3과목 데이터분석

데이터마트(DM)

데이터 웨어하우스의 한 분야로 특정목적을 위해 사용 (소규모 데이터웨어하우스)

EDA(탐색적 자료 분석)

데이터의 의미를 찾기위해 통계, 시각화를 통해 파악

EDA의 4가지 주제 저향성의 강조, 잔차계산, 자료변수의 재표현, 그래프를 통한 현시성

(저잔재현)

결측값

존재하지 않는 데이터 null로 표시

1)단순대치법

Complet case 함수로 False데이터에 결측값 제거

2)평균대치법

평균으로대치

3)단순확률대치법

가까운 값으로 변경

4)다중 대치법

여러번 대치 (대치 -> 분석 -> 결합)

이상값

이상값을 항상 제거하는것은 아님

1)ESD 평균으로부터 표준편차의 3배가 넘어가는 데이터는 이상값으로 판단

2)사분위수

통계분석

전수조사와 표본조사

전수조사: 전체를 다 조사, 시간과 비용 많이 소모

표본조사: 일부만 추출하여 모집단을 분석

표본 추출 방법

1)랜덤 추출법: 무작위로 표본추출

2)계통 추출법: 번호 부여하여 일정 간격으로 추출

3)집략 추출법: 여러 군집으로 나눈 뒤 군집을 랜덤추출

군집 내 이질적 특징, 군집 간 동질적 특징

4)층화 추출법: 군집 내 동질적 특징, 군집 간 이질적 특징,

같은 비율로 추출 시 비례층화 추출법

5)복원, 비복원 추출법:

자료의 척도구분

등간척도: 구간사이에 간격이 의미가 있으며 덧셈 뺄셈만 가능(온도, 지수 등)

비율척도: 절대적 기준 0이 존재하고 사칙연산 가능한 자료(무게, 나이 등)

Summary함수 결과의 해석

Mean, median 등 존재: 수치형 변수(회귀분석)

집단의 빈도 수: 범주형 변수(분류분석)

기초확률 이론

1)조건확률 특정사건 B가 발생했을 때 A가 발생할 확률

2)독립사건: A,B가 서로 영향을 주지 않는 사건

3)배반사건: A,B가 서로 동시에 일어나지 않는 사건

확률분포

1. 이산 확률 분포

이산균등분포, 베르노이분포, 이항분포, 기하분포, 다항분포, 포아송분포

(배포항항하)

추정

1. 점추정: 모집단이 특정한 값
2. 구간추정: 모집단이 특정한 구간(95%, 99%를 가장 많이 사용)

가설검정

1. 귀무가설: 일반적으로 생각하는 가설(차이가 없다)

차이가 없다 혹은 동일하다

1. 대립가설: 귀무가설을 기각하는 가설, 증명하고자하는 가설(차이가 있다, 크다/작다)
2. 유의수준: 귀무가설이 참일때 기각하는 1종오류를 범할 확률의 허용 한계 일반적 5%
3. 유의확률: 귀무가설을 지지하는 정도를 나타내는 확률

회귀분석 종류

단순회기: 1개의 독립변수와 종속변수의 선형관계

다중회귀: 2개 이상의 독립변수와 종속변수의 선형관계

다항회귀: 2개 이상의 독립변수와 종속변수가 2차함수 이상의관계

릿지회귀: L2규제를 포함하는 회귀모형

라쏘회귀: L1규제를 포함하는 회귀모형

상관분석

두 수변수간의 선형적 관계가 존재하는 파악하는 분석

피어슨 상관분석: 양적, 척도, 연속형 변수, 선형관계 크기 측정

스피어만 상관분석: 서열, 척도, 순서형 변수, 선형/비선형적 관계 나타냄

다중공신성: 다중회귀분석에서 설명변수들 사이에 상관관계가 클 때 모델을 불안정하게 만듬

회귀분석

개념: 독립변수들이 종속변수에 영향을 미치는 파악하는 분석방법

1. 독립변수: 원인을 나타내는 변수
2. 종속변수: 결과를 나타내는 변수
3. 잔차: 계산값과 예측값의 차이

회귀계수 추정방법

최소제곱법: 잔차의 제곱합이 최소가 되는 회귀계수와 절편을 구하는 방법

주성분 분석 (pca)

상관성 높은 변수들의 선형 결합으로 차원을 축소하여 새로운 변수를 생성

자료의 분산이 가장 큰 축이 첫번째 주성분

70~90%의 설명력을 갖는 수를 결정

분해시계열

시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

1. 추세요인: 장기적으로 증가, 감소하는 추세
2. 계절요인: 계절과 같이 고정된 주기에 따라 변화
3. 순환요인: 알려지지 않은 주기를 갖고 변화(경제 전반, 특정산업)
4. 불규칙요인: 위 3가지로 설명 불가능한 요인

(추운 계절의 순환이 불규칙하다)

데이터 마이닝의 유형

지도학습: 정답이 있는 데이터를 사용

비지도학습: 정답이 없는 데이터들 사이의 규칙을 파악

과대적합과 과소적합

과대적합: 모델이 지나치게 데이터를 학습하여 매우 복잡해진 모델

과소적합: 데이터를 충분히 설명하지 못하는 단순한 모델

데이터 분할

과대적합과 과소적합을 방지하고, 데이터가 불균형한 문제를 해결하기 위해 사용

분할된 데이터 셋 종류

1. 훈련용: 모델을 학습하는데 활용 50%
2. 검증용: 모델의 과대,과소 적합을 조정하는데 활용 30%
3. 평가용: 모델을 평가하는데 활용 20%

분할된 데이터의 학습 및 검증방법

1. 홀드아웃: 훈련용과 평가용 2개의 셋으로 분할
2. K-flod 교차검증 데이터를 k개의 집단으로 구분하여 k-1개학습 나머지 1개로 평가
3. LOOCV 1개의 데이터로만 평가, 나머지로 학습
4. 부트스트래핑: 복원추출을 활용하여 데이터 셋을 생성, 데이터 부족, 불균형 문제해소

--------------------------R언어-------------------------------

R의 특징

오픈소스이며 다양한 운영체제에서 사용할 수 있다

S언어의 기반의 프로그래밍 언어이다

다양한 데이터 구조형을 지원하므로 분석 대응력이 좋다

백터연산

NA결측치는 연산의 대상이 아니다 NA를 포함한 데이터의 평균, 합계 함수 적용결과는 Nark ehlsek

숫자형, 문자형, 논리형 백터를 병합하면 문자형이 되며, 숫자열, 논리형을 병합하면 숫자 형이 된다

--------------------------통계------------------------------

결측치 개요

결측치의 추정량 표준오차와 과소추정 문제를 보안하기 위해 단순 확률 대치법과 다중 대치법 등의 다양한 방법이 고려되었다

단순대치법

1. 완전히 응답한 개체 분석

불완전 자료는 모두 무시하는 방법

1. 평균 대치법

관측 또는 실험을 통해 얻어진 데이터의 평균으로 결측값을 대치하는 방법

1. 단순확률 다채법

평균 대치법에서 추정한 표준오차의 과소추정 문제를 보안하고자 고안된 방법

다중 대치법

추정량 표준오차의 과소추정, 계산의 난해성 문제등을 가지고 있음

이상값처리

일반적인 다른값과 멀리 떨어져 있는 값(거리, 밀도)을 의미한다

반드시 제거해야하는 것은 아니며, 분석의 목적이나 종류에따라 적절한 판단 필요

통계 관련 기본언어

모집단: 잘 정의된 연구목적과 이와 연계된 명확한 연구 대상 데이터 전체 집합

모수: 모집단의 평균, 분산 같은 수치들을 모수라고 함

표본: 모집단의 데이터를 전부 조사하기 힘들때 모집단에서 추출 한것

통계량: 표본의 특성을 나타내는 수치

표본의 평균, 분산과 같은 수치를 통계량이라고 함

확률표본 추출법의 종류

1. 단순 무작위 추출

모집단의 각 개체가 표본으로 선택될 확률이 동일하게 추출되는 경우

1. 계통추출

첫 번째 표본을 임의로 선택하고 일정 간격으로 다음 표본을 선택

1. 층화추출

모집단을 서로 겹치치 않게 몇개 집단 또는 층으로 나누고 각 집단 내에서 원하는 크기의 표본을 단순 무작위 추출법으로 추출한다

1. 군집추출

모집단을 차이가 없는 여러개의 집단으로 나눈다

명목척도

단순히 측정 대상의 특성을 분류하거나 확인하기 위한 목적으로 사용(성별, 혈액형)

등간척도 (구간척도)

순위를 부여하되 순위 사이의 간격이 동일하여 양적인 비교가 가능, 절대0점은 존재하지 않는다(온도계, 수치, 물가지수)

비율척도

절대 0점이 존재하며 측정값 사이의 비율 계산이 가능한 척도

서열척도

높고 낮음 등의 순위만 제공할 뿐 양적인 비교는 할 수 없다

항목들 사이 서열이나 순서가 존재

범위(Range)

범위 = 최댓값 – 최솟값

최솟값과 최댓값의 차이를 의미

사분위수 범위(Interqualtile Range)

IQR = Q3 – Q1

가운데 50% 데이터가 위치하는 범위를 의미

편차(Deviation)

편차 = 변량 – 평균

어떤 자료의 변량에서 평균을 뺀 값을 의미

표본분산(Vaiance)

편차 제곱의 합을 자유도 (N-1)로 나눈 것

데이터 집합이 얼마나 퍼져있는지 알아볼 수 있는 수치

표본표준편차(Standard Deviation)

자료의 산포도를 나타내는 수치로, 분산의 양의 제곱근을 의미

평균으로부터 각 데이터의 관찰 값까지 평균 거리를 나타냄

변동계수(CV) – 데이터 퍼짐의 정도 측정

CV = X 분의 S (S= 표본 표준편차, X= 표본 평균)

이상현 확률분포 종류

베르누이분포: 모수가 하나이며, 서로 반복되는 사건이 일어나는 실험을 확률분포로 나타 낸 것

이항분포: 베르누이분포를 n회 반복할떄 성공한 횟수 x에 대한것을 나타낸 것

기하분포: 베루누이 시행에서 처음 성공까지 시도한 횟수 x를 확률변수로 가지는 분포

포아송분포: 단위 시간이나 단위 공간에서 어떤 사건이 몇번 발생할 것인지를 표현한것

연속균등분포 등

연속 확률분포 종류

정규분포, 지수분표, 연속균등분포, 카이제곱분포, F분포, G분포 등

정규분포 개념

가우스 분포라고도 하며 수집된 자료의 분포를 표현하는 데 자주 사용

분포도 언제 사용되는지

T분포, Z분포: 한 집단 또는 두 집단의 평균 차이를 검정

카이제곱 분포: 한집단의 모분산 검정, 범주형 변수의 적합도, 동질성, 독립성 검정

F분포: 두 집단의 분산이 같은지를 검정

모순적 추론

모집단에 특정 분포를 가정하고 분포의 특성을 결정하는 모수에 대해 추론하는 방법

자료가 정규분포이면서 등간척도 또는 비율척도인 경우에만 사용

비모순적 추론

모집단에 대해 특정 분포 가정을 하지 않는다

분포 형태 검정을 위해 중앙값, 순위 등을 사용한다.

회귀모형의 가정

잔차와 관련된 회귀모형의 가정으로 독립성, 정규성, 등분산성, 비상관성이있다

회귀모형 해석 방법

1. 모형이 통계적으로 유의미한가?를 F통계량의 유의확률로 확인한다
2. 회귀 계수들이 통계적으로 유의미한가?를 회귀 계수의 t값에 대한 유의 확률로 확인
3. 모형이 얼마나 설명력을 갖는가?를 결정 계수의 크기로 학인한다
4. 모형이 데이터를 잘 적합하고 있는지 잔차를 그래프로 그려 회귀 진단을 한다

상관계수의 이해

상관계수는 두 변수의 관련성의 정도를 의미하며 -1 ~ 1의 값으로 나타낸다

두 변수의 상관관계가 존재하지 않을 경우 상관계수는 0이다

상관관계가 높다고 해서 인과관계가 있다고 할 수는 없다

공분산

공분산의 공식

2개 확률변수의 선형관계를 나타내는 값이다

하나의 변수가 상승하는 경향을 보일 때 다른 값도 상승하는 선형 상관성이 있다면 양의 공분산을 갖는다

피어슨 상관계수

X, Y의 공분산을 x,y, 표준편차의 곱으로 나눈 값이다

차원축소

주성분 분석: 선형 결합으로, 상관관계가 없는 새로운 변수를 만들고 분산을 극대화 하는 변수로 축약된다

다차원 척도법: 상대적 거리 또는 유사성 정보를 활용하여 고차원 데이터를 저차원 공산 상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 분석방법이다

요인분석: 다차원 데이터 집합에서 숨겨진 구조를 파악하고 변수들 간의 상관관계를 이 해하는 통계적 기법

판별분석: 여러 그룹 또는 범주간의 차이를 감지하고 설명하는 통계적 기술이다

군집분석: 데이터를 유사한 특성을 갖는 그룹 또는 군집으로 나누는 비지도 학습 기법

정준상관 분석: 두개 이상의 다른변수 세트 간의 상관관계 파악 후 이를 최대화

다차원 척도법: 상대적 거리 또는 유사성 정보를 활용하여 고차원 데이터를 저차원 공산 상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 분석 방법

주성분 분석

데이터를 분석할때 변수의 개수가 많다고 모든 데이터를 활용하는 것이 좋은것은 아니다

공분산행렬 또는 상관계수행렬을 사용해 모든 변수를 가장 잘 설명하는 주성분을 찾는다

선형 결합으로 상관관계가 없는 새로운 변수를 만들고 분산을 극대화 하는 변수로 축약

시계열 자료

시간의 흐름에 따라 관측된 데이터

시계열 분석을 위해서는 정상성을 만족해야 함

정상 시계열의 조건

평균은 모든 시점에 대해 일정하다

분산은 모든 시점에 대해 일정하다

공분산은 시점에 의존하지않고 단지 시차에만 의존한다

시계열 모형 종류

자기 회귀 모형: 백색 잡음의 현재값과 자기 자신의 과거값의 선형 가중 값으로 이루어 진 정상 확률 모형이다

이동평균 모형: 과거 q시점 이전 오차들에서 현재 항의 상태를 추론한다

자기회귀 누적 이동평균 모형: 현재와 추세간의 관계를 정의한 것으로 많은 시계열 자료 가 arima모형을 따른다

데이터 마이닝

모든 사용 가능한 원천 데이터를 기반으로 감춰진 지식, 기대하지 못햇던 경향 또는 새 로운 규칙 등을 발견한 후 실제 비즈니스 의사결정 등에 유용한 정보로 활용하는 작업

의사결정 나무

지니지수, 엔트로피지수, 카이제곱통계량의 p-value, ANOVA F-통계량의 P-value는 작을 수록 순수도가 높고 분산 감소량은 클 수록 순수도가 높다

앙상블 모형 개요

여러개의 분류 모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법

앙상블 모형 특징

1. 성능을 분산시키기에 과적합 감소효과 있음
2. 보팅, 배깅, 부스팅, 랜덤 포레스트, 스태킹 있다
3. 상호 연관성이 낮은 모델을 결합할 때 가장 효과 적

경사 하강법

함수 기울기를 낮은 쪽으로 계속 이동시켜 극값에 이를 때까지 반복시키는 것

K-means 군집의 절차

1. 초기 군집의 중심으로 k개의 객체를 임의 방식으로 선택
2. 가장 가까운 군집의 중심으로 할당
3. 자료들의 평균을 계산하여 군집의 중심을 갱신
4. 군집 변화가 없을때까지 2와 3을 반복

DBSCAN

1. 밀도 기반 군집으로 점이 몰려있어 밀도가 높은 부분을 클러스터링 한다
2. 어느 점을 기준으로 반경 x내에 점이 n개이상 있으면 하나의 군집으로 인식
3. 임의적 모양의 군집 분석에 적합
4. K값을 정할 필요가 없으며, 이상치에 의한 성능 하락을 완화

혼합분포 군집

1. 복잡한 형태를 가진 분포의 경우 여러 분포를 확률적으로 선형 결합으로 설명
2. 데이터가 k개의 모수적 모형의 가중합으로 표현된다
3. 군집 크기가 작으면 추정의 정도가 떨어지고 정밀한 추정이 힘듬
4. 모수로 표현 할 수 있으며, 확률 분포를 도입하여 군집 수행
5. 모수와 가중치 추정에 EM알고리즘 사용
6. 데이터가 커지면 수렴에 시간이 더 많이 걸림

로지스틱 회귀 분석

승산(Odds)

성공률 / 실패율

Sigmoid 비선형 함수

연속형 0~1 사이의 값으로 바꾸는 함수

1 + e-x 분의 1

승산비(Odds ratio)

나머지 변수가 주어질 때, 예측변수 x1이 한 단위 증가할 때마다 성공(Y=1)의 오즈가 몇배 증가하는지를 나타냄

Odds\_a / Odds\_b

의사결정 나무 모형

의사결정 규칙을 나무 구조로 나타내 전체 자료를 몇개의 소집단으로 분류하거나 예측

을 수행하는 분석 방법, 과정이 직관적이고 이해하기 쉬움

분리기준

불확실성이 낮아지는 방향, 불순도가 낮아지는 방향, 즉 순수도가 높아지는 방향으로 감

앙상블 모형

보팅, 배깅, 부스팅, 랜덤 포레스트, 스태킹

상호 연관성이 낮은 모델을 결합할 때 가장 효과적임

배깅

문제에 붓스트랩 들어가면 - 배깅

원 데이터에서 중복을 허용하는 크기가 같은 표본을 여러번 단순 임의 복원 추출

하여 각 표본에 대해 모델을 생성하는 기법

같은 데이터가 여러번 추출될 수 있으며 어떤 데이터는 추출 안될 수 있음

보팅

서로다른 여러개의 모형을 생성하고 결과를 집계하여 많은 표를 받은 것을 답으로

하는 방식

부스팅

이전 모델의 결과에 따라 다음 모델 표본 추출에서 분류가 잘못된 데이터에 가중치를

부여하여 표본을 추출함

여러모델을 순차적으로 학습

랜덤 포레스트

배깅에 랜덤 과정을 추가한 방법

노드 내 데이터를 자식 노드로 나누는 기준을 정할 때 모든 예측 변수에서 최적의

분할을 선택하는 대신, 설명 변수의 일부분만을 고려함으로써 성능을 높임

스태킹

다양한 기본모델의 결과를 사용하여 메타 모델을 학습하게 됨

SVM

서로 다른 분류에 속한 데이터 간격이 최대가 되는 선을 찾아 이를 기준으로 분류

ROC 그래프

X축은 FP Rate, Y축은 민감도를 나타내어 이 두 평가 값의 관계로 모형을 평가함

Kappa 상관계수

모델 예측값과 실제값의 일치 여부를 판단하는 통계량으로 사용됨

0~1 사이의 값을 가지며, 1에 가까울 수록 정도가 높다고 판단

향상된 도표

분류 모형의 예측 성능을 평가하기 위한 척도, 주로 불균형 데이터 집합에 사용됨

군집분석

집단간 이질화, 집단내 동질화가 모두 높은것을 군집으로 선택