2.1 데이터 정제

개념 1

● 데이터 전처리

데이터 전처리에는 데이터를 정제하는 과정과 분석 변수를 처리하는 과정을 말한다.

데이터 전처리는 분석결과에 직접적으로 영향을 줄 수 있고, 순서는 다음과 같이 진행된다.

데이터 정제 ->결측값 처리 ->이상값 처리 ->분석변수 처리

개념 2

● 데이터 정제 개요

1. 데이터 정제

데이터 정제란 결측값, 잡음, 이상값 등을 제거하는 것을 말한다.

2. 결측값

"NA"로 표현되거나 아예 빈칸으로 존재하는 데이터

3. 이상값

일반적인 데이터 값의 범위를 벗어난 값을 말한다.

개념 3

● 데이터 세분화

1. 계층적 방법

사전에 군집수를 결정하지 않는다.

각 개체를 하나의 소집단으로 간주하고 단계적으로 유사한 소집단들을 합쳐 새로운 소집단을 구성하는 방법

2. 비계층적 방법

군집수를 사전에 결정한다.

전체 집단으로부터 시작하여, 유사성이 떨어지는 개체들을 분리해가는 방법

개념 4

● 결측값의 유형

결측값이 결과에 영향을 주는 경우 비무작위 결측, 영향을 주지 않는 경우 무작위 결측이라고 한다.

1.완전 무작위 결측

다른 변수와 무관하게 랜덤으로 발생한 결측

예) 설문조사 시 특정 항목에 대답하지 않은 경우

2. 무작위 결측

다른 변수와 연관이 있지만, 그 자체가 결과에 영향을 미치지 않는 결측

예) 성별에 따라 응답 확률이 달라서 생기는 결측

3. 비무작위 결측

결과에 영향을 미치는 결측 값

예) 임금을 조사할때, 임금이 낮은 사람이 임금에 대해 응답할 확률이 낮아서 생기는 결측

개념 5

● 결측값 처리하는 방법

1. 결측값을 삭제

2. 목록 삭제

3. 특정 값으로 대체

4. 단순 확률 대치법

개념 6

● 이상값 검출 방법

1. 분산을 이용해서 이상값 검출

2. 가능도를 이용해서 이상값 검출

3. 근접 이웃 기반 이상치 탐지

4. 밀도를 기반으로 한 탐지

5. 사분위수

이상값을 처리하는 방법으로는 삭제, 대체, 스케일링, 정규화 방법 등이 있다.

이상값의 원인 : 표본추출 오류, 고의적인 이상값, 데이터 입력 오류, 실험오류, 측정오류 등

개념 7

● 변수선택

변수 선택이란 종속변수에 유의미한 영향을 미칠 것으로 생각되는 독립변수를 선택하여 변수의 개수를 줄이는 방법을 의미한다.

변수선택의 장점은 다음과 같다.

1. 모델의 학습속도가 빨라진다.

2. 모델의 복잡성이 줄어들고, 사용자가 모델을 해석하기 더 쉽다.

3. 모델의 정확성이 향상될 수 있다.

4. 과적합을 줄일 수 있다.

개념 8

● 변수선택 방법

1. 필터 방법

전처리 과정 중에 각종 통계량을 이용해서 불필요한 특징들을 걸러내는 방법

2. 래퍼 방법

2-1 전진 선택법

가장 유의미한 변수를 하나씩 추가하는 방법

2-2 후진 제거법

모두 적합한 모형에서 변수를 하나씩 제거하는 방법

2-3 단계적 방법

아무것도 적합하지 않은 모형에서 변수를 하나씩 적합하면서 그 전 단계에서 적합된 변수들의 유의미성을 다시 한번 더 검증하는 방법

2-4 AIC

AIC=-2ln(L)+2k 값이 작을수록 더 좋은 모형이라고 판단

3. 임베디드 방법

모형 학습과정에서 변수 선택을 같이 포함하는 방법

예시) 라쏘(LASSO)

개념 9

● 차원의 저주

데이터의 차원이 많아 질 수록, 모형의 성능은 하락하게 되는 현상을 말한다.

개념 10

● 다중공선성

독립변수들 간의 상관관계가 있을 경우를 의미한다.

위와 같은 이유들로 인해, 차원축소가 필요하다

차원축소 방법은 변수선택과 변수추출로 나뉜다.

개념 11

● PCA(주성분 분석)

PCA는 여러 변수들 간의 존재하는 상관관계를 이용해서 선형 연관성이 없는 저차원 공간으로 축소하는 방법을 말한다.

개념 12

● 선형판별분석(LDA)

LDA는 지도학습으로 데이터의 분포를 학습하여 결정경계를 만들어 데이터를 분류한다. LDA는 클래스의 정보를 보호하면서 차원을 최소로 줄이는 방법이다.

개념 13

● t-SNE

T-분포를 이용하여 확률적 차원축소 하는 방법이다.

개념 14

● SVD

행렬분해로 차원축소를 하는 방법 중 하나이다.

개념 15

● 비음수 행렬 분해(NNMF)

행렬의 원소들이 음수가 되지 않게 하면서 행렬분해를 하는 방법이다.

개념 16

● 파생변수

기존 변수들을 조합하여 새롭게 만들어진 변수를 파생변수라고 말한다.

파생변수 생성 방법

1. 하나의 변수에서 정보를 추축해 새로운 변수를 생성

예) 주민등록번호에서 나이와 성별을 추출

2. 한 레코드의 값을 결합하여 파생변수를 생성한다.

3. 조건문을 이용해 파생변수를 생성한다.

단위변환, 표현형식 변환, 요약 통계량 변환, 정보추출, 변수결합, 조건문 이용해서 파생변수 생성 가능하다.

개념 17

● 변수변환

분석 목적에 맞게 데이터를 변환하는 것을 변수변환이라고 부른다.

1. 카테고리 임베딩

범주형에서 연속형으로 변환, 예를 들어 학생 이름의 범주값을 학생의 나이, 주민등록번호 학생번호 등으로 변환하여 사용

2. 더미변수화

남자면 0, 여자면 1로 변환하는 방법

3. 데이터 구간화

나이를 10대,20대,30대 등 범주화 하는 방법

4. 데이터 정형화/정량화

문서와 같은 데이터를 단어 빈도수로 정형화 하는 방법

5. 정규화

최소-최대 표준화, Z-표준화 등등

6. 표준화

로그변환, 루트 제곱근 변환 등..

7. 비닝(Binning)

데이터 값을 몇개의 Bin으로 분할하여 계산하는 방법

* 박스-콕스 변환

=> Data를 정규분포에 가깝게 만들기 위해 사용되는 변환.

개념 18

● 클래스 불균형

학습데이터에서 클래스(label) 에서 불균형이 있는 경우를 말한다.

클래스 불균형 문제 해결 방법

1. 과소표집

무작위로 정상 데이터를 일부만 선택해 유익한 데이터만 남기는 방법이다.

예시: CNN,OSS,ENN,랜덤과소표집
2. 과대표집

무작위로 소수의 데이터를 복제하는 방법

예시: 랜덤과대표집, SMOTE, ADASTN

* 기출 정리

분류형에 대해서 정규화 가능.
최소-최대 정규화는 0~1 사이의 값을 갖는다.

CNN (Condensed Nearest Neighbor) : 데이터 제거하여 대표적인 데이터만 남기는 방법.

임계값 이동 ⇒ test 할때 적용.

불균형 Data 학습하게 되면 잘게쪼갠 다수의 Data class 쪽으로 이동한다.

차원 축소하면 머신러닝 모형의 정확도는 높아진다.