




Statistical Machine Learning

1주차
담당: 18기 방서연



KUBIG 24-2 Summer ML Session

Orientation

- 분반장 : 방서연(통계 21), 신인수(생공 17, 통계대학원 23)
- 진행방식
 - 매주 목요일 오후 7~9시 (5~10분 휴식)
 - 7주차 Coursework
 - 스터디 전반부 : 이전 주차 과제 우수자 발표 (5분 내외)
 - 스터디 중반부 : 강의진행
 - 스터디 후반부 : 과제 공지

.ipynb 파일로 제출, Python / google colab 권장

- 과제 마감 : 세션 전날 22시까지 (ex. 7월 10일 22시)
- 과제제출: KUBIG github (추후 안내 예정)
- PPT 자료 : Slack & github>방학분반>머신러닝
- Projects
 - 4주차 이후 조별 프로젝트 진행
 - 8/29 KUBIG Contest

Aa 주차	📅 날짜	👤 담당자	📖 주제	📑 과제
1주차	2024년 7월 4일	방서연	OT & What is ML?	
2주차	2024년 7월 11일	신인수	Supervised Learning	2주차 과제
3주차	2024년 7월 18일	방서연	Classification	3주차 과제
4주차	2024년 7월 25일	신인수	Regression	4주차 과제 & 프로젝트
5주차	2024년 8월 1일	신인수	Dimension Reduction	5주차 과제 & 프로젝트
6주차	2024년 8월 8일	방서연	SVM & Decision Tree	6주차 과제 & 프로젝트
7주차	2024년 8월 15일	신인수	Ensemble	7주차 과제 & 프로젝트
휴강	2024년 8월 22일		Contest 준비	
contest	2024년 8월 29일			

KUBIG 24-2 Summer ML Session

1) 수강 권장 대상

24-2 ML Session은 Data Science에 입문하시는 분들 & 개념을 탄탄히 다지고 싶은 분들을 위한 세션입니다.

2) 기초부터, 프로젝트까지!

분류예측 모델링은 환경/사회/정책/금융/산업 등 셀 수 없이 많은 분야와 접목시킬 수 있습니다.

머신러닝 개념에 대한 학습과 추후 커리어에 도움이 될 프로젝트까지 모두 경험해볼 수 있는 세션을 지향합니다.

KUBIG 24-2 Summer ML Session

20기 면접 질문 中 발췌

Q. Regression과 Classification의 차이를 loss 관점에서 설명해주세요.

Q. k-means clustering algorithm에 대해 설명할 수 있나요?

Q. Confusion Matrix이란 무엇이며, 이 행렬을 통해 얻을 수 있는 주요 지표에 대해 설명해주세요.

Q. 머신러닝에서 overfitting을 방지하기 위한 방법들을 소개해 주세요.

Q. 부스팅 모델이 무엇이고, 어떤 것들이 있나요? 그중 하나를 선택하여 모델의 주요 특징에 대해 설명해 주세요.

Q. 로지스틱 회귀 모델을 쓴 이유와, 로지스틱 회귀를 어떻게 해석해야하는지 구체적으로 설명해주세요.

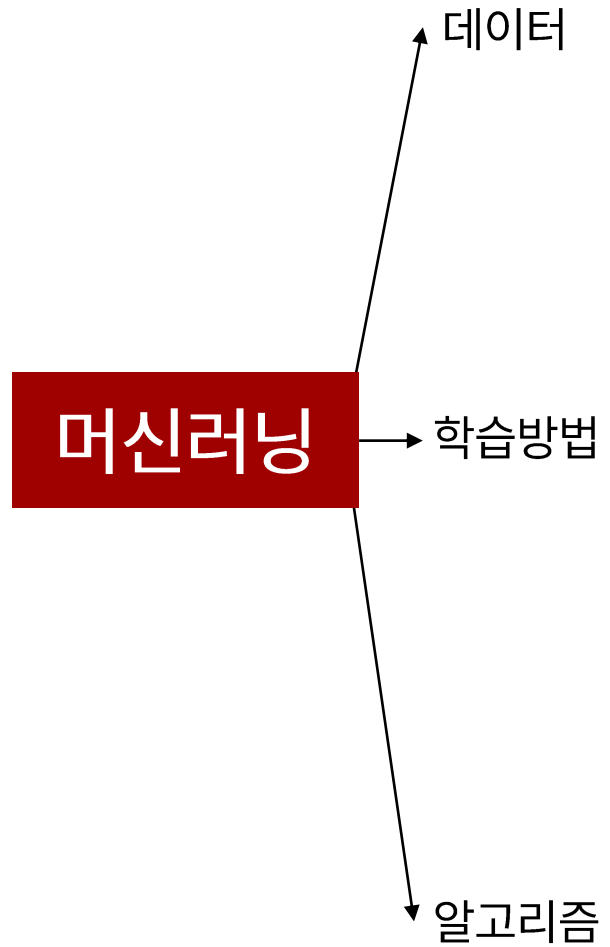
Q. 지도학습과 비지도학습의 차이를 설명해주세요.

Q. 적절한 군집의 개수를 찾는 방법이 팔꿈치 기법이라면, 군집이 잘 되었는지 사후에 평가하는 지표도 있습니다. 무엇인지 아시나요?

Q. 결손값을 처리하셨다고 했는데, 처리 방법과 해당 방법을 선택한 이유를 알려주세요.

Q. 어떤 상황에 oversampling/undersampling을 실시할까요?

KUBIG 24-2 Summer ML Session



KUBIG 24-2 Summer ML Session

PM

Data Engineer

Data Analyst

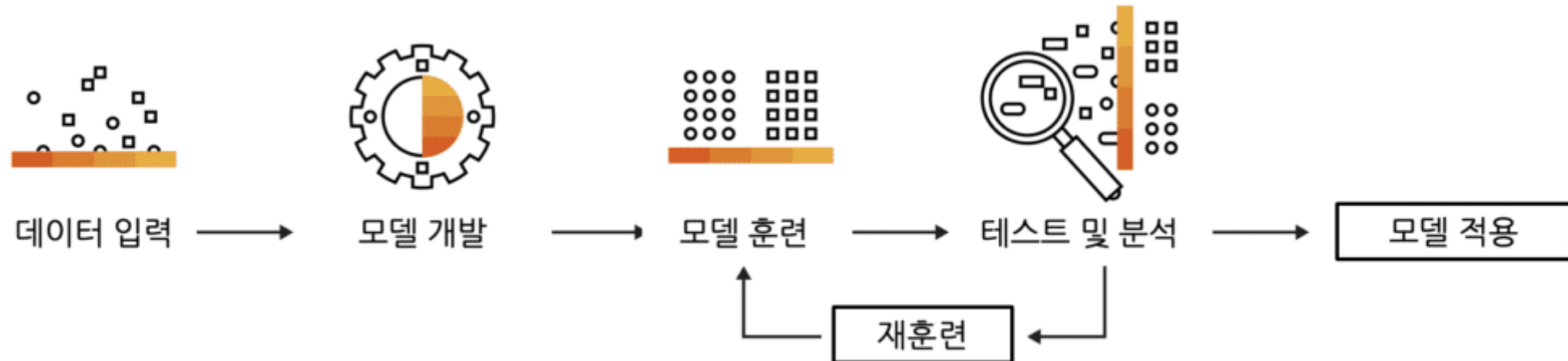
ML/DL Engineer

Data Scientist

BI

KUBIG 24-2 Summer ML Session

[ML Process]



1. How does a machine Learn?

Artificial Intelligence?

Intelligence?

- Gather Information
- Save Memory (and recall)
- Use it to **Learn (Inference?)**

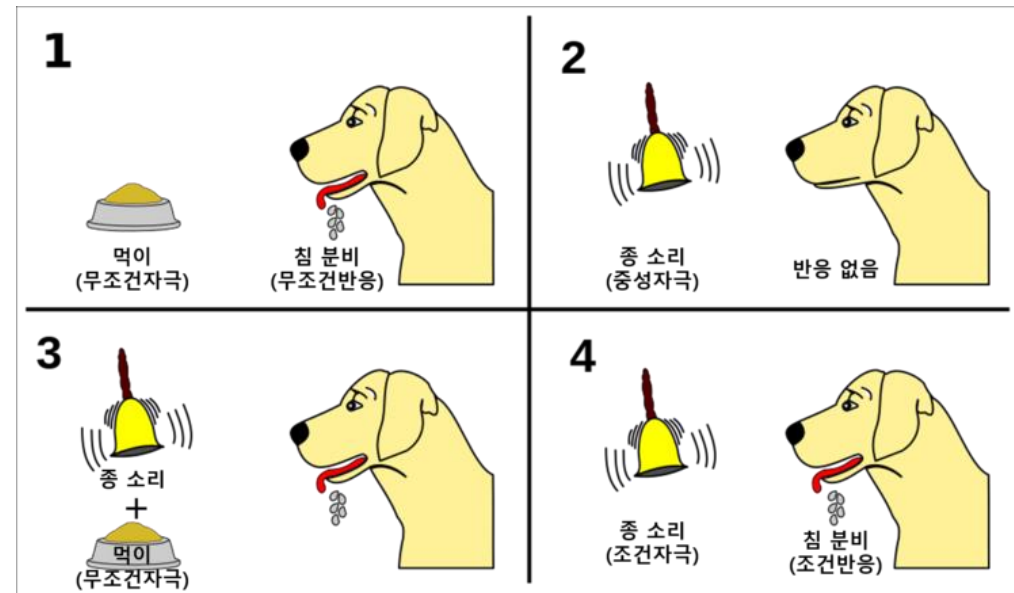
[국어사전]

학습 (學習)

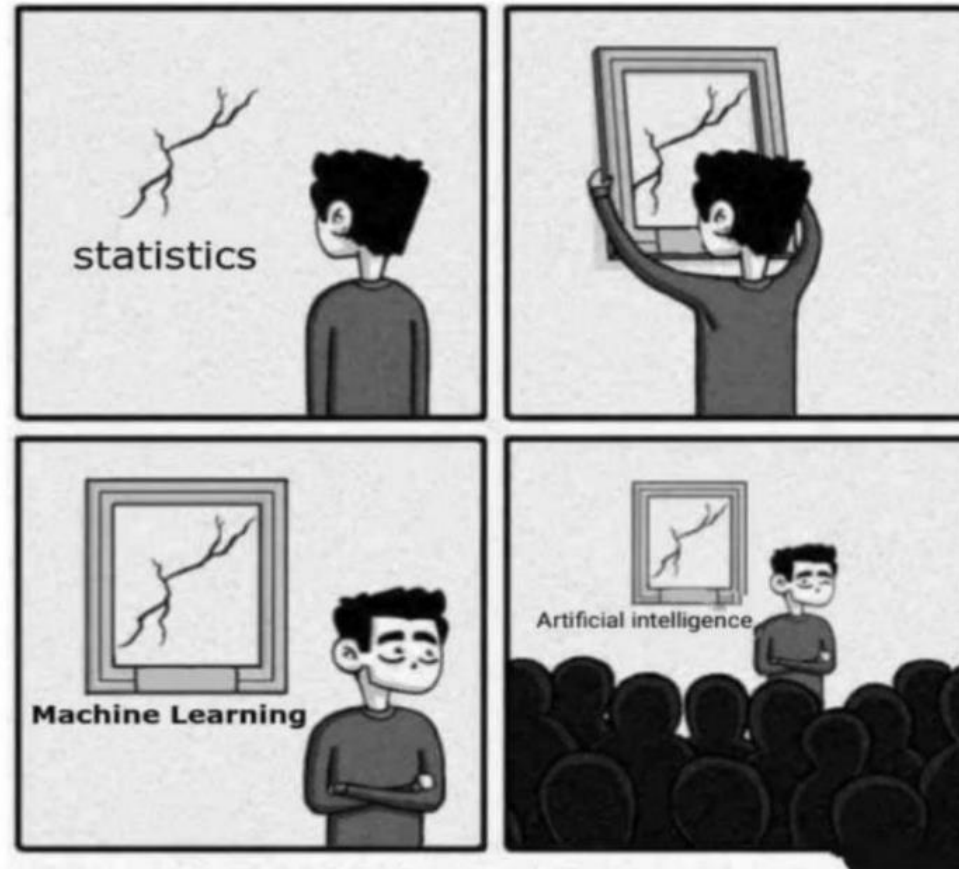
[학습] ㄱㄹ

1 배워서 익힘.

2 경험의 결과로 나타나는, 비교적 지속적인 행동의 변화나 그 잠재력의 변화. 또는 지식을 습득하는 과정.



Statistics? Machine Learning? AI?



What is Machine Learning?

"Machine Learning is the science of programming computers so they can **learn from data**"

"field of study that gives computers the ability to learn without being explicitly programmed" -
Arthur Samuel, 1959

"A computer program is said to **learn from experience E** with respect to some **task T** and some **performance measure P**, if its performance on T, as measured by P, improves with experience E" -
Tom Mitchell, 1997

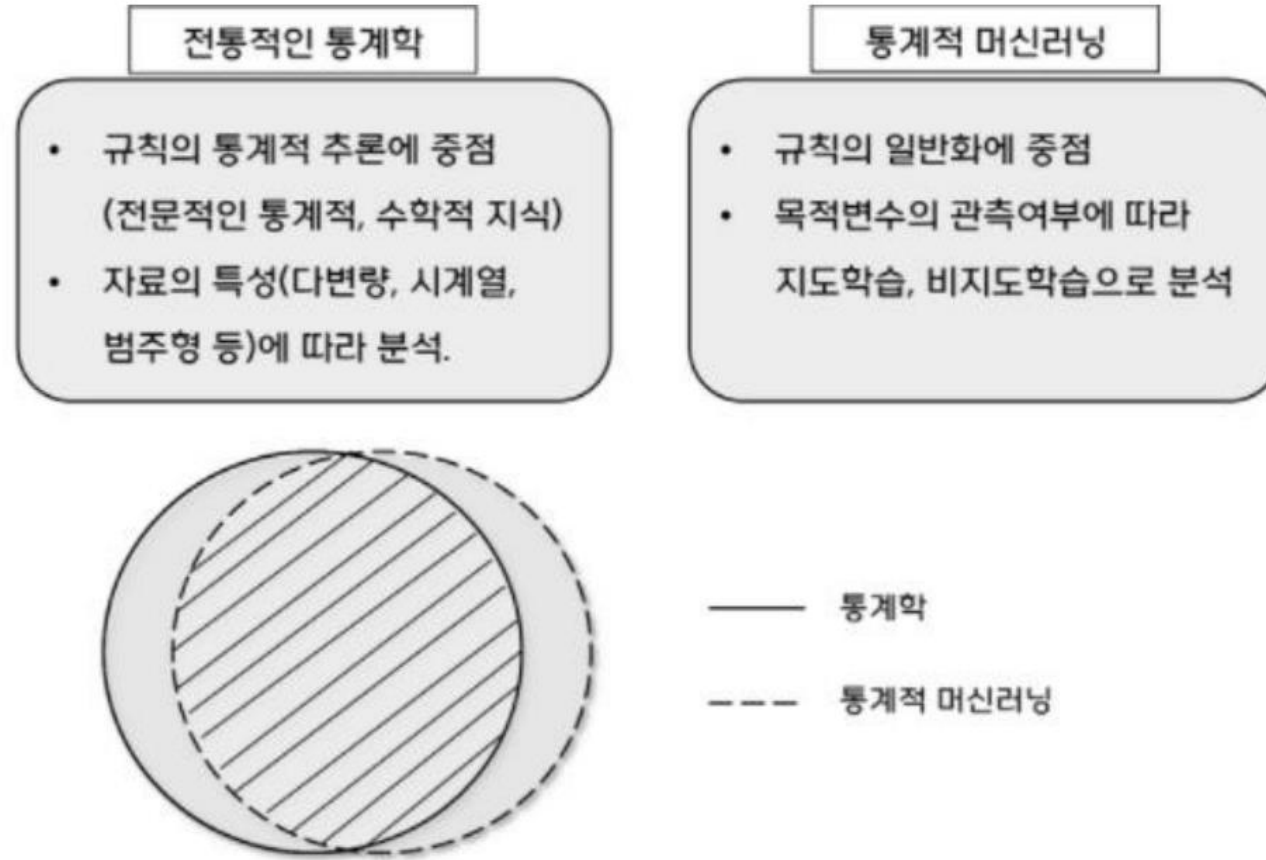
Experience (E) : spam인지 아닌지 사람이 정답을 매겨줌(데이터)

Task (T) : spam mail인지 아닌지 분류하는 일

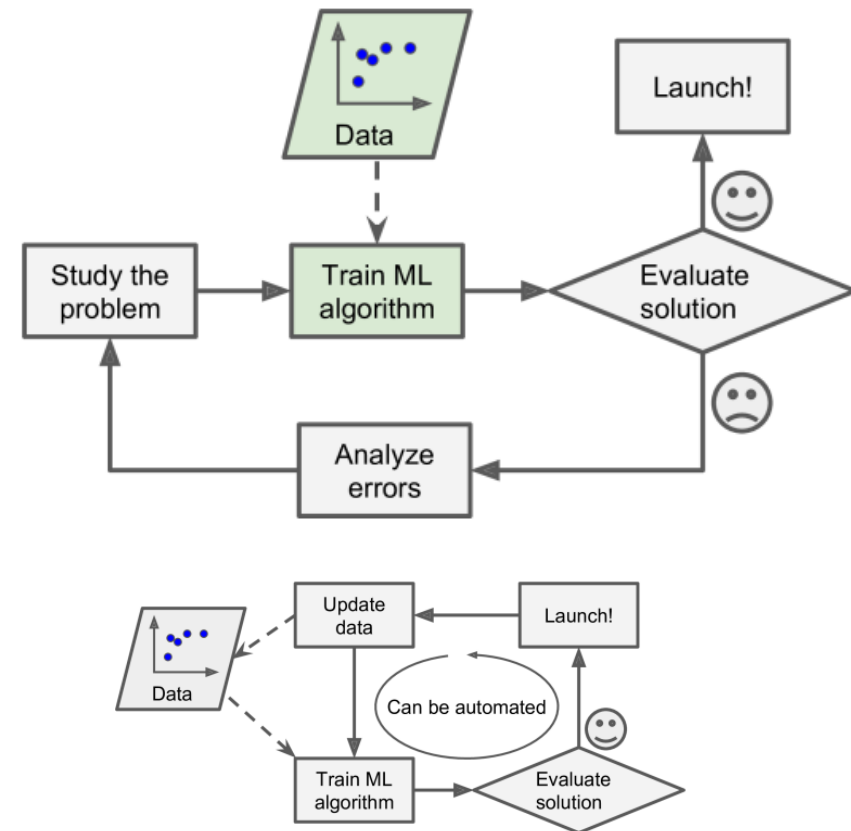
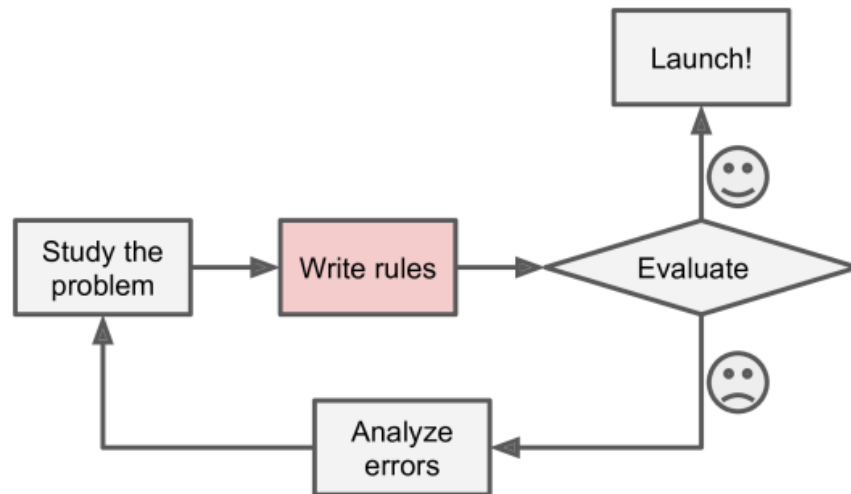
Performance (P) : 실제 spam인데 computer program이 실제 spam으로 잘 분류한 정도



What is Statistical Machine Learning?



Why use ML?



Types of ML systems

Whether or not they are trained with human supervision

⇒ Supervised, Unsupervised, Semi-Supervised, Reinforcement Learning

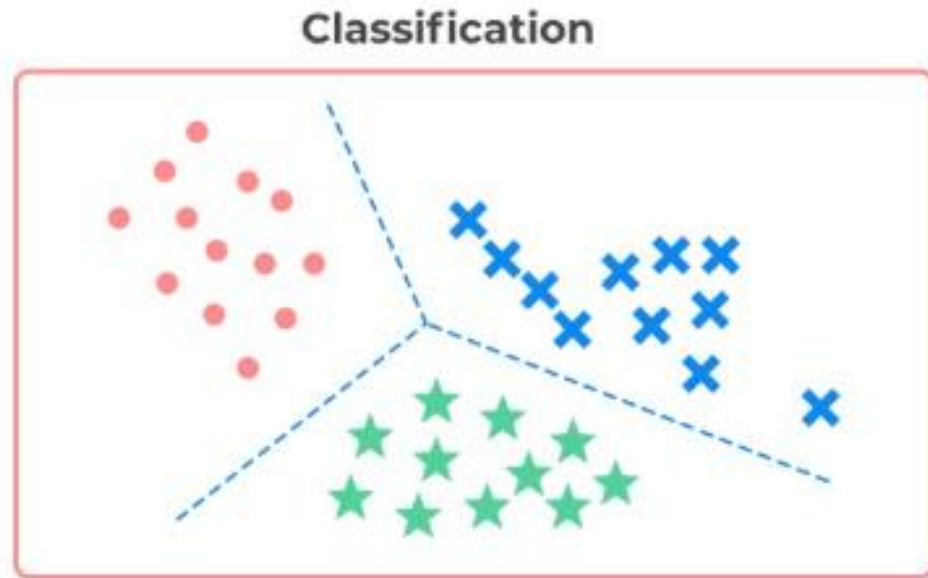
Whether or not they can learn incrementally on the fly

⇒ Online Learning, Batch(Offline) Learning

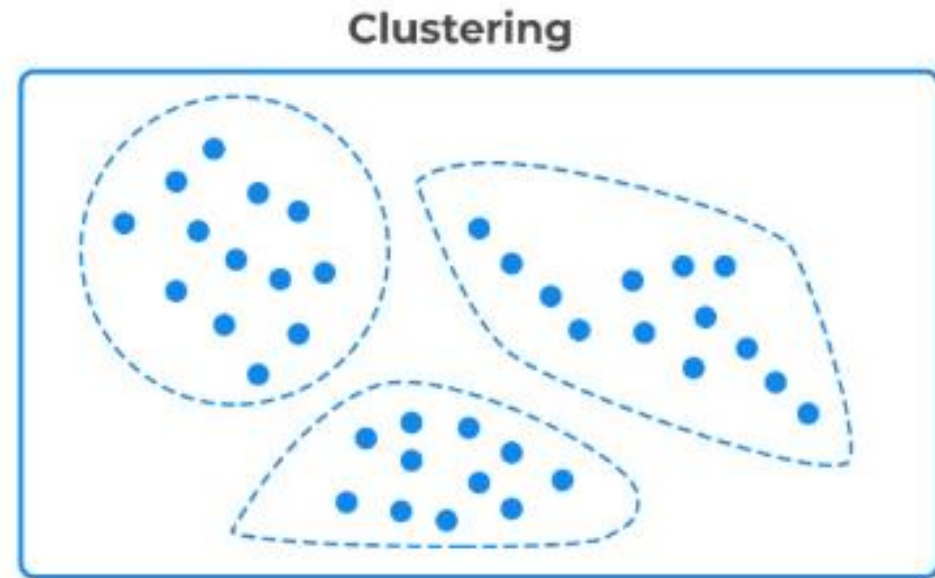
Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do

⇒ Instance-based Learning, Model-based Learning

Supervised vs. Unsupervised



Supervised learning



Unsupervised learning

$$\mathbb{P}_{train} \simeq \mathbb{P}_{test}$$



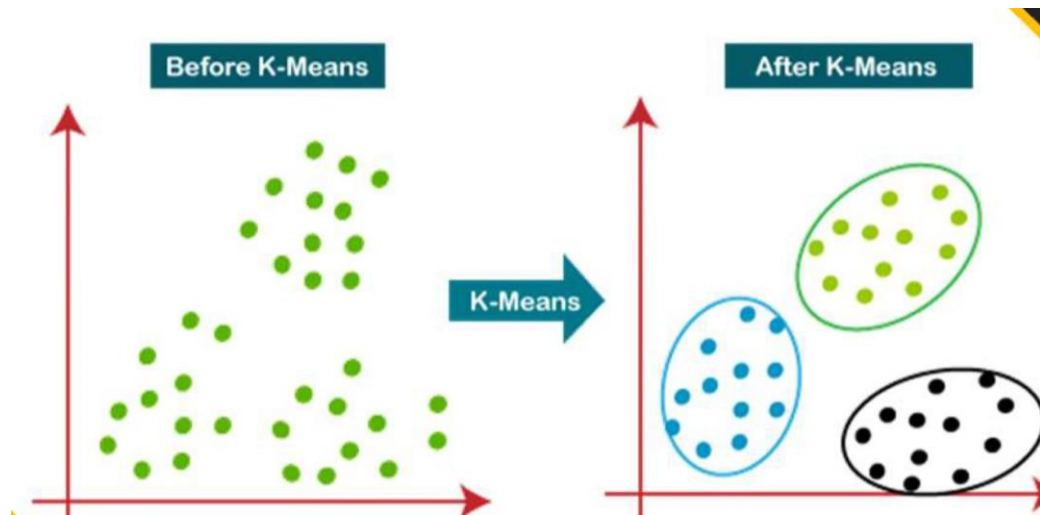
Critical Assumption in ML (중요)

학습하려는 데이터의 분포가
우리가 적용하려는 데이터의 분포와 유사하다

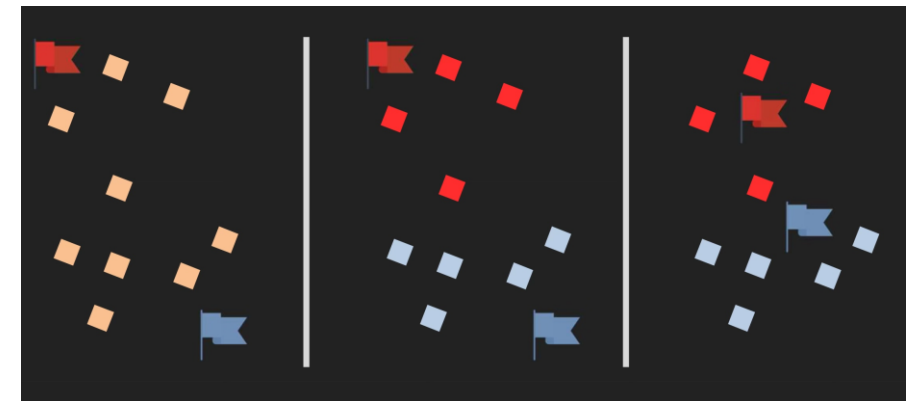
Unsupervised - Clustering

Clustering (군집화)

- K-means clustering



- 비지도 학습, 초기 중심점 선택, 중심점 재계산 및 반복
- 사전에 군집의 개수 k 결정
- 각 군집에는 중심이 존재하게 되는데, 중심과 군집 내 데이터 거리 차의 제곱합을 최소화 하는 최적 군집 탐색



```
Import sklearn.cluster import Kmeans
```

```
model = KMeans(n_clusters=3, random_state=0)
```

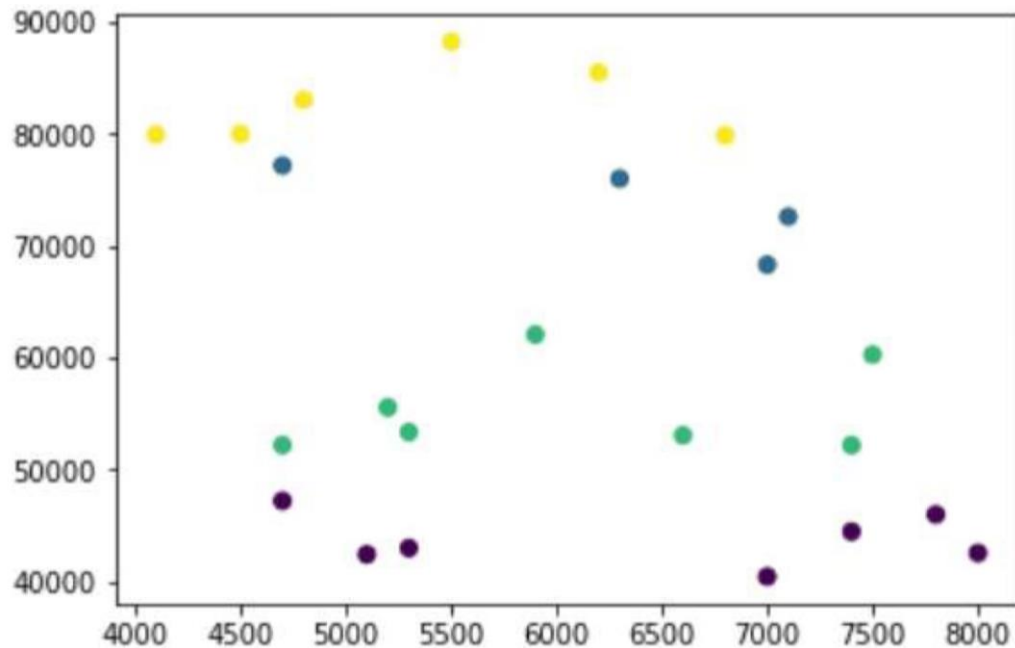
```
model.fit(df)
```

```
result = model.predict(df)
```


Unsupervised - Clustering

Clustering (군집화)

- K-means clustering

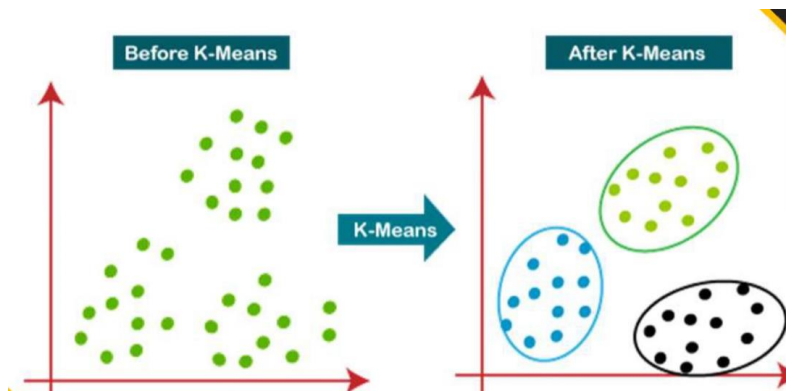


- Clustering 결과는 Scaling의 영향을 매우 많이 받음
- 거리 기반 계산

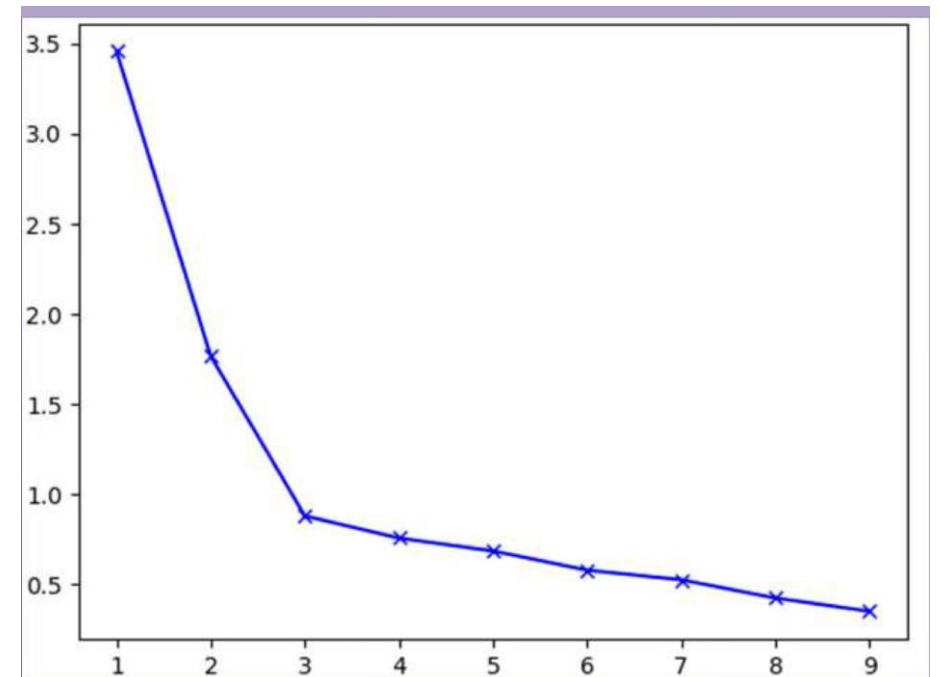
Unsupervised - Clustering

Clustering (군집화)

- K-means clustering, 적절한 군집의 개수?



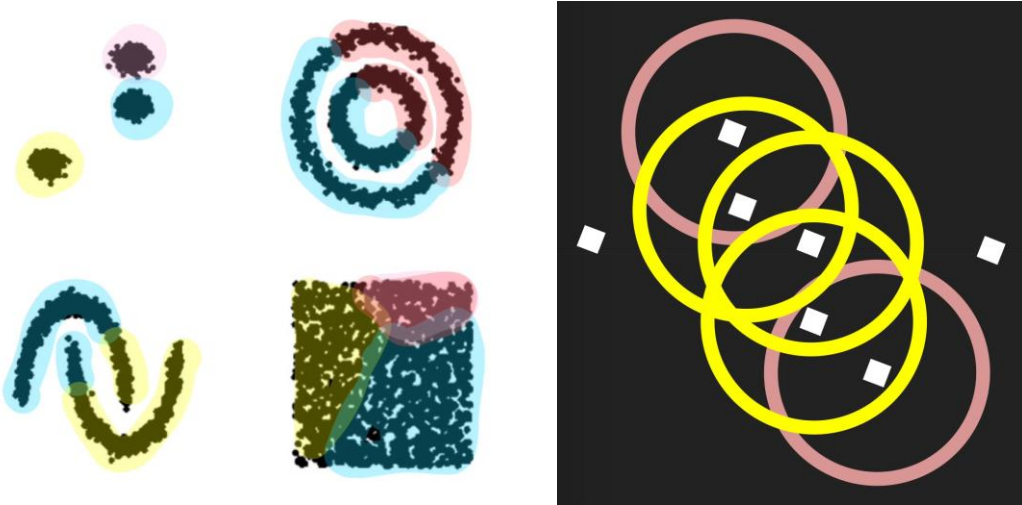
- K=3이 아닐 경우?
- Elbow Method



Unsupervised - Clustering

Clustering (군집화)

- DBSCAN

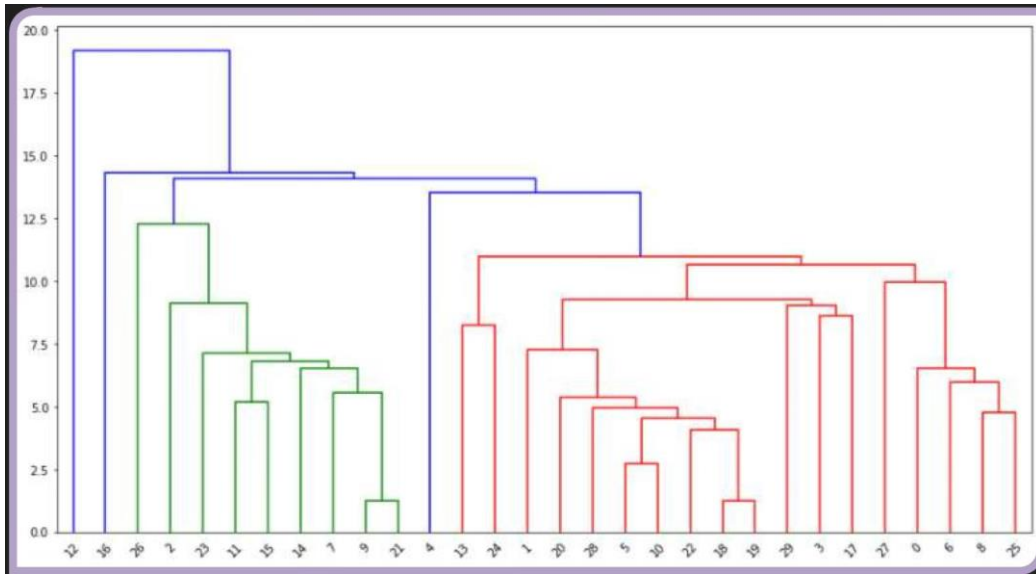


- 원의 반지름과 원 안에 포함될 최소 점 개수를 사전에 지정
- 최소 개수보다 더 많은 점이 포함될 경우, 원으로 점들을 묶음
- 만들어진 원들을 하나로 이어 최종 군집 결정
- 노이즈 구별 가능 (k-means clustering은 불가)

Unsupervised - Clustering

Clustering (군집화)

- Hierarchical clustering



Y축?

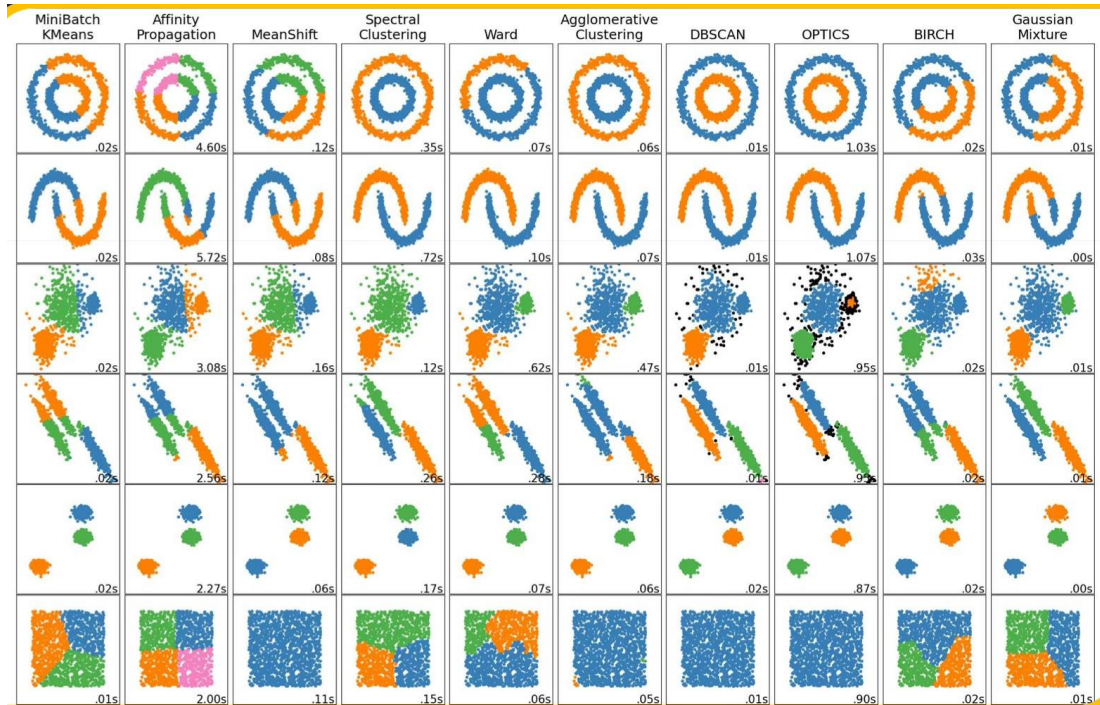
- 데이터로 주어진 모든 개체끼리의 유사도 계산
- 유사도가 가장 높은 두 개체를 하나의 군집으로
- 같은 과정을 반복해 Tree 형태로 최종 군집화
- Dendrogram 시각화 가능

군집 간 유사도 계산 방법

- 1) **Single Linkage**: 서로 다른 그룹의 점 사이의 최솟값이 유사
- 2) **Complete Linkage**: 서로 다른 그룹의 점 사이 거리의 최댓값이 유사도
- 3) **Average Linkage**: 서로 다른 그룹의 점 사이의 거리의 평균값이 유사도

Unsupervised - Clustering

Clustering (군집화) 정리



K-means	군집의 수를 정해놓고, 각 군집의 중심에 데이터가 몰리게끔 군집화
DBSCAN	같은 군집에 속한 데이터끼리는 밀도 있게 모여 있을 것
Hierarchical Clustering	유사도가 높은 개체끼리 묶고, 그 군집을 다른 개체와 묶는 과정 반복

Unsupervised - Clustering

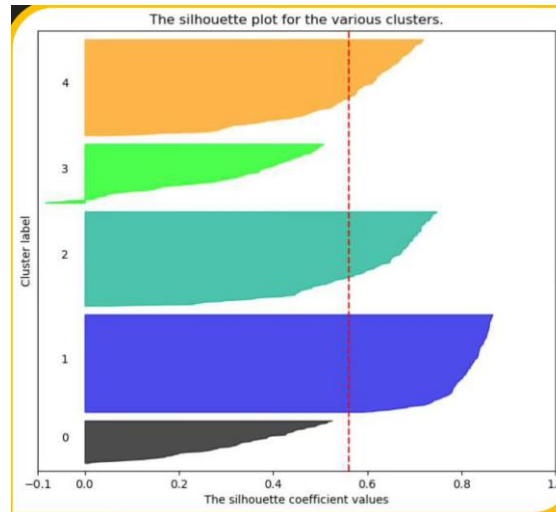
Clustering (군집화)

군집화 결과에 대한 평가 (사후) : Silhouette Analysis

군집화의 필요성에 대해 알아보자 (사전) : Hopkins Statistic

Silhouette Analysis (실루엣 계수)

- 군집화를 마친 뒤, 모든 데이터에 대하여 실루엣 계수 계산 가능
- 범위는 -1 이상 1 이하
- 값이 클 수록 군집화가 잘 이루어짐



Hopkins Statistic (홉킨스 통계량)

- 군집화 수행 전, 해당 데이터셋이 군집화에 적절한지 판단하기 위한 지표
- 0 이상 1 이하

2. End to end ML Project

EDA (Exploratory Data Analysis)

- Definition

수집한 데이터를 다양한 시각에서 이해하는 과정

모델링 전에 데이터가 어떻게 생겼는지 시각화, 통계적 검정, 통계량 등을 통해 살펴보는 과정

- 필요성

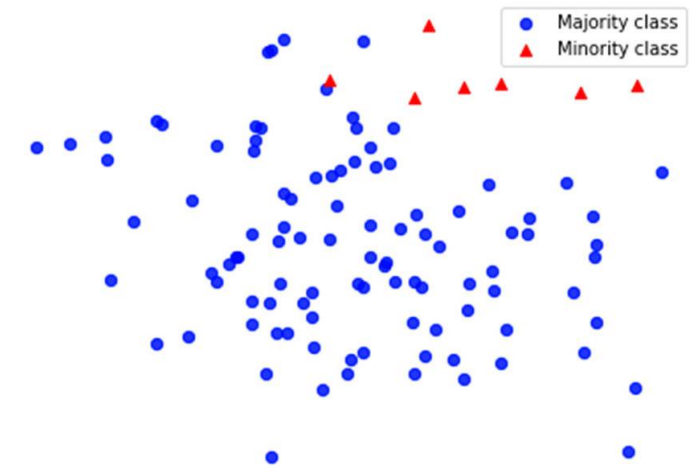
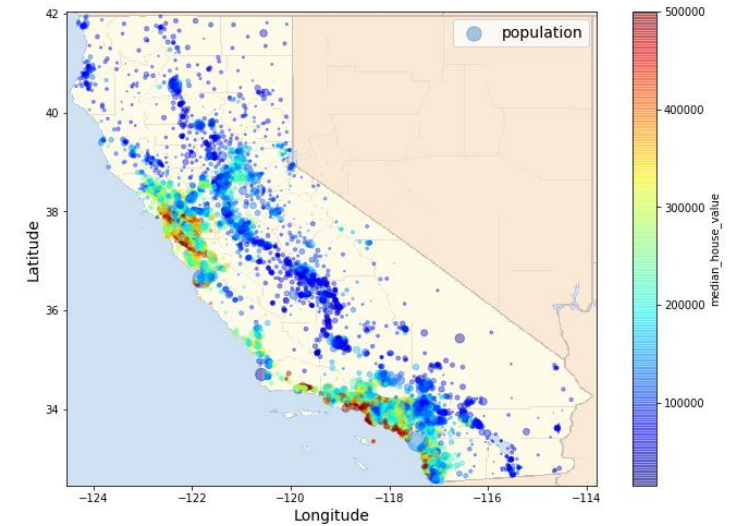
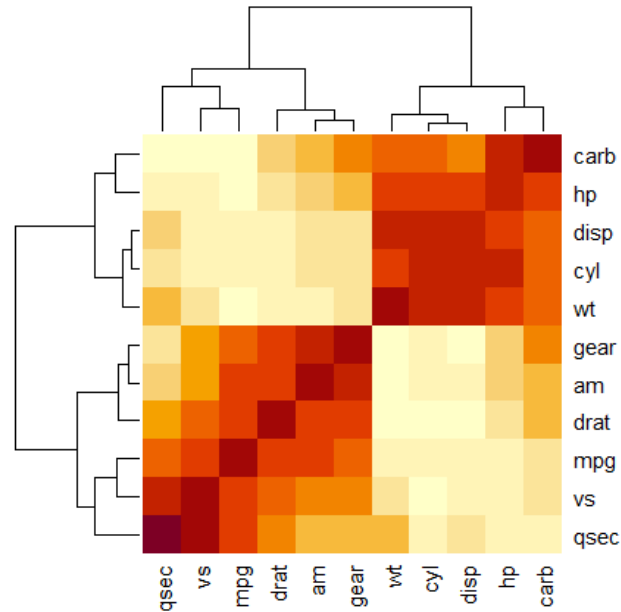
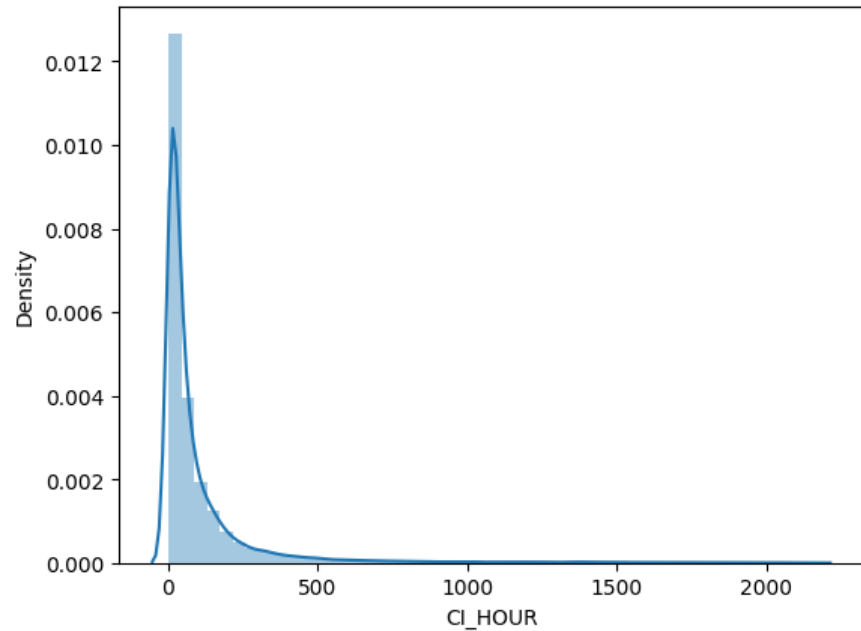
어떤 모델을 사용하고, 어떤 전처리를 할지 결정하는 것은 데이터의 형태가 결정함.

- 각 변수 분포/ 변수간 상관관계/ 이상치/ 결측치/ 통계적 (분포)검정/ class imbalance ...

EDA (Exploratory Data Analysis)

```
# CI_HOUR ==0 인 행들은 모두 DIST == 0 인가? # yes :)  
train[train.CI_HOUR == 0].loc[:,['DIST']].sum()
```

```
DIST    0.0  
dtype: float64
```



Preprocessing (데이터 전처리)

정규화(Normalization) 표준화(Standardization)

나이 (X_1)	월별 소득 (X_2)
30	3,620,000
13	0
21	600,000
61	500,000

scale차이에서 오는 변수간
영향력 차이를 제거하기
위해 정규화/표준화 진행

Dimension Reduction

Feature(basis)가 지나치게
많아 발생하는 curse of
dimension을 해결
⇒ PCA, SDR, tSNE

One-Hot Encoding

Scikit-Learn ML 모델들은
문자열 값을 인식하지 못함

각각 범주에 매칭되는
column에 1, 아니면 0 입력

이상치, 결측치

ML모델은 직접 결측치를
처리할 수 없음
이상치 때문에 모델 전반의
성능이 왜곡될 수 있음

Label Encoding

각각 범주에 매칭되는
column에 범주의 개수만큼
고유 index를 입력

Regression task에 적용시
dummy variable로 적용 필요

Class Imbalance

데이터가 부족한 특정
불균형 클래스에 대해
oversampling 진행
⇒ SMOTE, ADASYN

Learning Process

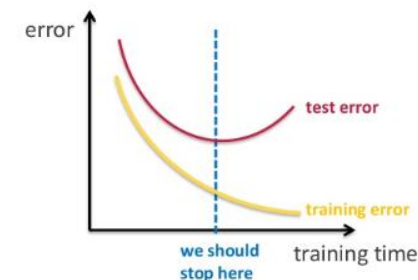
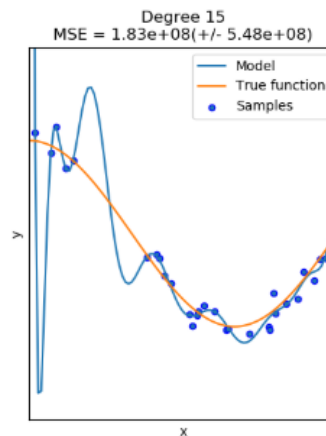
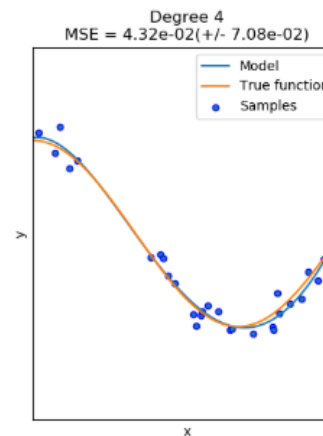
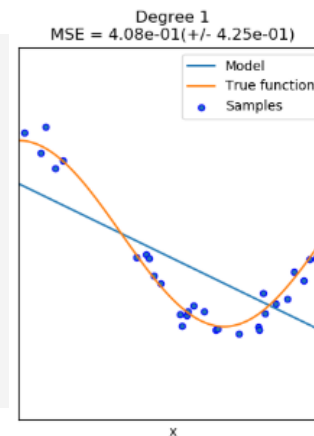
Train Set : 모델을 학습하기 위한 데이터

Validation Set : 학습된 모델의 성능 평가 이전에 검증하기 위한 데이터

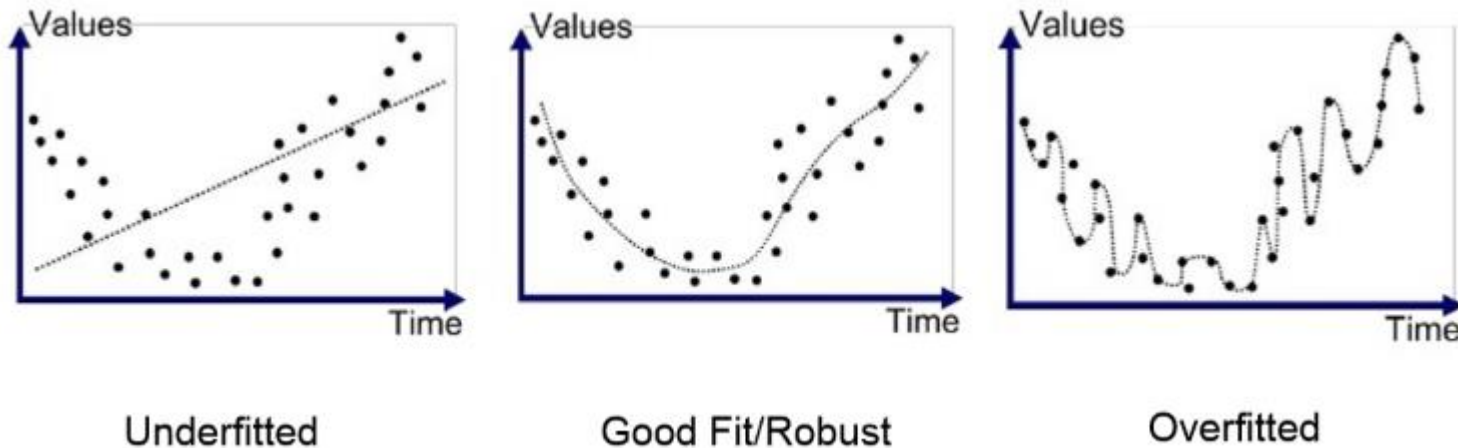
(Cross Validation)을 통해 모델의 generalization performance 및 overfitting 방지, 초모수 조정

Test Set : 학습,검증이 완료된 모델의 최종 성능을 평가하기 위한 데이터

모델이 학습하는 weight가 아닌,
사용자 지정에 따라 달라지는 값



Overfitting (과적합)



과적합: '학습' 데이터에 과하게 적합한 상태가 되어, 모델이 정확한 예측이나 결론을 내릴 수 없는 경우에 발생
어떻게 하면 과적합을 줄일 수 있을까?

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 + \theta_9 x_9 + \dots$$

- 독립 변수의 종류를 줄인다 (불필요한 독립변수 제거)
- 계수 theta의 값을 줄인다 (독립변수의 영향력을 낮춤)

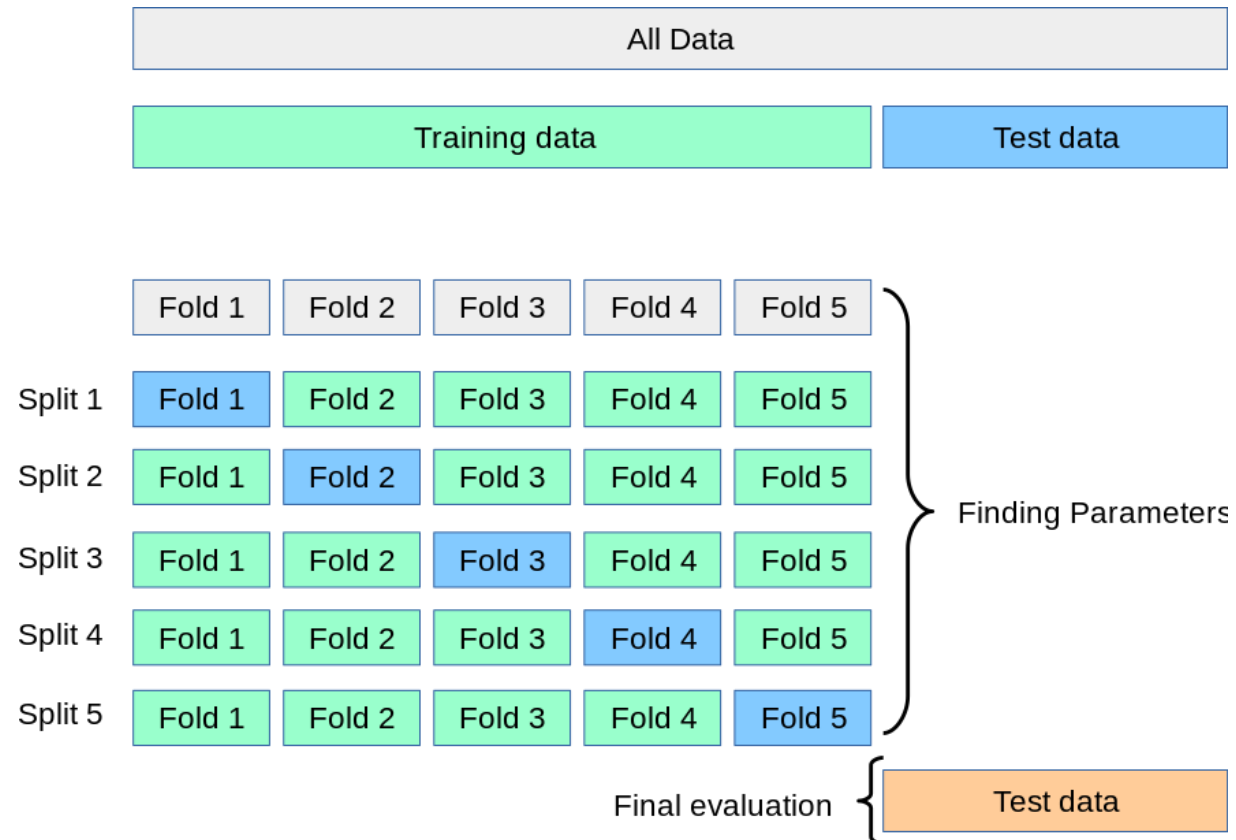
Minimize (MSE

+ $\alpha ||\theta||_1$ Lasso 정규화

+ $\alpha ||\theta||_2^2$ Ridge 정규화

Learning Process

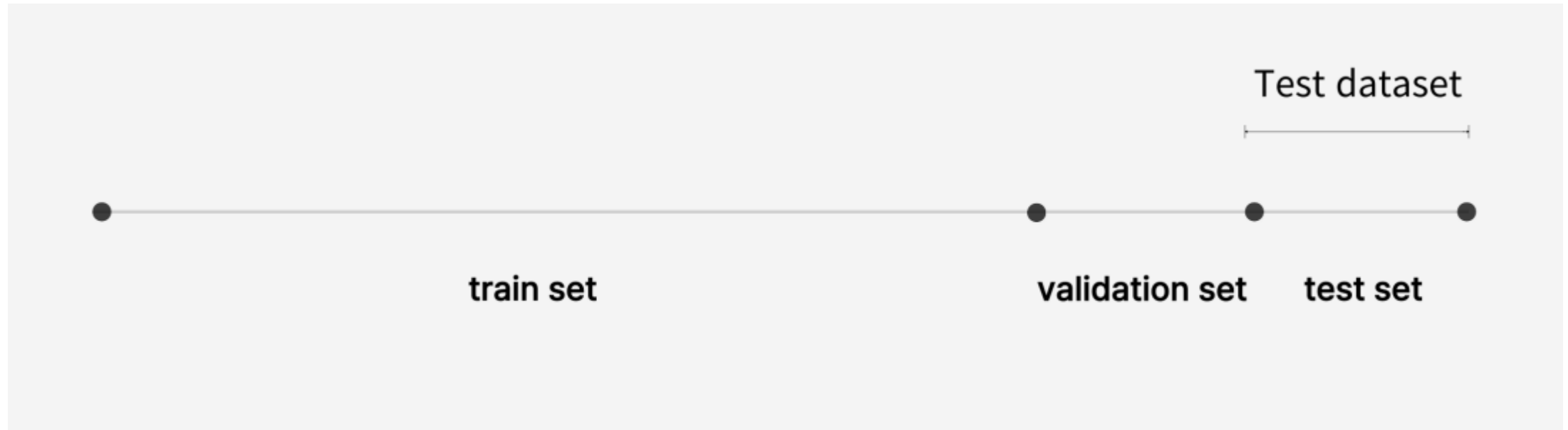
K-Fold Cross Validation



Model Evaluation

Test Set 에 대하여 최종 모델의 성능을 평가

데이터 수 증대, 규제화, 앙상블 등의 방법을 통해 최종성능을 더 향상시킬 수 있다.



평가

Task	Error type	Loss function	Note
Regression	Mean-squared error	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$	Easy to learn but sensitive to outliers (MSE, L2 loss)
	Mean absolute error	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	Robust to outliers but not differentiable (MAE, L1 loss)
Classification	Cross entropy = Log loss	$-\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] =$	Quantify the difference between two probability

주어진 데이터에, 얼마나 꼭 맞는가(=fit 한가?)

R-squared: 결정계수

- 종속 변수의 변동 중 몇 퍼센트를 이 회귀 모형이 설명할 수 있는가?
- 독립 변수를 더 다양하게 투입할 수록 값이 늘어나는 경향 → Adjusted R squared

분류를 평가하는 여러 가지 방법

Confusion Matrix

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Prediction



KUBIG 24-2 Summer ML Session

20기 면접 질문 中 발췌

Q. Regression과 Classification의 차이를 loss 관점에서 설명해주세요.

Q. k-means clustering algorithm에 대해 설명할 수 있나요?

Q. Confusion Matrix이란 무엇이며, 이 행렬을 통해 얻을 수 있는 주요 지표에 대해 설명해주세요.

Q. 머신러닝에서 overfitting을 방지하기 위한 방법들을 소개해 주세요.

Q. 부스팅 모델이 무엇이고, 어떤 것들이 있나요? 그중 하나를 선택하여 모델의 주요 특징에 대해 설명해 주세요.

Q. 로지스틱 회귀 모델을 쓴 이유와, 로지스틱 회귀를 어떻게 해석해야하는지 구체적으로 설명해주세요.

Q. 지도학습과 비지도학습의 차이를 설명해주세요.

Q. 적절한 군집의 개수를 찾는 방법이 팔꿈치 기법이라면, 군집이 잘 되었는지 사후에 평가하는 지표도 있습니다. 무엇인지 아시나요?

Q. 결손값을 처리하셨다고 했는데, 처리 방법과 해당 방법을 선택한 이유를 알려주세요.

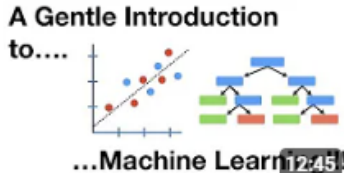
Q. 어떤 상황에 oversampling/undersampling을 실시할까요?

추천 자료

<https://www.youtube.com/@statquest>

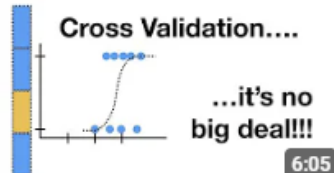
Machine Learning ▶ 모두 재생

Machine Learning covers a lot of topics and this can be intimidating. However, there is no reason to fear, this play list will help you trough it all, one step at a time.



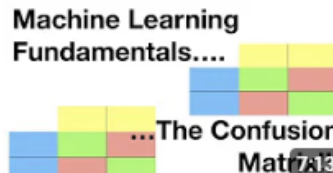
A Gentle Introduction to Machine Learning

StatQuest with Josh Starmer ✓
조회수 99만회 · 5년 전
자막



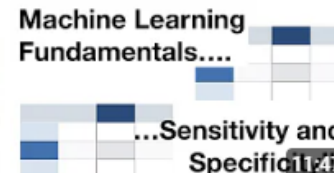
Machine Learning Fundamentals: Cross Validation....
...it's no big deal!!!

StatQuest with Josh Starmer ✓
조회수 108만회 · 6년 전
자막




Machine Learning Fundamentals....
...The Confusion Matrix

StatQuest with Josh Starmer ✓
조회수 66만회 · 5년 전
자막



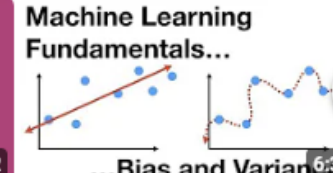
Machine Learning Fundamentals....
...Sensitivity and Specificity

StatQuest with Josh Starmer ✓
조회수 33만회 · 4년 전
자막



The Sensitivity, Specificity, Precision and Recall Sing-a-Long Song!!!
StatQuest!!!

StatQuest with Josh Starmer ✓
조회수 7.5만회 · 2년 전



Machine Learning Fundamentals...
...Bias and Variance

StatQuest with Josh Starmer ✓
조회수 127만회 · 5년 전
자막



고생하셨습니다 :)

해당 자료는 Slack & KUBIG github에서 보실 수 있습니다
1주차는 과제가 없습니다.

