Machine Learning

1장 Machine Learning이란?

고려대학교 통계학과
박유성

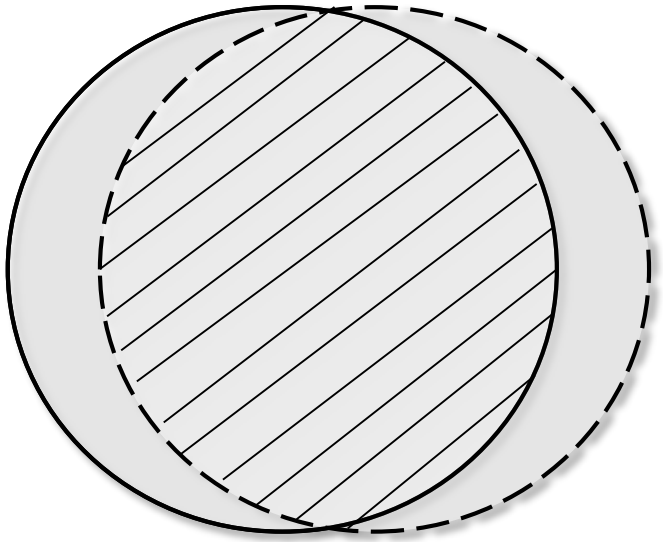
Machine Learning

전통적인 통계학

- 규칙의 통계적 추론에 중점(전문적인 통계적, 수학적 지식)
- 자료의 특성(다변량, 시계열, 범주형 등)에 따라 분석.

통계적 머신러닝

- 규칙의 일반화에 중점
- 목적변수의 관측여부에 따라 지도학습, 비지도학습으로 분석



——— 통계

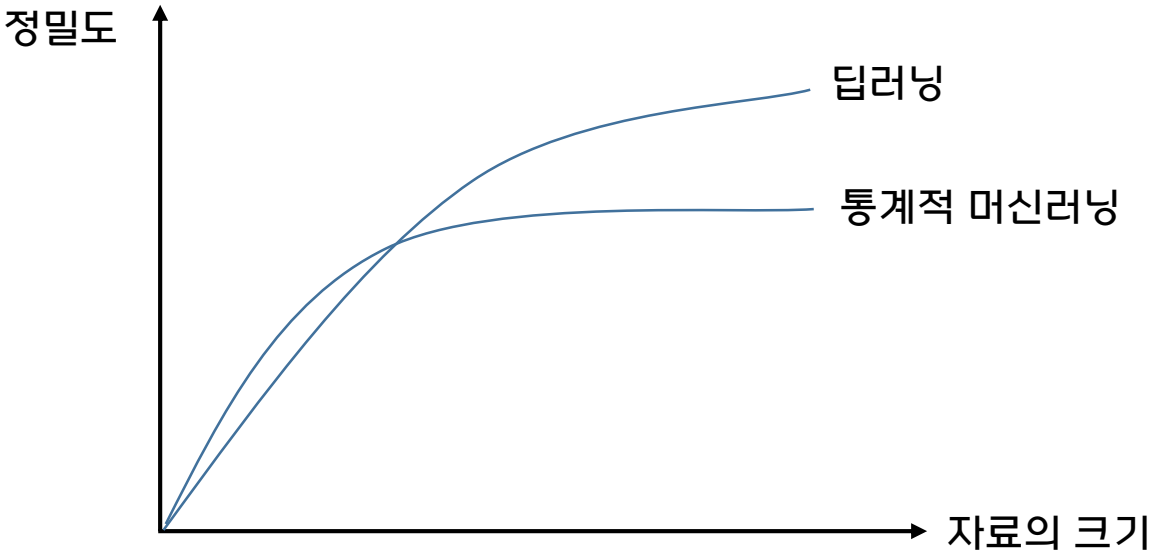
– – – 통계적 머신러닝

통계적 머신러닝과 딥러닝

구분	통계적머신러닝	딥러닝
데이터 크기	중/소 크기	빅데이터
분석자료 형태	2차원 텐서	2차원 텐서이상
강점을 갖는 자료	정형화된 자료	비정형자료
특성변수	특성변수를 만들어야 함	특성변수가 만들어짐
특성변수의 정규화 및 표준화	선택	필요
모형	매우 많음	기본적으로 3 개의 모형
최적화	일반적으로 전체 데이터 사용	배치데이터
해석여부	해석이 쉬움(단, SVM과 boosting 제외)	어렵거나 불가능
하드웨어	중급	고성능(GPU 요구)
실행요구시간	최대 시간 단위	최대 주단위 시간

통계적 머신러닝과 딥러닝

자료의 크기와 정밀도



02 Machine Learning의 분류

- 지도학습(supervised learning)

  비지도학습(unsupervised learning)

  강화학습(Reinforcement learning)

- 배치학습(Batch learning)

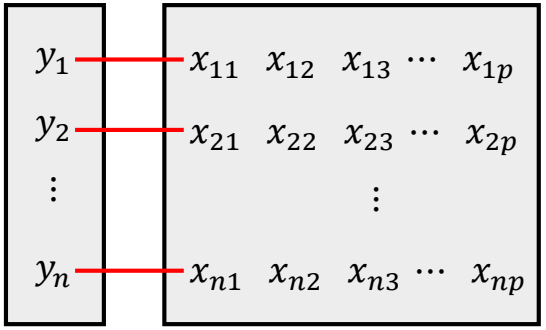
  온라인학습(Online learning)

- 사례기반(Instance-based learning)

  모형기반(Model-based learning)

지도학습(Supervised learning)

- 표식이 있는 자료(labeled data)를 대상으로 함.



- 입력자료 x로부터 표식이 있는 자료 y로의 mapping을 학습하는 방법을 말함. y 의 예측이 주목적임.
- 즉, y의 적합 값이 $\hat{y} = \emptyset(x)$로 표현된다고 할 때 y의 가장 좋은 적합 값을 구해주는 패턴 $\emptyset$을 찾아내는 것이 목적임.
- 표식이 있는 y가 있기 때문에 패턴의 정밀도 점검 가능.

지도학습(Supervised learning)

- y가 범주형인 경우 → 분류(Classification)

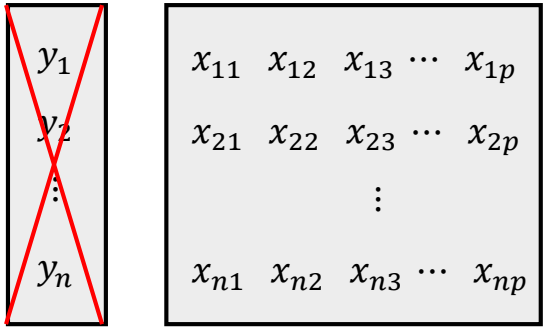
  예) e-mail이 spam인지 아닌지, 우편번호를 구별해내는 handwriting recognition, 사람 얼굴을 구별하는 face detection,
  X-ray 사진을 통해 특정 질병의 유무
  ⋮

- y가 연속형인 경우 → 회귀(Regression)

  예) 내일의 주식 가격 예측, 특정 제품을 선호하는 소비자의 연령 예측
  검진 자료를 이용한 특정 질병 항체의 양 예측
  특정 지역의 기온 예측
  ⋮

비지도학습(unsupervised learning)

- 표식이 없는 자료(unlabeled data)를 대상으로 함.



- 자료의 숨겨진 구조(hidden structure)를 찾고자 함.
- 하지만 표식이 있는 y 가 없기 때문에 찾아낸 structure가 정밀한지에 대한 검증은 불가능함.
- 추출된 요인을 이용해 2차 분석을 할 때 주로 사용함.

비지도학습(unsupervised learning)

- 비지도학습의 대표적인 방법에는

  – 군집분석(Clustering)

  : feature x 변수의 유사성을 이용해 몇 개의 그룹으로 분류

  예) 광고대상자를 고객의 특성을 바탕으로 분류

  – 잠재요인추출(Latent factor extraction)

  : feature x 변수에서 관측되지 않은 잠재요인을 추출하는 방법

  예) 차원축소:PCA (Principal Component Analysis), Embedding, AutoEncoders

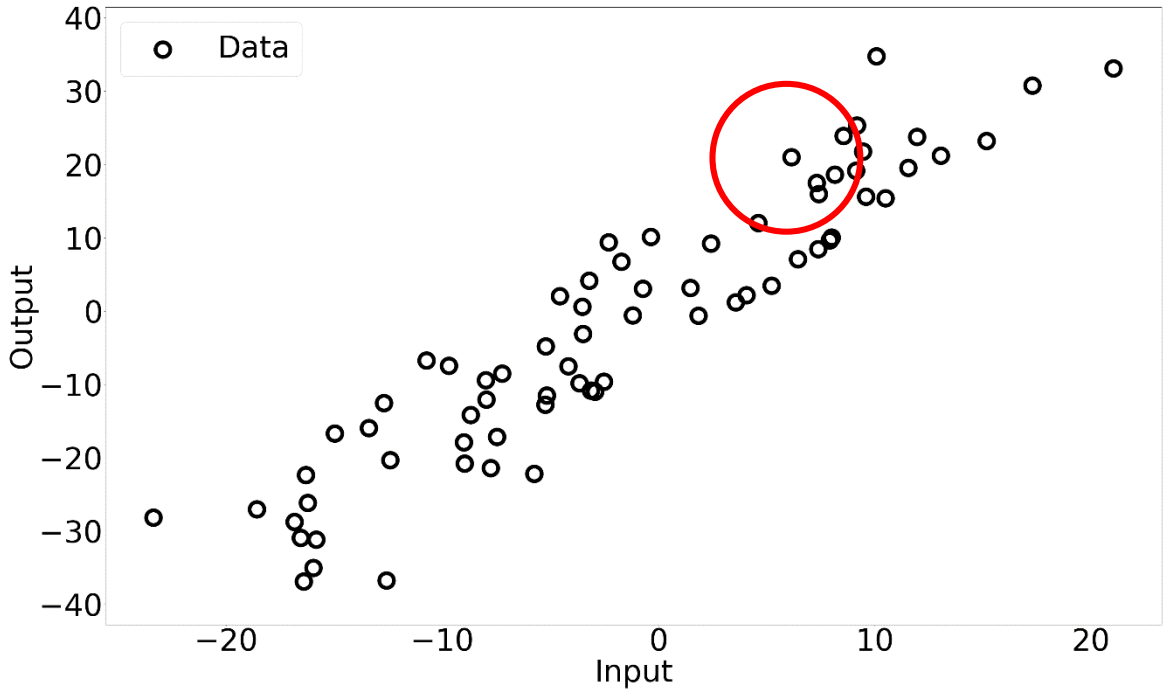
  차원 증대: Kernel method(비선형), Deep learning

강화학습(Reinforcement learning)

- 주어진 환경에 의해 시스템의 성능을 향상시키는 learning algorithm이며 게임이나 로봇공학에서 주로 사용됨.
- 어떤 행동을 했을 때 받는 보상에 따라 미래의 행동을 바꿔가는 강화라는 개념을 기본 아이디어로 함.

사례기반 learning과 모형기반 learning

- 사례기반(instance-based) learning

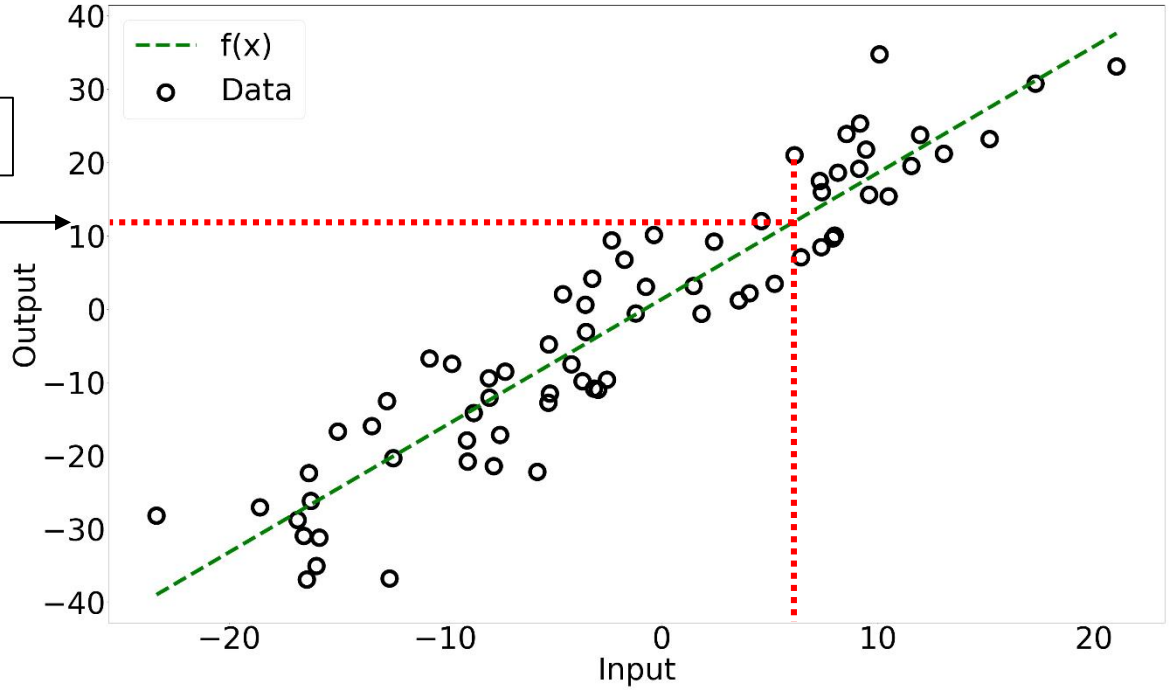


– x 주변의 y의 평균 값을 x 에 대응하는 y의 예측치로 하거나 x 주변에 가장 많이 있는 class로 관측치 x 의 class를 부여하는 learning 방법임.

사례기반 learning과 모형기반 learning

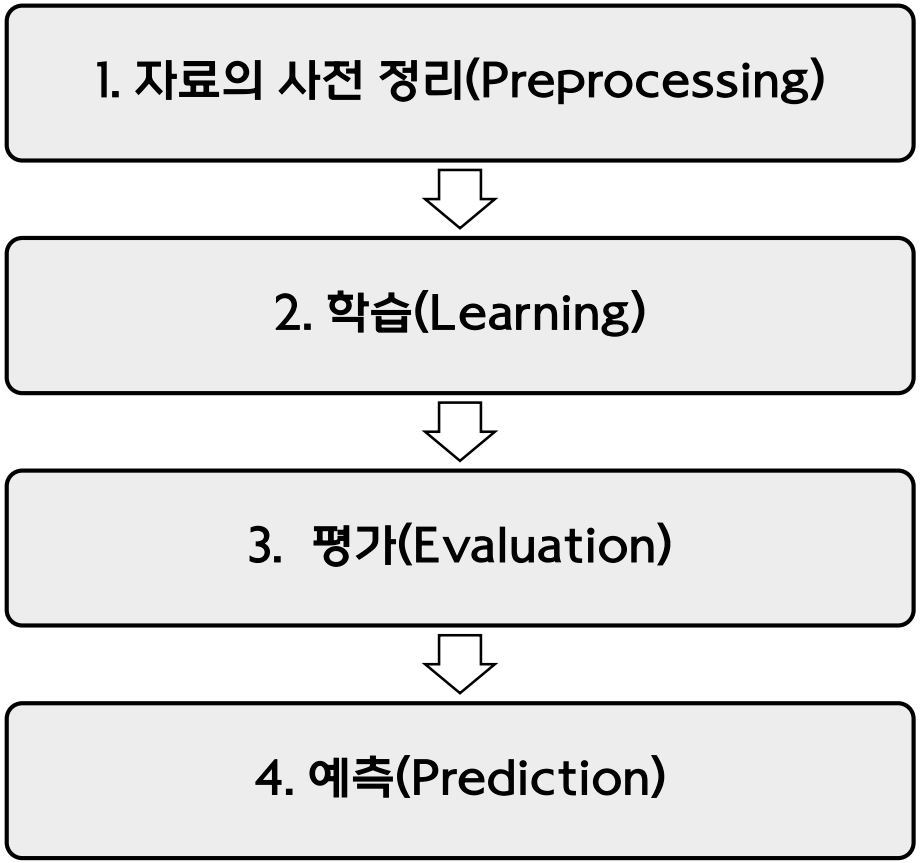
- 모형 기반(model-based) learning

$\hat{y}_i = f(x_i)$



– 어떤 함수 f를 이용해 $f(x)$를 y의 예측치로 하는 learning 방법을 말함.

03 Machine Learning 분석 절차



자료의 사전 정리(preprocessing)

- 주어진 자료를 실수로 구성된 텐서(tensor) 자료로 전환
- 결측치(missing data) 처리 : 대체(imputation) 혹은 제외시킴.
- 이상치(outlier) 처리
- 자료의 표준화(standardization): 딥러닝에서는 정규화 또는 0~1로 rescaling
- One-hot vectorization, Word2Vec, Glove(natural language)
- 불균형자료의 처리: default 차주 사례가 정상차주에 비해 아주 작음.

자료의 사전 정리(preprocessing)

- Training dataset과 Test dataset으로 분할.

  Training dataset에서는 분석 모형의 선택과 모수추정을 하고

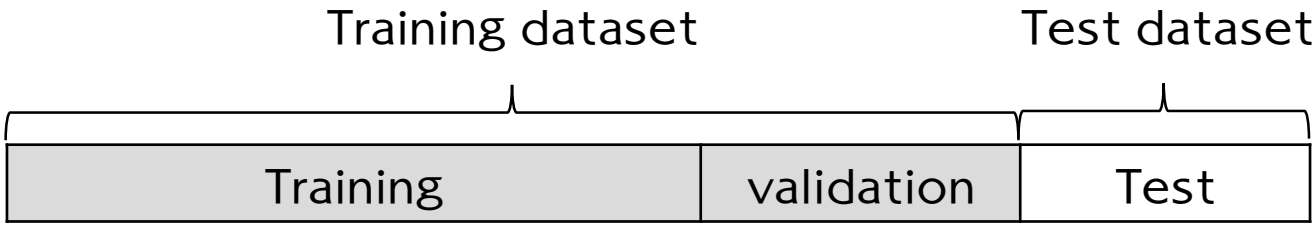
  Test dataset에서는 추정된 모형의 일반화 성능(generalization performance)를 측정하여 모형을 평가하게 됨.

학습(Learning)

- 모형 선택 (통계적 머신러닝의 경우): 통계적 머신러닝은 전통적 통계학의 확률론, 수리통계, 선형모형, 다변량 등의 이론적 분류를, 단순하게 분류와 회귀로 나눔.
  - 분류(Classification)
    : Naïve Bayesian, Logistic analysis, Linear discriminant analysis (LDA), Support vector machine (SVM), Decision tree, K-nearest neighbors (KNN) 등
  - 회귀(Regression)
    : Ordinary least squares, Ridge regression, LASSO, Elastic net linear regression, KNN with decision tree, Kernelized SVM 등

학습(Learning)

- 모형 적합

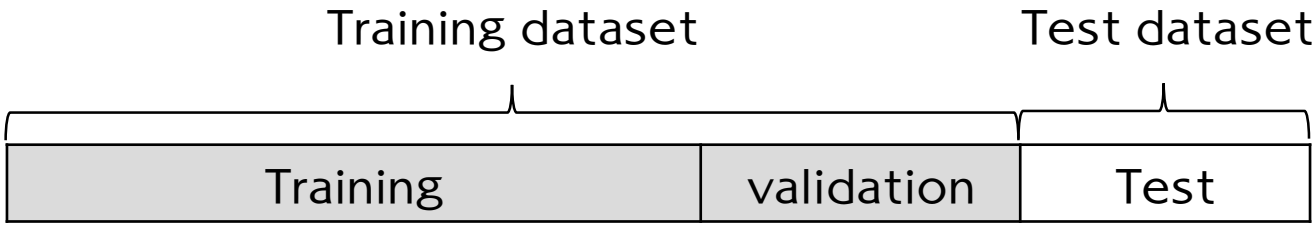


– Training 데이터를 이용하여 참값과 적합값의 거리로 정의되는 손실함수를 계산하고 이 함수에 대한 모수의 미분 값에 아주 작은 값을 곱한다.

– 전 단계의 모수값에 앞에서 산출된 값을 차감하여 모수를 update한다. 이를 gradient descent라고 한다.

학습(Learning)

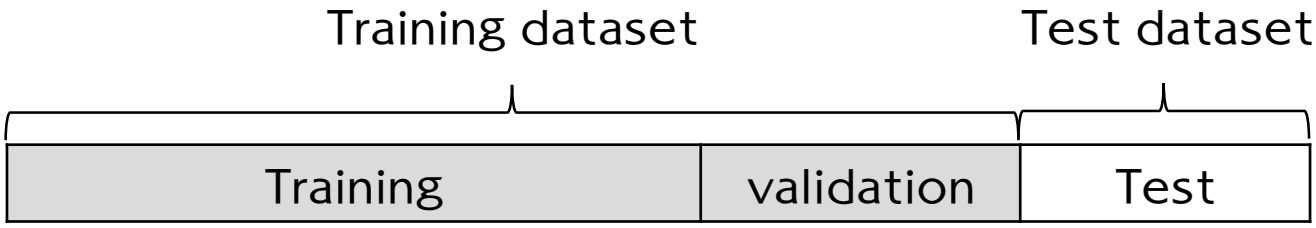
- 모형 적합



  - Deep learning에서는 input data에 대해 weights(모수)를 곱한 후 이를 비선형으로 전환하고 이 값에 또 다른 weights를 취하는 과정을 반복하므로 초기 layer에 있는 weight를 최신화하기 위해
  - 손실함수에 가장 가까운 weight로부터 chain rule의 반대 방향으로 미분을 순차적으로 구해 원하는 weight의 미분값을 구하게 된다. 이를 backpropagation이라고 한다.

학습(Learning)

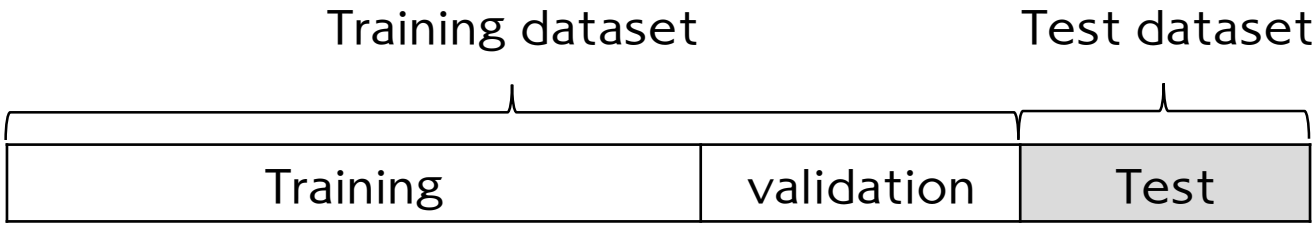
- 모형 적합



  - 이를 학습(learning)이라고 하며 손실함수가 최소가 되도록 위의 과정을 반복한다. 그러므로 learning 알고리즘은 머신러닝의 엔진과 같다.
  - 딥러닝에서는 layer나 layer안에 있는 노드수를 늘려 손실함수 값을 0으로 접근시킬 수 있다.
  - Validation 데이터는 초모수를 선택하게 하며 과대적합(overfitting)문제를 파악할 수 있도록 한다.

평가(Evaluation)와 예측(Prediction)

- 머신러닝은 과대적합(over-fitting)이 발생할 가능성이 높음.



- 앞선 일련의 과정을 통해 선택된 모형을 학습에 이용하지 않았던 test dataset에 적합 시켜 generalization error를 계산하여 over-fitting(또는 under-fitting) 여부를 평가함.

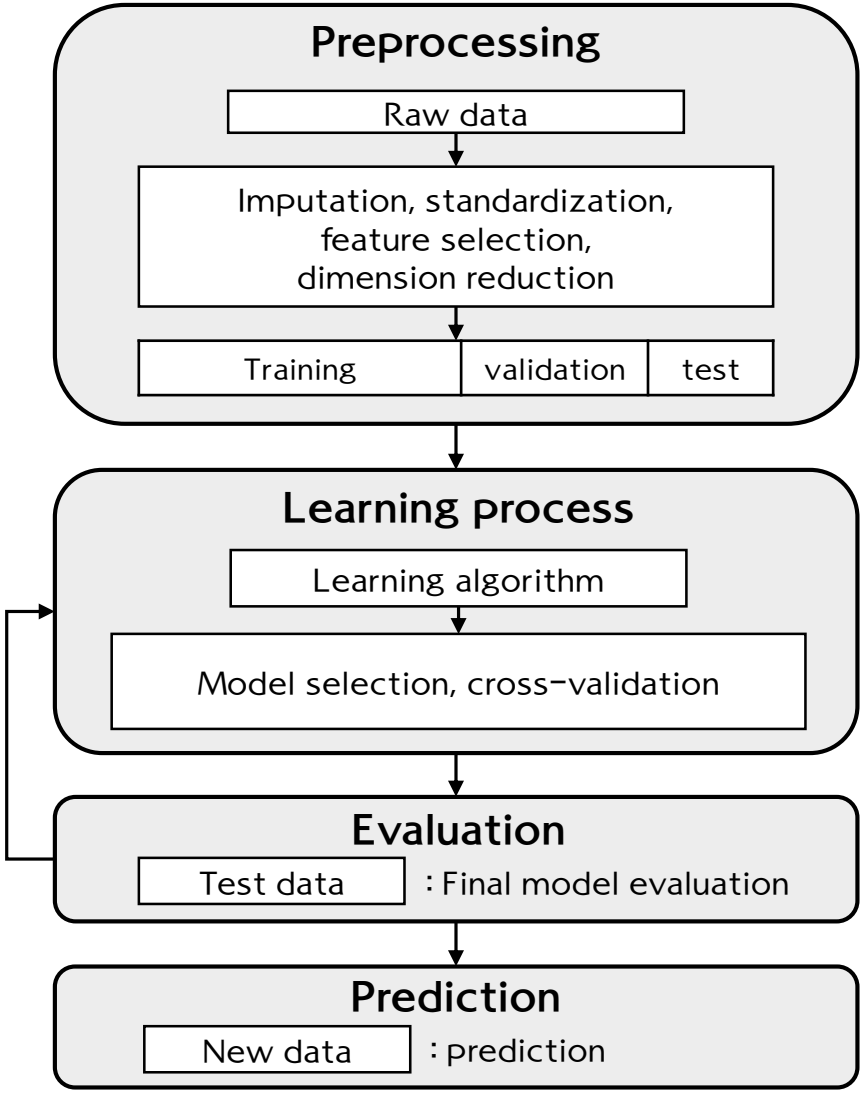
평가(Evaluation)와 예측(Prediction)

- 과대적합의 해결책

  통계적 머신러닝: L1 또는 L2 규제화

  딥러닝: L1, L2 규제화, batch normalization, dropout, downsizing

Machine learning 분석 절차의 도식화



04 Machine Learning 방법들

- 목적
  - 분류(Classification) / 비선형(Nonlinear) 분류
  - 회귀(Regression) / 비선형(Nonlinear) 회귀 / 로버스트(Robust) 회귀
  - 군집(Clustering)
  - 차원 축소(Dimension reduction) / 비선형(Nonlinear) 차원 축소
  - 앙상블(Ensemble)
  - 문서 분석(Documents analysis)

Machine learning 방법별 목적 및 구분

learning	목적	구분
K-nearest neighbors (KNN)	분류, 회귀(3장)	지도학습, 사례기반, 배치
Kernel smoothing	density estimation(3장)	
Adaptive linear neuron	분류(4장)	지도학습, 모형기반, 배치
Logistic regression	분류(4장)	지도학습, 모형기반, 배치, online
Discriminant analysis	분류(5장)	지도학습, 모형기반, 배치
Naive Bayes	분류(5장)	지도학습, 모형기반, 배치
Classification and Regression Tree (CART)	분류, 회귀(6장)	지도학습, 배치, 비모수
Support vector machine (SVM)	분류(7장), 회귀(10장)	지도학습, 모형기반, 배치, online
Kernelized SVM (kernel trick)	비선형분류(7장), 비선형회귀(10장)	지도학습, 모형기반, 배치, online
Principal component analysis (PCA)	차원축소(8장)	비지도학습, 모형기반, 배치
Kernelized PCA	비선형 차원축소(8장)	비지도학습, 모형기반, 배치
Linear discriminant analysis (LDA)	차원축소(8장)	비지도학습, 모형기반, 배치
Regression (OLS)	회귀(10장)	지도학습, 모형기반, 배치, online
RANSAC	로버스트 회귀(10장)	지도학습, 모형기반, 배치

Machine learning 방법 별 목적 및 구분

learning	목적	구분
Bagging	분류, Ensemble(11장)	지도학습, 모형기반, 배치
Boosting, XGboost	분류, 회귀, Ensemble(11장)	지도학습, 모형기반, 배치
Random forest	분류, 회귀, Ensemble(11장)	지도학습, 모형기반, 배치
K-means clustering	군집(12장)	비지도학습, 사례기반, 배치
Hierarchical clustering	군집(12장)	비지도학습, 사례기반, 배치
DBSCAN	군집(12장)	비지도학습, 사례기반, 배치
Sentiment analysis	분류, 회귀, 문서분석(13장)	지도학습, 모형기반, 배치, online
Multilayer Neural Network/backpropagation	딥러닝의 기초이론	지도학습, 모형기반, 온라인
Convolutional Neural Network	비정형데이터(이미지, 텍스트, 오디오, 음성)	지도학습, 모형기반,온라인
Recurrent Neural Network/LSTM	자연어 처리(언어번역, 감성분석, 고객서비스 자동화, 웹 검색)	지도학습, 모형기반, 온라인

모형 진단과 초모수 선택을 위한 주제와 그에 따른 목적

주제	목적
L_1, L_2 규제화	과대적합 방지, 특성변수 선택(4장)
Gradient decent	손실함수 최소(2장)
Cross-validation (CV)	과대 및 과소 적합 진단, 초모수 선택(9장)
Nested cross-validation	machine learning의 최종 성능 점검(9장)
Grid search CV	CV를 이용한 최적 초모수의 선택(9장, 13장)
pipeline	machine learning을 위한 사전자료 정리의 통합처리(9장)

Q & A