

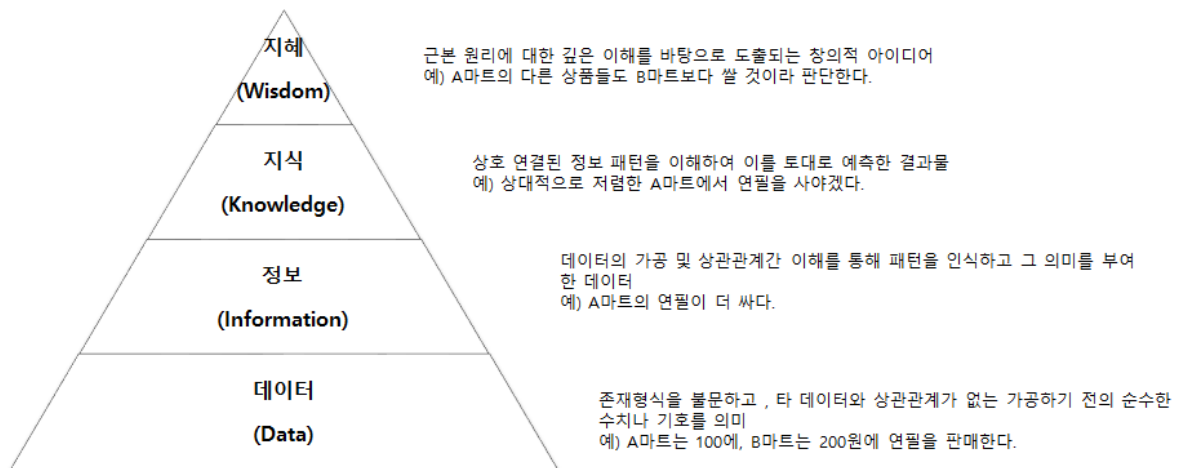


빅데이터 분석기사

빅데이터 특징

1. 데이터 정의 - 현실 세계로부터 관찰되거나 측정되어 수집된 사실 또는 값
의미 있는 정보를 가진 모든 값
객관적 사실을 의미(당위적 특징)

2. DIKW 피라미드



*지식의 종류

암묵지 - 개인에게 축적된 내면화된 지식 (ex.노하우)

형식지 - 언어나 문자로 표준화 및 형상화된 지식 (ex.교재)

암묵지와 형식지의 상호작용



공통화 (암묵지 → 암묵지) : 개인 혹은 집단의 경험 공유

표준화 (암묵지 → 형식지) : 개인의 지식을 공유하기 위해 문자나 매체로 표현

내면화 (형식지 → 암묵지) : 교육, 훈련등으로 형식지를 개인이 취득

연결화 (형식지 → 형식지) : 개인 혹은 집단의 형식지를 상호 결합

3.빅데이터 정의

일반적으로 관리할 수 있는 규모를 넘어선 데이터를 의미

가트너 - 높은 시사점을 제공하고 더 좋은 의사결정을 위해 사용하는 정보자산

매킨지 - 일반적인 DB가 저장,관리,분석 가능한 범위를 넘어서는 규모의 데이터

IDC - 다양하고 대규모의 데이터에서 저비용으로 가치를 추출하고 데이터의 수집과 발굴을 초고속으로 처리할 수 있는 차세대 기술 및 아키텍처

빅데이터의 특징

V3 - Volume(크기), Variety(다양성), Velocity(속도)

V5 - Veracity(신뢰성(노이즈 제거)), Value(가치)

빅데이터가 만드는 변화

사전처리 → 사후처리 : 필요한 데이터만 수집 → 가능한 많은 데이터를 수집 후 조합

표본조사 → 전수조사 : 데이터 처리 비용감소 → Sampling 활용

질 → 양 : 데이터가 지속적으로 추가될 경우 양질의 정보가 증가

인과관계 → 상관관계 : 상관관계 분석이 미래 예측을 압도

빅데이터 사회적 현상 변화

산업 구조가 디버전스 패러다임으로 변화
생산 → 시장창조로 변화
제품 생산 → 서비스 중심

빅데이터 가치

1. 빅데이터 활용 및 가치
2. 빅데이터 가치 선정 - 다양한 활용방식, 새로운 가치, 기술발전으로 가치 선정이 어려움
3. 빅데이터의 영향
 - 기업 - 비즈니스 모델 혁신
 - 정부 - 미래 대응 전략 수립
 - 개인 - 기회비용 절약(생활의 스마트화)
4. 빅데이터 위기요인과 통제방안
 - 사생활 침해 - 동의제 → 책임제
 - 책임 훼손 원칙 - 기존 책임 원칙 강화
 - 데이터 오용 - 분석 알고리즘 접근 허용

데이터 산업의 이해

1. 데이터 산업과 분석 인사이트

제조 : 제품 생산 분량률 개선(공급망 최적화, 수요 예측, 재고 보충 등)
쇼핑 : 구매 분석을 통한 소비 예측
물류 : 물류 관리 최적화
의료 : 헬스케어 서비스를 통한 의료 복지 향상
정보 : 사기탐지, 사례관리, 범죄방지, 수익 최적화
커뮤니케이션 : 가격 계획 최적화, 고객 보유, 수요 예측
에너지 : 트레이딩, 공급 및 수요 예측

2. 데이터 사이언스

다양한 유형의 데이터로부터 의미 있는 정보를 추출하는 분야

3. 데이터 사이언스의 구성요소

분석 + IT + 비즈니스 분석(Hard Skill - 분석 기술, 지식 / Soft Skill - 대화, 통찰력 있는 분석)

4. 기업 관리 시스템

CRM(고객관리시스템) : 기업이 고객 내 외부 데이터 통합관리 및 분석을 통한 마케팅 활

동

SCM(공급망관리) : 외부 공급업체부터의 자재 구매관리에서부터 생산, 재고, 고객관리, 판매

빅데이터 조직 및 인력

1. 빅데이터 업무 프로세스

빅데이터 도입 단계 - 도입 기획, 기술검토, 도입 및 조직 구성, 예산 확보

빅데이터 구축 단계 - 요구사항 분석, 설계, 구현, test 진행

빅데이터 운영 단계 - 빅데이터 플랫폼, 분석 모델, 운영 조직, 운영 예산

2. 빅데이터 조직의 구성

데이터 분석의 가치를 발굴하고 이를 활용하여 비즈니스를 최적화하는 목표 구상

조직 목표를 달성하기 위하여 업무를 나눠서 수행

수직적 분할은 계획, 감독, 업무 수준에 따라 나눈다.

분업화는 수평적 분할과 수직적 분할로 나눌 수 있다.

3. 빅데이터 조직 구조

집중구조 - 부서별 분석 진행 & 전담 분석팀 보유(이원화, 이중화 가능성 존재)

기능구조 - 부서별 분석 진행(핵심 분석의 한계점이 명확)

분산구조 - 분산 조직을 협업부서로 배치

4. 빅데이터 조직의 직무별 역량 모델링 개발 단계

조직의 비전 → 행동 특성 도출 → 직무별 역량 도출 → 역량 모델 확정

빅데이터 기술

1. 데이터 수집

2. 데이터 저장 : 수집된 데이터를 목적에 맞는 형태로 저장

3. 데이터 분석 : 통계분석, 머신러닝, 딥러닝, 데이터 마이닝

4. 데이터 활용 : 시각화, 분석 리포트, 응용 프로그램 연계

*데이터 마이닝 : 현상 분석 / 머신러닝 : 현상 파악 후 미래 예측

빅데이터 플랫폼

데이터의 수집, 저장, 분석, 활용 등 분석 프로세스를 지원하는 구격화된 빅데이터 프로세스 기술

1. 방대하고 복잡한 데이터를 처리하기 위해 다양한 빅데이터 플랫폼이 개발되었다.
2. 오픈 소스 기반의 분산처리 환경에는 하둡 분석플랫폼이 있다.

빅데이터 에코시스템

특정 기술이나 솔루션에 국한되지 않고 수집, 변환, 적재, 분석, 시각화 여러 단계를 거치면서 사용되는 여러 가지 기술을 이용해 플랫폼을 구축하는 과정을 의미

빅데이터 분석 프로세스

수집 → 저장 및 관리 → 처리 → 분석 → 활용 → 폐기

개인정보 법

1. 개인식별 정보는 비식별화 조치 후 사용 가능
2. 빅데이터 처리 사실, 목적 등의 공개를 통한 투명화
3. 개인정보 재식별 시, 즉시 파기 및 비식별화 조치
4. 민감정보 및 통신 비밀 수집, 이용, 분석 등 처리금지
5. 수집된 정보의 저장, 관리 시 기술적 관리 보호조치 시행

개인정보 비식별 조치 가이드라인

데이터 활용이 증가함에 따라 개인정보 보호 강화에 대한 요구가 지속되어 개인정보 보호를 보장하면서 데이터를 활용하기 위해 만들어졌으며, 개인정보를 이용 또는 제공할 때 준수해야 할 조치

사전검토 → 비식별 조치 → 적정성 평가 → 사후관리

1. 가명처리 - 대체값 적용 (ex. 홍길동)
2. 총계처리 - 개인정보에 통계값을 적용
총합, 부분합, 라운딩, 재배열
3. 데이터 삭제 - 특정 데이터 값 삭제

4.데이터 범주화 - 그룹의 대표값, 구간값 변환 (ex.30대)

5.데이터 마스킹 - 전체 또는 부분을 대체값 적용

임의 잡음 추가, 공백 대체

데이터 분석 계획

분석 로드맵 설정

데이터 분석 체계 도입

데이터 분석 유효성 검증

데이터 분석 활용 및 고도화

분석 문제 정의

하향식 접근 방식 : 문제 정의 가능, 해결 방법을 찾기 위해 단계적으로 업무를 수행하는 방식

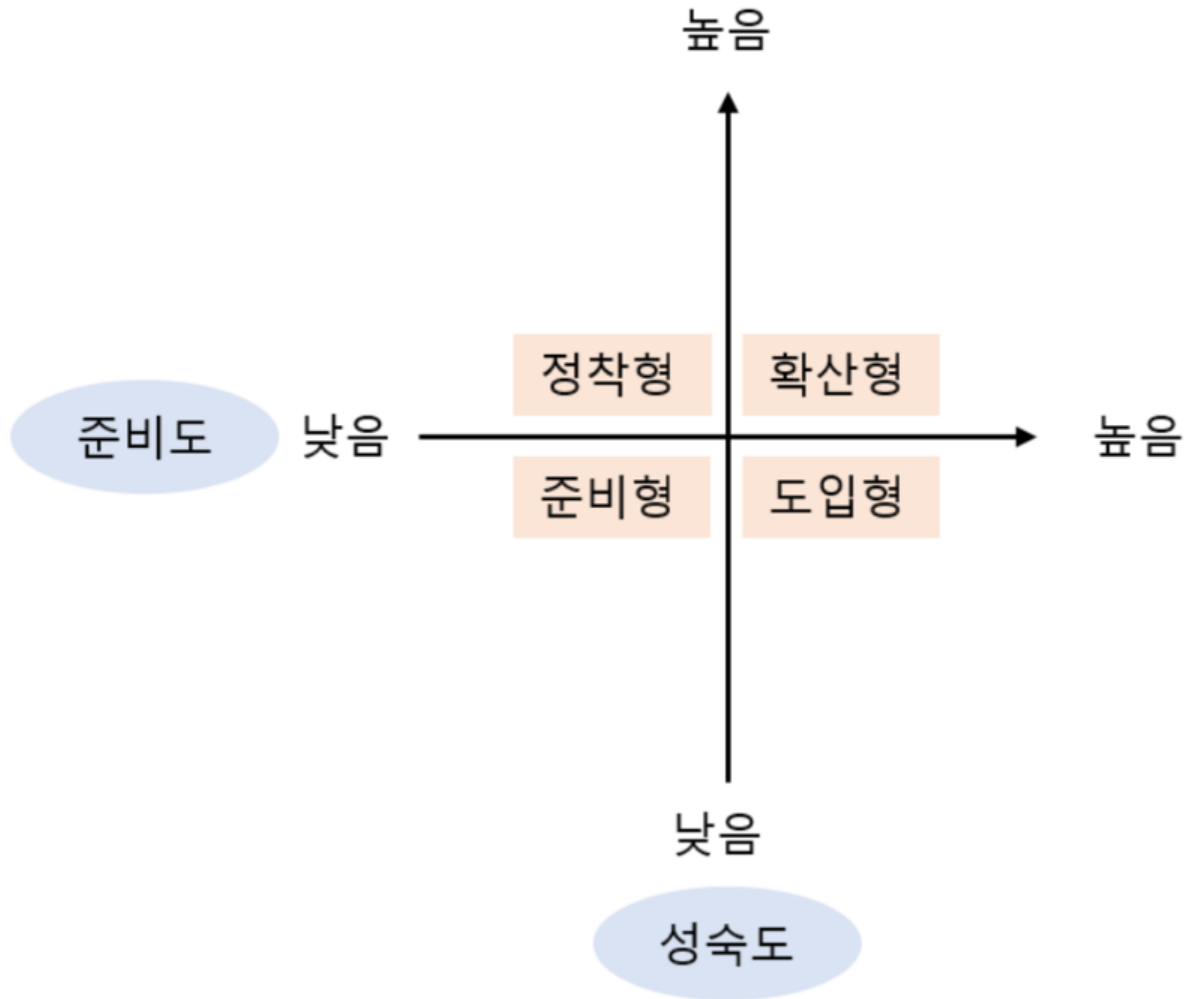
문제 탐색 → 분석 문제 정의 → 해결 방안 탐색 → 타당성 평가 및 과제 선정

상향식 접근 방식 : 문제 정의 불가, 데이터 기반으로 문제 정의 및 해결 방안 탐색

문제 해결 방안

Aa 구분	::: 분석 대상	::: 분석 방법
<u>최적화</u>	Known	Known
<u>솔루션</u>	Known	Un-Known
<u>통찰</u>	Un-Known	Known
<u>발견</u>	Un-Known	Un-Known

빅데이터 분석 준비도 및 성숙도



데이터 분석 방안

1. 데이터 분석 업무 흐름

데이터 수집

데이터 저장 : 수집된 데이터를 분석 아키텍처에 저장

데이터 처리 : 전처리와 후처리를 통해 데이터를 분석 환경과 목적에 적합하게 가공

EDA : 데이터 현황을 분포도, 평균과 분산 등 간단한 시각화

모형 및 알고리즘 설계 : 데이터 특성에 맞는 분석 모형과 알고리즘 설계

시각화 및 데이터 활용 : 분석 결과 이해를 위해 시각화

2. 빅데이터 분석 방법론

분석 기획 : 비즈니스 분석 및 문제 정의

데이터 탐색 : 분석 데이터 정의, 데이터 저장 설계

데이터 분석 : 분석 데이터 준비, 텍스트 분석
시스템 구현 : 설계 및 구현, 시스템 테스트 및 운영
평가 및 전개 : 모델 발전 계획 수립, 프로젝트 평가 보고

3.추가 분석 방법론 유형

KDD 분석 방법론

데이터 선택 → 전처리 → 변환 → 데이터 마이닝 → 해석과 평가

CRISP-DM

업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개

SEMMA - **프로파일링** 기법을 기반으로 한 방법론

샘플링 → 탐색 → 수정 → 모델링 → 검증

데이터 확보 계획

1.데이터 확보

2.데이터 확보 계획 수립

목표 정의 → 요구사항 도출 → 예산안 수립 → 계획 수립

빅데이터 분석 절차

문제 인식 → 현황 분석 → 모형화 → 데이터 수집 → 데이터 분석 → 분석 결과 활용

빅데이터 작업 계획

WBS

1.데이터 분석 과제 정의

2.데이터 준비 및 탐색

3.데이터 분석 모델링 및 검증

4.산출물 정리

데이터 수집 및 저장 계획

데이터 수집

1.데이터 수집 프로세스

수집 대상 선정 → 데이터 수집 세부계획 수립 → 테스트 수집

2.데이터 수집 기술

정형 데이터 - ETL, Open API, FTP

비정형 데이터 - 크롤링, Open API, RSS, FTP, Kafka, Scrapy

반정형 데이터 - Sensing, Streaming, Flume, Scribe, Chukwa

*반정형 데이터는 구조에 따른 메타데이터를 포함한다.

3.데이터 유형

구조 - 정형, 반정형, 비정형

존재형태 - 실시간 데이터, 비실시간 데이터

저장형태 - 파일 데이터, DB 데이터, 콘텐츠 데이터, 스트림 데이터

데이터 속성 파악

1.범주형 데이터 (명목척도, 서열척도, 등간척도(절대영점))

명목형 - 변수의 크기가 순서와 상관없이 의미만 구분

순서형 - 변수의 값이 기준에 따라 순서를 의미

2.수치형 데이터

이산형 - 수치로 측정되는 변수(연산가능)

연속형 - 변수가 구간 안의 모든 값을 가질 경우

데이터 처리 기술

1.데이터 필터링

오류 탐색, 보정, 삭제 및 중복 확인

2.데이터 변환

평활화 - 노이즈 제거, 군집화

집계 - 다양한 차원으로 데이터를 요약

일반화 - 특정 구간으로 값을 스케일링

정규화 - 정해진 구간으로 전환

속성 생성 - 새로운 속성 값을 생성하는 기법

3.데이터 정제

결측치 대체

4.데이터 통합

5.데이터 축소

불필요한 항목 제거.

데이터 비식별화

1.데이터 보안 관리

사용자 인증 - ID와 비밀번호 인증

접근 제어 - 권한 부여 및 확인

암호화 - 암호화 알고리즘을 통해 데이터를 변경

개인정보 비식별화

개인정보 암호화 - 개인정보 데이터를 암호화하는 방법

2.개인정보 비식별 조치 가이드라인

기초 자료 작성 → 평가단 구성 → 평가 수행 → 추가 비식별 조치 → 활용

3.데이터 비식별화

데이터 품질 검증

1.데이터 품질

정확성 - 현식의 값과 정의된 값의 일치 여부

유효성 - 데이터가 정해진 유효기준을 충족하는 특성

완전성 - 데이터의 필수 항목에 누락이 없는 특성

정합성 - 시스템 내의 동일한 데이터가 서로 일치하는 특성

유일성 - 데이터의 구분 기준에 따라 중복이 없는 특성

유용성 - 사용자가 만족할 만한 수준의 최신 데이터가 쉽게 접근할 수 있는 특성

적시성 - 사용자가 필요한 시점에 지연 없이 데이터를 제공하는 특성

보안성 - 데이터의 접근이 적절히 통제되고 개인정보에 대한 보안 특성

안정성 - 에러, 장애의 발생 가능성을 최소화할 수 있는 특성

일관성 - 데이터의 구조, 값, 형태가 일관되게 정의되어 있는 특성

2.데이터 변환 품질 검증

메타데이터 수집 : 테이블 정의서, 컬럼 정의서, 도메인 정의서

메타데이터 분석

데이터 속성 분석

3.수집 데이터의 저장

ETL - 데이터의 추출 → 변환 → 적재 과정을 의미(정형 데이터 기준)

RDB

NoSQL(Not Only SQL)

분산 파일 시스템(HDFS, GFS)

데이터 저장 플랫폼

1.데이터 웨어하우스

기업의 업무시스템에서 발생하는 방대한 데이터를 통합 관리하여 의사결정 도구의 기초 데이터로 사용되는 데이터의 집합체

2.데이터 레이크

전통적인 기업 활동의 정형 데이터뿐만 아니라 비정형, 반정형 등 다양한 유형의 데이터를 저장할 수 있는 저장소

데이터 전처리

데이터 정제

1. 데이터 전처리

데이터 분석을 위한 필수 과정

분석 과정에서 가장 많은 작업시간을 필요로 함

데이터 전처리 과정에서 발생한 오류는 분석의 신뢰성에 영향을 미칠 수 있다.

2. 데이터 정제

분석 작업이 시작되 전 오류를 일으킬 수 있는 결측치, 이상치를 제거하는 사전 작업

결측값 : 필수 데이터가 입력되지 않은 누락된 값

이상값 : 관측된 데이터 범주에서 일반적인 데이터 값의 범위를 벗어난 값

데이터 결측값 처리

데이터 결측값의 유형

1. 완전 무작위 결측

다른 변수와 무관하게 발생한 결측값

2. 무작위 결측

결측값이 다른 변수와 연관이 있음

3. 비무작위 결측

결측값이 다른 변수와 연관이 있으며 분포에 영향을 미침

데이터 결측값 처리 방법

1. 단순 대체법

특정 대푯값으로 결측값을 대체하는 통계적 기법

2. 다중 대체법

여러 번의 단순 대체법을 통해 통계 분석하는 방법

3. 단순 확률 대체법

3-1. Hot-Deck

기존 값으로 결측치 대체

3-2. Cold-Deck

외부 데이터를 통해 대체

3-3. 혼합 방법

여러 가지 대체 방법을 혼합하는 방식

데이터 이상값 검출 방법

ESD - 평균으로부터 표준편차*3 만큼 떨어진 값을 이상값 판단

기하평균 활용 - 기하평균으로부터 표준편차*2.5 만큼 떨어진 값을 이상치 판단

사분위수 활용 - IQR * 1.5만큼 떨어진 값을 이상치 판단

시각화 (히스토그램, 밀도 그래프, 상자 그림)

비지도 학습 - 군집화를 통한 이상치 탐색

마할라노비스 거리

평균과의 거리가 표준편차의 몇 배인지 나타내는 방법

데이터 관측치가 평균으로부터 벗어난 정도를 측정하는 기법

LOF - 관측치 주변의 밀도와 근접한 관측치 주변 밀도의 상대적인 비교로 이상치 탐색

IForest - 의사결정나무를 이용하여 모든 관측치를 고립시켜나가면서 분할 횟수로 탐색

데이터 이상값 처리 방법

1.삭제 : 이상값으로 판단되는 관측값을 모두 삭제하는 방법

2.대체 : 결측값 대체와 동일한 방식으로 이루어지는 대체 방법

3.변환 : 극단적임 값으로 인해 이상값이 발생한 경우 로그변환하는 방법

*오른쪽으로 꼬리가 긴 분포를 가진 경우 로그변환을 이용

데이터 변수 종류

범주형

1.명목형 - 변수나 변수의 크기가 순서와 상관없는 명사형 변수

2.순서형 - 명사형으로써 의미를 갖고 순서에도 의미부여 가능

수치형

1.이산형 - 정수의 형태로 표현할 수 있는 변수

2.연속형 - 소수 형태로 표현할 수 있는 변수

데이터 변수 선택 방법

필터 기법(Filter Method)

데이터의 통계적 측정 방법을 사용하여 변수들의 상관관계를 탐색하는 방법

래퍼 기법(Wrapper Method)

하위 집합을 반복하여 선택하는 방법으로 탐색

1. 전진 선택법 - 모든 변수 중 유의미한 변수 순으로 추가하는 방법
2. 후진 제거법 - 모든 변수 중 불필요한 변수 순으로 제거하는 방법
3. 단계적 방법 - 전진 선택법 + 후진제거법

임베디드 기법

모델 자체에 변수 선택이 포함된 기법

모델의 학습 및 생성 과정에서 최적의 변수를 선택

차원축소 방법

1. PCA - 주성분 분석

여러 변수 중에서 중요한 몇 개의 주성분으로 전체 변동의 대부분을 설명하고자 하는 알고리즘

2. 선형판별분석

데이터를 최적으로 분류하여 차원을 축소하는 방법

3. 요인분석

데이터 안에 관찰할 수 없는 잠재적인 변수를 모형에 세운 뒤 잠재 요인 도출 후 해석

4. 특이값 분해

행렬 데이터를 적용하여 특이값을 추출하고 이를 통해 차원을 축소하는 방법

5. 독립성분 분석

다변량의 신호를 통계적으로 독립적인 하부성분으로 분리하여 차원을 축소하는 방법

6. 다차원 척도법

군집 분석과 마찬가지로 데이터에 내재된 구조를 찾아내 함축적으로 표현하는 방법

변수 변환

1.로그/지수 변환

한쪽으로 치우친 변수를 로그/지수 변환하여 분석 모형을 적합하게 하는 모형

2.비닝(Binning)

연속형 데이터를 범주형 데이터로 변환하기 위해 사용(데이터 평활화)

3.더미 변수호(One Hot Encoding)

4.스케일링

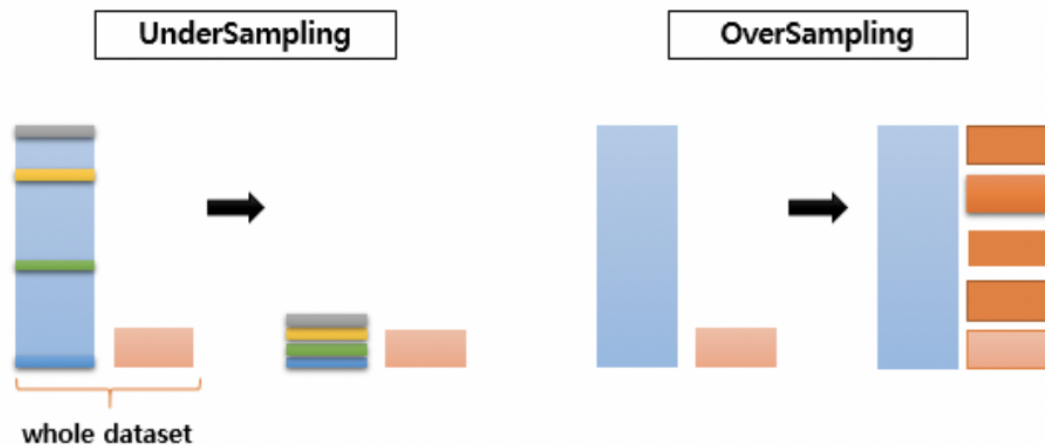
데이터를 특정 구간으로 바꾸는 척도법

불균형 데이터 처리

1.Under Sampling & Over Sampling

Under Sampling - 데이터 손실이 매우 큼

Over Sampling - 과적합 위험 존재



2.SMOTE

과대표집 방법으로, KNN or SVM 등의 알고리즘을 통해 소수 데이터를 토대로 새로운 데이터를 생성하는 방법(데이터 손실과 과적합 위험을 보완한 방법)

데이터 탐색

EDA 이해

데이터를 이해하고 의미있는 관계를 찾아내는 과정
데이터의 통계량과 분포 등을 통해 데이터의 형태를 확인
분석가의 인사이트를 최대화하는 것을 목표로 함.

EDA 특성

- 1.저항성 - 자료의 일부가 기존과 현격히 다른 값으로 대체되었을 때 영향을 적게 받는 것
- 2.잔차 - 관찰값들이 주 경향으로부터 얼마나 벗어났는지를 나타내는 성질
- 3.자료 재표현 - 데이터 분석과 해석을 단순화할 수 있도록 원래 변수를 적당한 척도 변경
- 4.현시성 - 자료를 그래프를 활용해서 시각적으로 표현함으로써 자료의 구조를 효율적으로 파악하는 성질

개별 데이터 탐색 방법

1.범주형 데이터(질적 데이터)

명목형 변수와 순서형 변수에 대한 데이터 탐색
빈도수, 최빈값, 비율, 백분율 등을 이용하여 데이터 특성 파악
시각화는 막대형 그래프를 주로 사용

2.수치형 데이터(양적 데이터)

이산형 변수와 연속형 변수에 대한 데이터 탐색
평균, 분산, 표준편차, 첨도, 왜도 등을 이용하여 데이터 분포 특성 파악
시각화는 박스 플롯이나 히스토그램 주로 이용

다차원 데이터 탐색 방법

1.범주형 - 범주형

빈도수와 비율을 활용한 교차 빈도, 비율, 백분율 분석 등을 활용

2.수치형 - 수치형

산점도와 기울기를 통하여 변수 간의 상관성을 분석

3.범주형 - 수치형

범주형 데이터의 항목들을 그룹으로 간주하고 각 그룹에 따라 수치형 변수의 기술 통계량 차이를 상호 비교

상관관계 분석의 이해

상관관계 - 두 개 이상의 변수 사이에 존재하는 상호 연관성의 존재여부와 연관성 강도

상관관계 종류

양의 상관관계 - 두 변수의 증감이 비례

음의 상관관계 - 두 변수의 증감이 반비례

상관관계 분석 유형

수치형 데이터 - 피어슨 상관관계 분석

순서형 데이터 - 스피어만 상관관계 분석

명목형 데이터 - 카이제곱 검정

산포도의 통계량

- 1.범위 - min - max
- 2.분산 - 편차를 활용하여 데이터의 흩어진 정도를 표현하는 대표적인 산포도 통계량
- 3.표준편차 - 분산의 제곱근, 평균에서 흩어진 정도
- 4.변동계수 - 측정 단위가 서로 다른 자료의 흩어진 정도를 상대적으로 비교할 때 사용
- 5.IQR

데이터 분포를 나타내는 통계량

- 1.왜도 - 데이터 분포의 비대칭성을 표현하는 통계량
- 2.첨도 - 데이터의 분포가 중심에 어느정도 모여있는가를 표현하는 통계량

<https://www.notion.so/417c362b2b8244a885a4a51bdc28226e>

다변량 데이터 탐색

변량 - 조사 대상의 특성, 성질을 숫자 또는 문자로 나타내는 값

일변량 데이터 - 단위에 대해 하나의 속성만 측정하여 얻게 된 변수에 대한 자료
이변량 데이터 - 각 단위에 대해 두 개의 특성을 측정하여 얻어진 데이터
다변량 데이터 - 하나의 단위에 대해 두 가지 이상의 특성을 측정하는 경우

다변량 데이터 탐색

일변량 데이터 탐색

기술통계량과 그래프 통계량을 활용하여 탐색

이변량 데이터 탐색

조사 대상의 각 객체로부터 두 개의 특성을 동시에 관측

다변량 데이터 탐색

분석 시행 전 산점도 행렬, 별 그림, 등고선 그림 등을 통해 시각적 탐색

비정형 데이터 탐색 플랫폼 구성

HDFS - 마스터 / 슬레이브 구조를 가지는 분산형 파일 시스템

맵리듀스 - 맵 함수에서 데이터를 처리하고 리듀스 함수에서 원하는 값을 계산

주키퍼 - 분산 환경에서 노드 간의 정보를 공유하고 락, 이벤트 등 보조기능 프레임워크

Avro - 이기종 간 데이터 타입을 교환할 수 있는 체계를 제공하는 기술

Hive - SQL과 유사한 구조, 데이터 웨어하우징 솔루션

Pig - 대규모 데이터에 대한 분석을 위한 쿼리 인터페이스

Hcatalog - 하둡 데이터 용 테이블 및 스토리지 관리 시스템

기술통계의 이해

데이터 분석의 목적으로 수집된 데이터를 통계적으로 정리하고 요약하는 기초적인 통계 기법

데이터 요약

1. 중심경향 통계량

평균, 중위수, 최빈수

2.산포도 통계량

범위

분산 - 평균으로부터 관측치들이 평균적으로 얼마나 떨어져있는지 요약해 주는 값

표준편차 - 관측치와 동일한 단위로 변화 파악이 가능

평균의 표준오차 - 모집단으로부터 수 많은 표본들을 추출한 후 표본들에 평균을 구한 후 각 평균들에 대한 전체 평균을 다시 구한 값으로 떨어져 있는 정도를 파악

3.분포 통계량

첨도 - 데이터의 분포가 뾰족한 정도

왜도 - 데이터의 분포가 기울어진 정도

표본 추출

1.자료 측정

질적 자료 - 명목 척도 : 단순히 집단의 분류를 목적으로 사용되는 척도

- 순서 척도 : 측정 대상 사이에 대소 관계를 나타내기 위한 척도

양적 자료 - 구간 척도(등간척도) - 온도, 지능지수

- 비율 척도(구간척도) - 절대 0점 존재(몸무게)

2.표본 추출 기법

단순 무작위 추출

계통 추출 - 특정한 특징을 가진 무작위 추출

층화 추출 - 특징별 분리 후 각 모집단 무작위 추출

군집 추출 - 여러 군집으로 나눈 후 일부 군집의 전체 또는 일부

3.확률 분포

어떤 일이 일어날 가능성의 척도

확률 변수 - 결과를 예측할 수 없는 확률 실험에서 나타낼 수 있는 확률적 결과

확률 분포 함수 - 확률 변수를 일직선 상의 공간에 표현한 함수

확률 질량 함수 - 셀 수 있는 사건이 존재하는 경우 각 단순 사건에 대한 확률 정의

확률 밀도 함수 - 임의의 지점에서 밀도를 함수형태로 표시

4.확률 분포의 종류

이산확률분포 - x가 1, 2, 3처럼 셀 수 있는 값

이항 분포

포아송 분포 - 특정시간 동안 특정사건의 평균 발생횟수

연속확률분포

가우시안 분포

표준 정규 분포 - 평균0, 표준편차 1

t - 분포 - 정규분포 평균해석에 많이 쓰이는 분포
카이제곱 분포 - k개의 서로 독립적인 표준 정규 확률 변수를 각 제곱한 다음
합으로 얻은 분포(표준정규분포의 제곱)

F - 분포 - 두 집단 간 분산의 동일성 검정에 활용

5. 표본 분포

모집단에서 추출한 크기가 일정한(n개) 표본이 가지는 추정량의 확률 분포

기본 용어

모집단 - 조사하고자 하는 대상 집단 전체

모수 - 표본 관측에 의해 구하고자 하는 모집단의 정보

표본 - 조사하기 위해 추출한 모집단의 일부 원소

통계량 - 표본에서 얻은 평균이나 표준 오차 같은 값

추정량 - 모수의 추정을 위해 구해진 통계량

표본 분포의 법칙

1. 큰 수의 법칙

데이터를 많이 뽑을수록 표본 평균의 분산은 0에 수렴한다.

2. 중심 극한 정리

표본의 개수가 커지면서 모집단의 분포에 상관없이 표본분포는 정규분포에 근사

구간 추정

추정값에 대한 신뢰도를 제시하면서 범위로 모수를 추정하는 방법

항상 추정량에 분포에 대한 전제가 주어져야 함

신뢰수준

추정값이 존재하는 구간에 모수가 포함될 확률

신뢰구간

신뢰수준을 기준으로 추정된 통계적으로 유의미한 모수의 범위

가설검정

모집단의 특성, 모수에 대한 가정 혹은 잠정적 결론

귀무가설 - 현재까지 주장한 내용이 기존과 변화가 없는 것

대립가설 - 표본을 통해 확실한 근거를 가지고 입증하고자 하는 가설

검정통계량

가설 검정의 대상이 되는 모수를 추론하기 위해 사용되는 표본 통계량

귀무가설이 참이라는 전제 하에 모집단으로부터 추출된 확률표본의 정보를 이용

P-Value - 귀무가설이 참이라는 가정 하에 주어진 표본 데이터를 희소 또는 극한값으로 얻은 확률

가설 검정 절차

P-value와 유의수준을 비교하여 귀무가설 혹은 대립가설 채택

$P\text{-value} < \text{유의수준} \rightarrow \text{대립가설 채택}$

$P\text{-value} > \text{유의수준} \rightarrow \text{귀무가설 채택}$

가설검정 방법

양측검정

모수에 대해 표본자료를 바탕으로 모수가 특정값과 통계적으로 같은지 확인
(같으면 귀무가설 채택 / 다르면 대립가설 채택)

단측검정

가설검정 오류

1종 오류 - 귀무가설이 참인데 기각하는 경우

유의수준(alpha) - 1종 오류를 범할 최대 허용 확률

신뢰수준 - $1-\alpha$

2종 오류 - 귀무가설이 거짓인데 채택하는 경우

베타수준(beta) - 2종 오류를 범할 최대 허용 확률

검정력 - $1-\beta$

분석 모형 설계

분석 모형 선정

데이터 분석 방법을 추상화하는 패턴, 계획, 설명 등을 의미
과거의 분석 경험을 추상화하여 패턴으로 만든 것

분석 방법

분석 목적에 따라 데이터 분석 방법이 달라짐
지도학습 or 비지도학습 등...

1.통계분석

특정 집단이나 불확실한 현상을 데이터를 통해 이해하고 추론을 통해 의사결정하는 과정을 말한다.

2.데이터마이닝

데이터에 숨어있는 유용한 정보를 찾아낸 나 과정
분류, 추정, 예측, 연관, 군집, 기술 등

3.머신러닝

머신 러닝 또는 기계학습은 분석 모형 알고리즘이 데이터를 학습하고 학습한 정보를 바탕으로 결과를 출력하는 분석 방법
지도학습, 비지도학습, 강화학습으로 분류

분석 방법에 따른 예시

- 1.통계 분석 - T-test, Chi-square, ANOVA
- 2.주요 인자 분석 - 많은 인자들 중에서 특정 현상에 영향을 미치는 중요한 인자 선별
- 3.상관관계 분석
- 4.연관 분석 - 장바구니 분석 등을 통해 두 인자간의 연관성을 도출
- 5.분류 분석
- 6.예측 분석
- 7.군집 분석
- 8.텍스트 분석
- 9.소셜 네트워크 분석 - 사회적 관계를 네트워크 구조로 분석하고 시각화 함.

분석 모형 정리

선정한 분석 모형을 데이터에 적용하기 위한 상세 정의 단계

1. 변수 선택

전진 선택법 - 상관관계가 큰 변수부터 순차적으로 모형에 추가

후진 제거법 - 상관관계가 낮은 변수부터 순차적으로 제거

단계적 선택법 - 전진 선택법 후 후진 제거법

2. 하이퍼파라미터 튜닝

모형의 성능에 간접적인 영향을 미치는 하이퍼파라미터 조정

3. 파라미터 vs 하이퍼파라미터

파라미터 - 모형 내부 요소로 모형의 성능에 직접적인 영향을 미침

ex) 인공신경망 가중치, SVM 서포트벡터, 선형회귀 모형의 결정계수

하이퍼파라미터 - 모형 외부 요소로 모형의 성능에 간접적인 영향을 미침

ex) 인공신경망 learning rate, KNN의 k의 개수

4. 하이퍼파라미터 튜닝 방법

1. 매뉴얼 서치 - 사용자가 직감 또는 경험에 근거하여 직접 하이퍼파라미터 조정
매우 비효율적인 방법,

2. Grid search - 하이퍼파라미터의 가능한 모든 조합을 시도 후 최적 값 도출
시간이 매우 오래 걸린다.

3. Random search - 값의 범위를 지정하고 무작위 표본추출을 통해 생성한 조합
시도

Grid search의 단점을 보완

분석 모형 구축의 절차

요건 정의 → 모델링 → 검증 및 테스트 → 적용

1. 요건 정의

기획 단계에서 도출한 내용을 요건 정의로 구체화하는 단계(추진 계획 수립)

2. 모델링

정의된 요건에 따라 본격적인 분석을 수행하는 단계(데이터 준비 및 EDA)

3. 검증 및 테스트

분석 모형을 가상 운영 환경에서 테스트하는 단계

4.적용

분석 결과를 실제 운영 환경에 적용하는 단계

분석 환경 구축

1.데이터 분할

데이터 분석 전에 수집된 데이터를 분할하는 과정

홀드 아웃 - 가정 보편적인 방법

train - 70, test - 30

교차 검증

test data를 제외한 데이터를 중복되지 않게 train data로 사용

부트스트랩

데이터의 분포가 치우쳐 있거나 데이터 건수가 너무 적을 때 사용

분석 기법 적용

회귀 분석 ($y = ax+b$) a-회귀계수, b-절편

하나 이상의 독립변수들이 종속변수에 미치는 여향을 추정하는 통계 분석 기법

회귀 분석에서 독립변수와 종속변수는 선형적인 관계를 가진다.

회귀 분석의 가정

1.선형성

독립변수와 종속변수는 선형적

종속변수는 독립변수와 회귀 계수이며 선형적 조합으로 표현이 가능

산점도를 통해 선형성을 확인 가능

2.독립성

단순 회귀분석에서는 잔차와 독립변수의 값이 서로 독립

다중 회귀분석에서는 독립변수 간에 상관성 없이 독립

3.등분산성

잔차와 분산이 독립변수와 무관하게 일정

잔차가 고르게 분포해야 함

4.정규성

잔차항이 정규분포의 형태를 띠

잔차항의 평균이 0이고 분산이 일정함
Q-Q plot에서 잔차가 우상향하는 형태

회귀 분석의 종류

1. 단순 회귀 ($y = ax+b$)
2. 다중 회귀 ($y = ax + bx^2 + \dots + c$)
3. 다항 회귀 - 독립변수가 2개이고 2차 함수인 경우
4. 곡선 회귀 - 독립변수가 1개이며, 종속변수와 관계가 곡선
5. 비선형 회귀 - 회귀식의 모양이 미지의 모수들의 선형관계로 이루어지지 않은 모형

단순선형 회귀분석

회귀계수의 추정

회귀계수는 최소제곱법을 사용하여 추정한다.

최소제곱법 - 측정값을 기초로 해서 제곱 합을 만들고 그것을 최소로 하는 값을 구하여 측정결과를 처리하는 방식

단순선형 회귀분석의 검정

회귀계수 검정

결정계수 - 회귀 모형의 설명력을 보여주는 지표

다중선형 회귀분석

다중선형 회귀 분석의 경우 단순선형 회귀 분석과 달리 독립변수가 k개이다.

독립변수와 종속변수의 관계는 1차 함수 이상인 경우이며 선형이다.

다중선형 회귀분석 역시 최소제곱법을 사용하여 각각의 독립변수의 가중치 역할을 하는 회귀계수를 구한다.

1. 모형의 통계적 유의성

F-통계량으로 확인이 가능, p-value가 0.05보다 작으면 통계적으로 유의하다고 판단.

F-통계량 - MSR / MSE

2. 회귀계수 유의성 - t-통계량으로 확인 가능

3. 결정계수

규제가 있는 회귀 분석

설명 변수들 사이에 선형 관계가 존재할 경우 다중공선성의 문제가 발생할 수 있기 때문에 규제를 통해 이러한 문제를 해결이 가능하다.

(규제 - 모델의 가중치를 제한하여 차수를 감소시키는 것)

(파라미터에 패널티를 크게 주어서 그 값을 감소시키는 것 - 정규화)

*규제를 하는 이유

최소제곱법 - bias가 없다는 가정하에 분산을 최소화 시키는 방법(과적합 위험 존재)

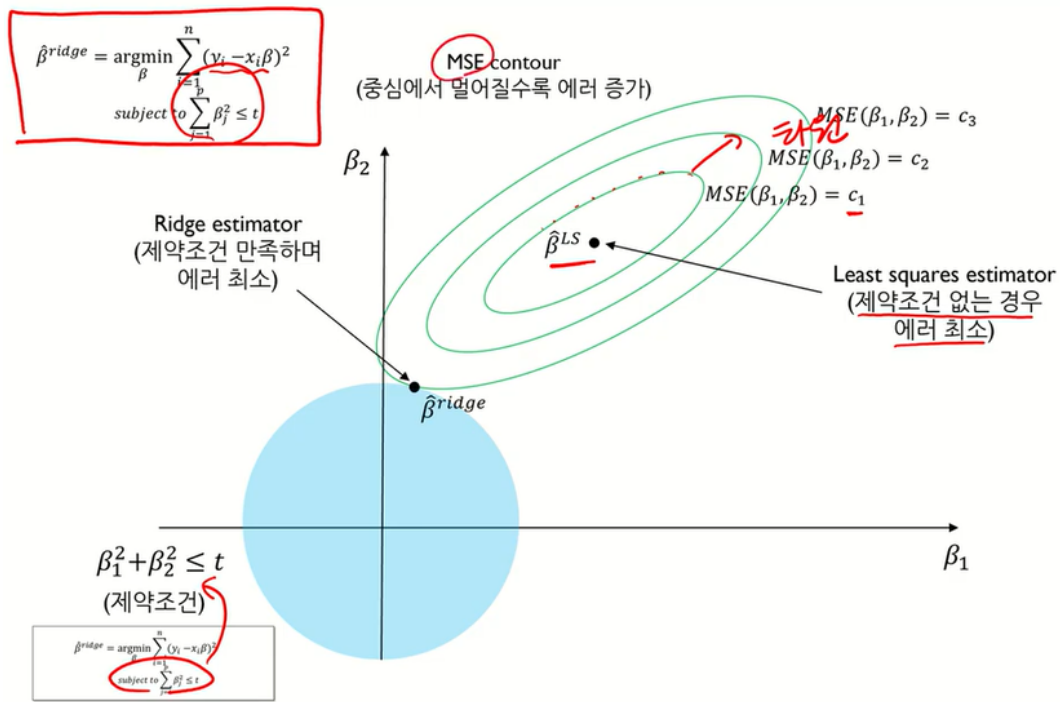
규제를 줄 경우 - bias가 조금 생길수는 있지만 분산을 더 최소화하기 위한 방법

1. Ridge 회귀 (L2 - 제곱의 합 규제) - 변수 선택 불가능

높은 상관관계가 있는 변수 간 검정오차(or MSE)가 최소인 모델을 찾는 것이 목적

변수 간 상관관계가 높을 때 좋은 성능을 보임

Ridge Regression

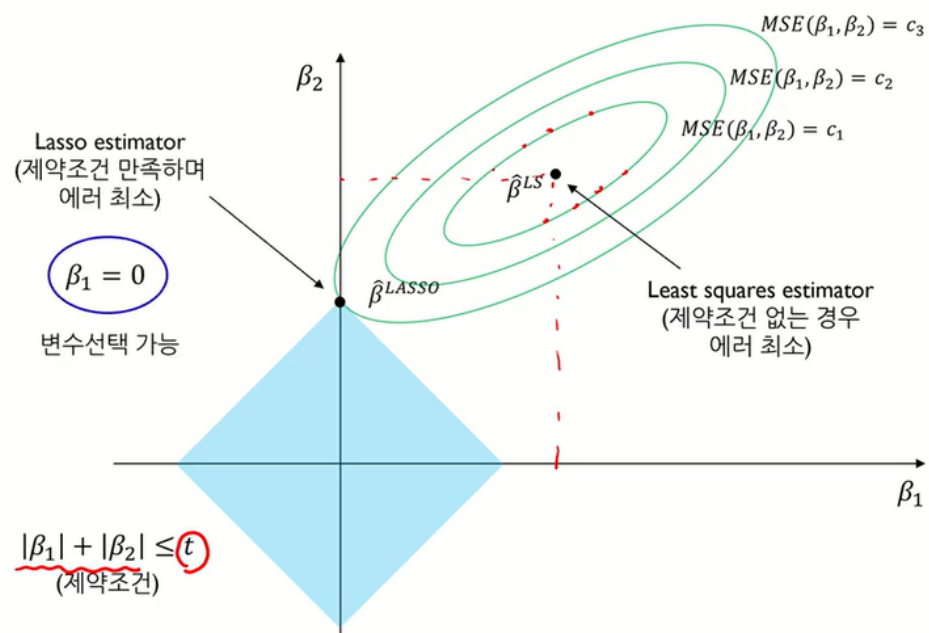


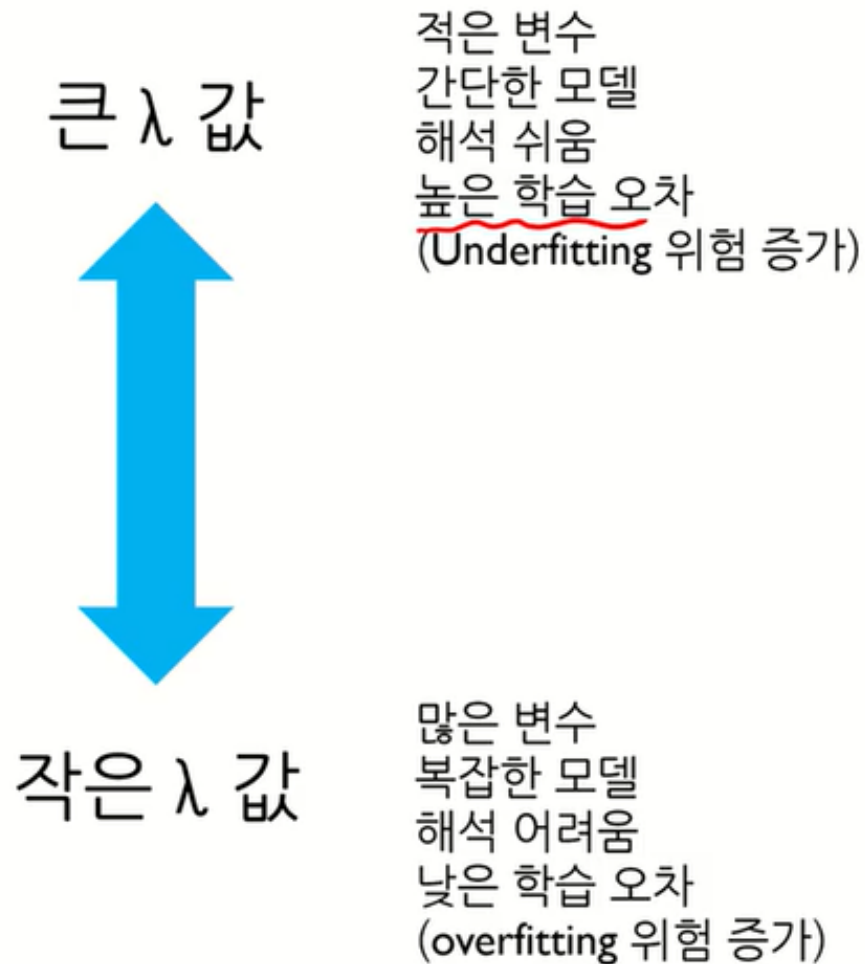
2. Lasso 회귀 (L1 - 절대값의 합 제약) - 변수 선택 가능

변수 선택을 통해 변수 간 검정오차가 최소인 모델을 찾는 것을 목적으로 가진다.

Beta1의 값을 0으로 맞춰줌

Lasso Solution Path





*다중공선성이 있는 경우에는 다중공선성이 발생하는 변수 그룹의 모든 변수가 제거 되는 경우가 발생할 수 있어 성능 저하 위험이 존재

3.Elastic net 회귀 (L1 + L2)

두 회귀의 규제항을 단순히 더해서 사용하며 혼합 비율을 조절하여 어느 방식의 비중을 크게 할 것인지 결정한다.

로지스틱 회귀 분석

로지스틱 개념

독립변수의 선형결합을 이용해 사건의 발생 여부를 예측하며, 종속변수(범주형) 종속변수의 범주가 두 개일 때 이항 로지스틱 회귀분석이라고 함

x, y값이 0~1사이에 위치한다.

선형회귀는 정규분포를 따르지만 로지스틱 회귀분석은 이항분포를 따른다.

모형 적합성

모형의 적합성 - 모형이 설명하지 못하는 데이터의 정도를 의미(이탈도)

계수의 유의성 - ward 검정을 통해 독립변수가 종속변수에 미치는 영향 확인

모형의 설명력 - 의사결정계수(AIC)값을 이용해 검증(결정계수는 사용 안함)