The Great Data Preprocessing Battle

https://product.kyobobook.co.kr/detail/S000001810192

데이터

데이터 레코드
멀티미디어 데이터
그래프 데이터

마스터 데이터
인덱스 데이터

머신러닝

지도학습
비지도 학습

지도 데이터
학습 데이터

테스트 데이터
적용 데이터

지표, 표, 그래프 작성용 전처리
지도학습용 전처리

비지도학습용 전처리

데이터 구조 대상의 전처리
데이터 내용 대상의 전처리

SQL
R → tidyverse 패키지
Python

데이터 열을 지정하여 추출
조건에 따른 데이터 행의 추출
데이터 값을 고려하지 않는 샘플링
집약 ID 에 기반한 샘플링

집계와 유니크 카운트, 합계, 최대와 최소, 대표값, 분포, 최빈값, 순위

마스터 테이블 결합

데이터 레코드 분할
교차 검증 ~ 과학습 배제
홀드아웃 검증

언더샘플링
오버샘플링 : SMTOE

희소 행렬

수치형 데이터 - 데이터 크기가 작고 가공에 용이

대수화

범주화

정규화

주성분 분석

수치 보완

: MCAR 우연히 발생한 완전 무작위 결손

: MAR 결손된 항목 데이터와 관계 없이 다른 항목 데이터에 의존한 결손

: MNAR 결손된 항목 데이터에 의존한 결손

결손 레코드 제거

정수 보완 - 제조 레코드 활용

평균값 보완

PMM 을 이용한 다중 대입

범주형 데이터 - 비선형한 변화 표현 가느앟나 대량의 데이터 필요

더미 변수화

Timstamp, DateTime ~ UNIXTIME

~ 일시, 시간대, 계절형, 평일과 휴일

문자형

언어 의존 분석

언어 비의존 분석 : N-gram

TF-IDF

위치 정보형

한국 측지계와 세계 측지계

- R : sp 패키지의 Spatial 오브젝트 활용
- Pyhton : pyproj 라이브러리를 사용하여 측지계

두 지점의 거리와 방향 계산

- R : geosphere 패키지
- Python : pyproj.Geod 오브젝트의 inv 함수 + geopy 라이브러리 활용