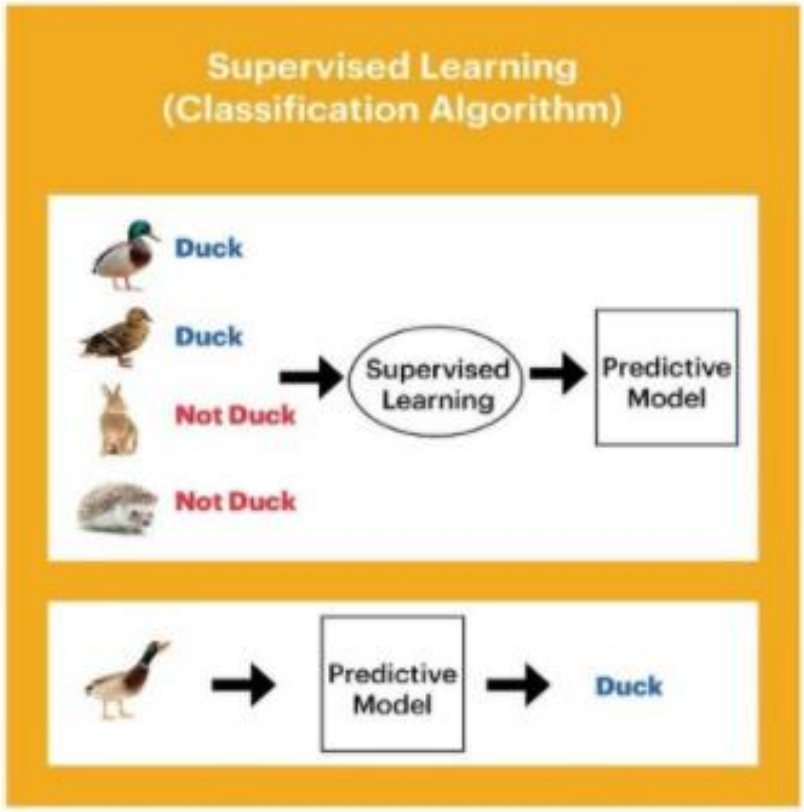


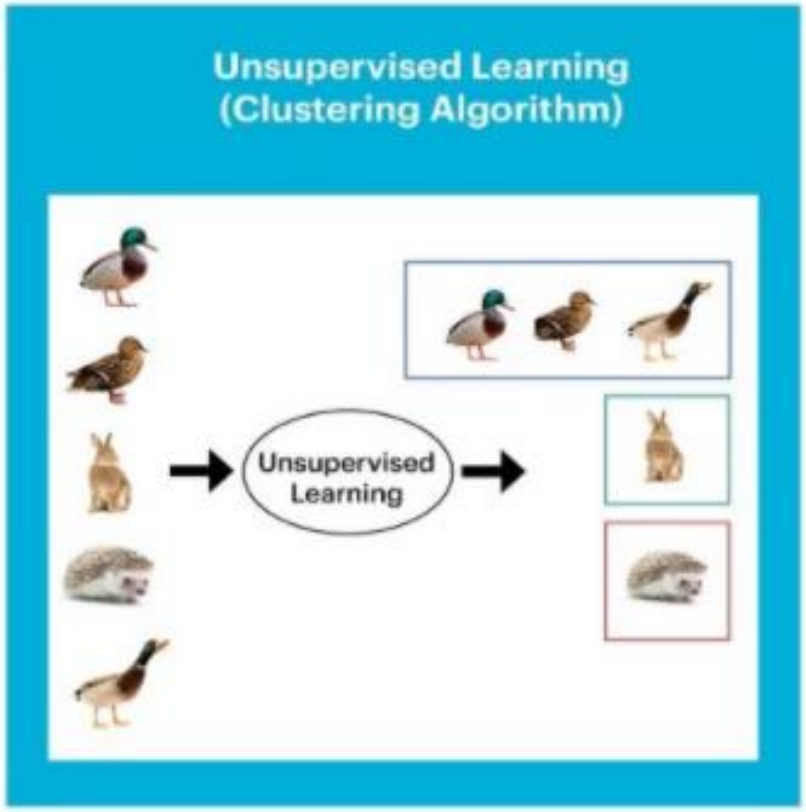
https://www.youtube.com/watch?v=aircAruvnKk&t=458s

머신러닝
머신러닝 종류

지도학습(Supervised learning)과 비지도학습(Unsupervised learning)

- 지도학습(Supervised learning) : Input과 Output 이 정해져 있음
- 비지도학습(Unsupervised learning) : Input의 특성만을 가지고 분류, Output 이 없음

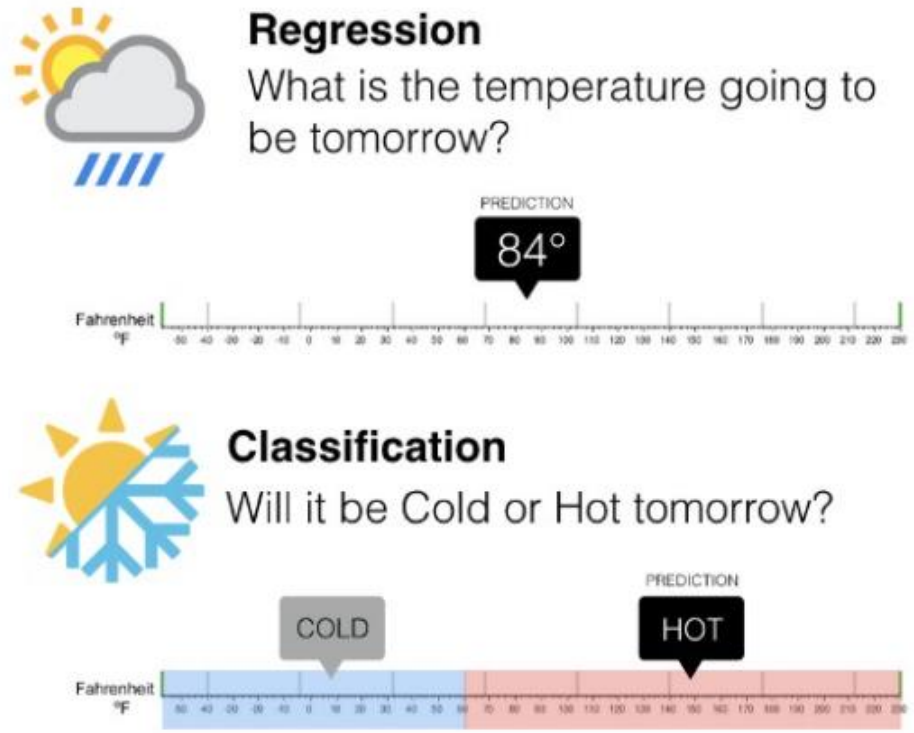




머신러닝
머신러닝 종류

분류(Classification)와 회귀생성(Regression)

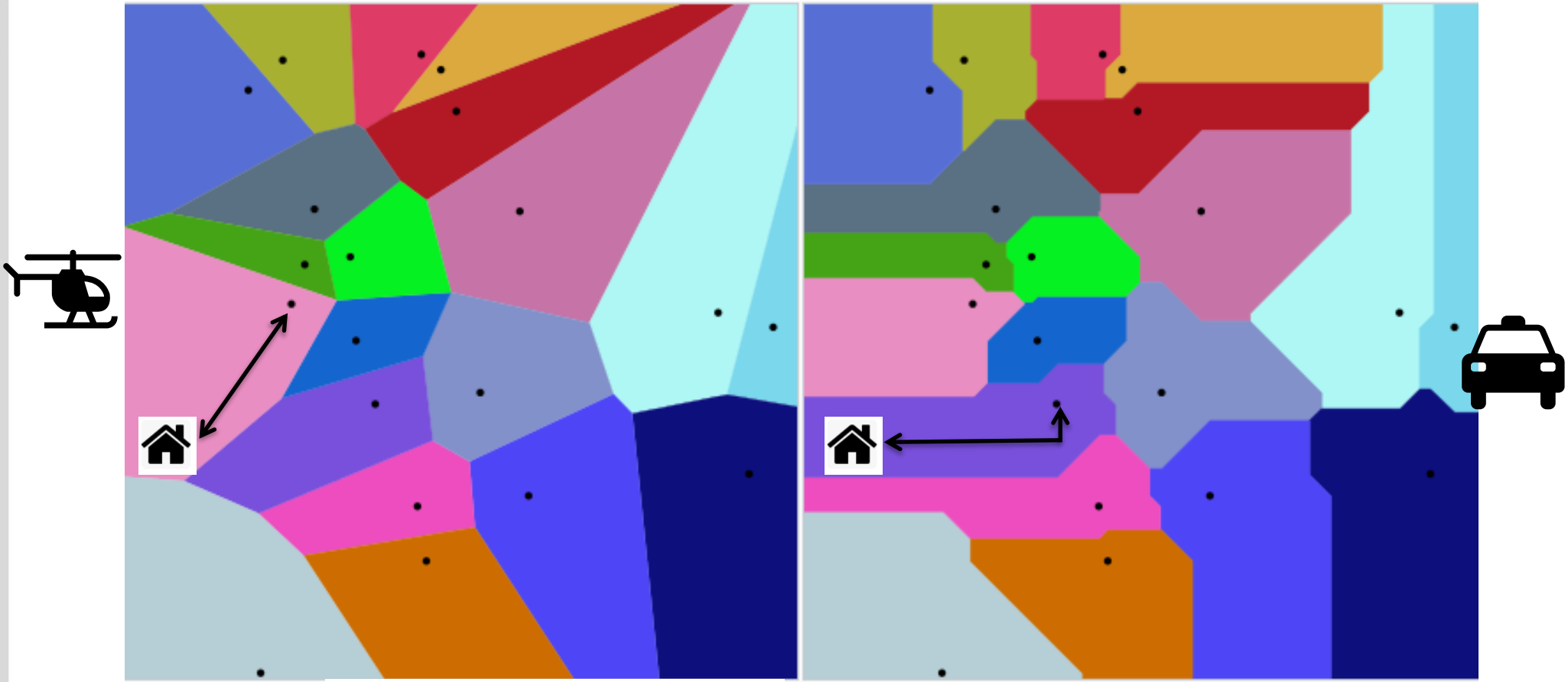
- 분류(Classification) : 명목형 변수(Categorical variables) 결과를 계측
- 회귀생성(Regression) : 수치형 변수결과를 예측



https://medium.com/@ali_88273/regression-vs-classification-87c224350d69

머신러닝
머신러닝 종류

Clustering은 유사한 데이터포인트(개체)들의 그룹을 판별하는 알고리즘



https://en.wikipedia.org/wiki/Voronoi_diagram