Statistical Machine Learning

2주차

담당: 13기 박주영

End-to-End Machine Learning Project

CONTENTS

머신러닝 분석 절차

Data Preprocessing

Learning Process

Model Evaluation

Prediction

01. Data Preprocessing

원자료를 학습 목적에 맞게 가공

보다 높은 정확성을 갖는 분석을 위해 원자료에 대해 전환 및 가공을 거치는 단계
→ AutoML의 등장으로 그 중요도 및 비중이 커지고 있음

정규화와 표준화

특성변수의 단위 등에서 나타나는 차이를 조정해주는 역할

One-hot encoding

범주형 변수를 수치형 변수로 변환

bag of words

텍스트를 수치형 변수로 변환해주는 방법(NLP)

차원축소: PCA / t-sne

feature가 너무 많으면 오버피팅 가능성이 있기 때문에 차원축소 필요

이상치 및 결측치 처리

머신러닝 모형은 직접 결측치를 처리할 수 없음

불균형 자료 처리: SMOTE / ADASYN

불균형 자료 문제를 해소하기 위한 과대표집 방법

01. Data Preprocessing

| PCA와 t-sne

PCA	t-sne
Principal Component Analysis	t-distributed stochastic neighbourhood embedding
linear dimensionality reduction technique	non-linear Dimensionality reduction technique
reduce the dimensionality of data that is highly correlated by transforming the original set of vectors to a new set	minimize the Kullback–Leibler divergence (KL divergence) between the two distributions

01. Data Preprocessing

| SMOTE와 ADASYN

SMOTE	ADASYN
Oversampling	Oversampling
K- nearest neighbors	SMOTE의 개선된 버전
$x_{syn} = x_i + \lambda\ (x_k - x_i),\ x_k \in S_i$	표본 수에 대한 weight를 통해 추출

02. Learning Process

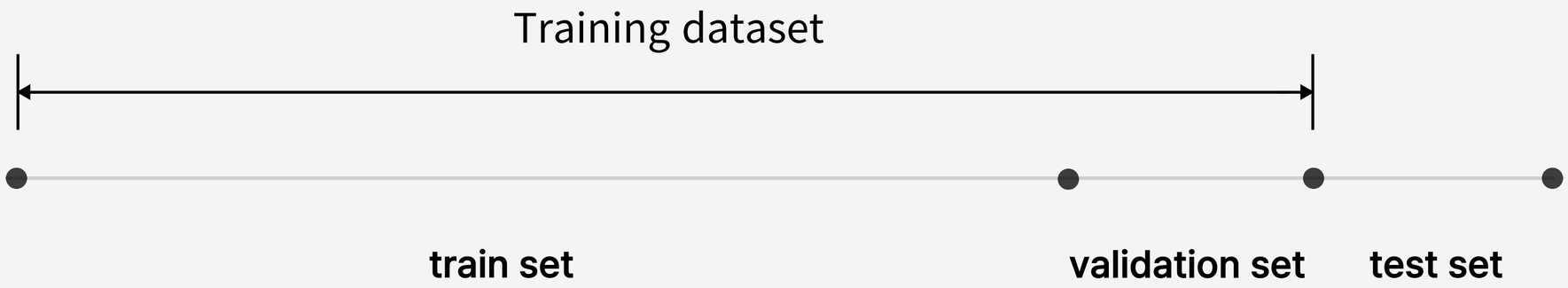
손실함수 정의 후 이를 최소화하는 모수를 추정하는 최적화 과정

learning	목적	구분
K-nearest neighbors (KNN)	분류, 회귀(4장)	지도학습, 사례기반, 배치
Kernel smoothing	density estimation(4장)	
Adaptive linear neuron	분류(5장)	지도학습, 모형기반, 배치
Logistic regression	분류(5장)	지도학습, 모형기반, 배치, online
Discriminant analysis (다변량)	분류(6장)	지도학습, 모형기반, 배치
Naive Bayes	분류(6장)	지도학습, 모형기반, 배치
Classification and Regression Tree (CART)	분류, 회귀(7장)	지도학습, 배치, 비모수
Support vector machine (SVM)	분류(8장), 회귀(11장)	지도학습, 모형기반, 배치, online
Kernelized SVM (kernel trick)	비선형분류(8장), 비선형회귀(11장)	지도학습, 모형기반, 배치, online
Principal component analysis (PCA)	차원축소(9장)	비지도학습, 모형기반, 배치
Kernelized PCA	비선형 차원축소(9장)	비지도학습, 모형기반, 배치
Linear discriminant analysis (LDA), MDS, Manifold Learning, t-SNE	차원축소(9장)	비지도학습, 모형기반, 배치
Regression (OLS)	회귀(11장)	지도학습, 모형기반, 배치, online
RANSAC → outlier에 민감하지 않은 회귀방법	로버스트 회귀(11장)	지도학습, 모형기반, 배치

learning	목적	구분
Bagging	분류, Ensemble(12장)	지도학습, 모형기반, 배치
Boosting, Random forest	분류, 회귀, Ensemble(12장) (= bootstrap)	지도학습, 모형기반, 배치
Xgboost, LightGBM, Catboost	분류, 회귀, Ensemble(12장)	지도학습, 모형기반, 배치
K-means clustering	군집(13장)	비지도학습, 사례기반, 배치
Hierarchical clustering	군집(13장)	비지도학습, 사례기반, 배치
DBSCAN, HDBSCAN	군집(13장)	비지도학습, 사례기반, 배치
Sentiment analysis	분류, 회귀, 문서분석(14장)	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

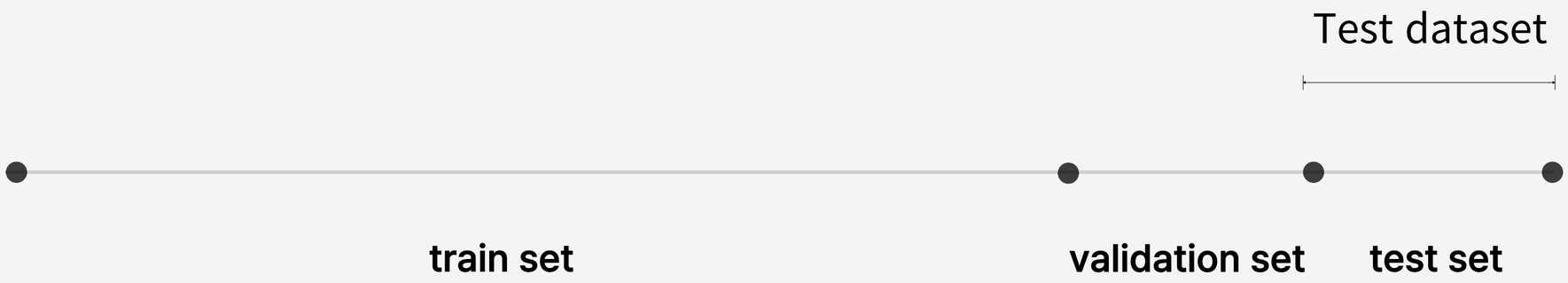
02. Learning Process

학습데이터를 이용해 학습된 모형을 검증데이터에 적합시켜 학습데이터에서 구현된 성능과 비슷한 수준인지 점검하고 초모수 조정



03. Model Evaluation

학습된 모형을 시험데이터에 적용시켜 성능 평가
: 과대적합 해결 → 자료크기 증대, 규제화, 앙상블 등



04. Prediction

일반화된 모형에 대해 예측 진행

Objective

Regresion

지도학습: y가 연속형

Classification

지도학습: y가 범주형

고생하셨습니다

이어서 구글코랩을 통한 실습을 진행하겠습니다.