3.2 분석기법 적용

개념 1

● 연관성 분석 개념

연관성 분석은 상품이나 서비스를 구매하는 등 일련의 거래나 사건의 데이터 안에 존재하는 항목 간의 일정한 연관규칙을 발견하는 과정이다.

예를 들어, 미국의 월마크는 연관성 분석을 통해 맥주를 구매하는 고객은 기저귀도 함께 구매하는 것과 샴푸를 구매하는 사람은 린스를 함께 구매하는 것과 사탕을 구매하는 사람은 키보드를 함께 구매하는 것을 발견했다.

이를 통한 비지니스 전략은 월마트의 매출을 증가 시킬 수 있었다.

개념 2

● 연관성 분석 측정지표

1. 지지도

지지도 = $P(A \cap B) = \frac{\text{A와B를 모두 포함하는 거래의 수}}{\text{전체 거래 수}}$

2. 신뢰도

신뢰도 = $\frac{P(A \cap B)}{P(A)} = \frac{\text{A와 B를 모두 포함하는 거래의 수}}{\text{A를 포함하는 거래의 수}}$

3. 향상도

향상도 = $\frac{P(A \cap B)}{P(A)P(B)} = \frac{\text{A와 B를 모두 포함하는 거래의 수}}{\text{A를 포함하는 거래의 수} \times \text{B를 포함하는 거래의 수}}$

개념 3

● 연관성 분석 측정지표

1. 지지도

지지도는 전체 거래 중에서 A와 B가 동시에 판매되는 거래의 비율을 의미한다.

2. 신뢰도

A의 거래 중에서 B가 포함된 거래의 비율로, 상품 간에 존재하는 연관성의 정도를 측정하는데 사용된다.

3. 향상도

향상도가 1이면 두 품목은 독립이고, 1보다 작으면 두 품목은 음의 상관관계로 A를 구매하면 B를 구매하지 않을 확률이 큼을 의미한다.

개념 4

● 연관성 분석 알고리즘 및 절차

1. Apriori 알고리즘

apriori 알고리즘은 지지도를 사용해 빈발 아이템 집합을 판별하고 이를 통해 계산의 복잡도를 감소시키는 알고리즘이다.

2. FP-growth

나중에 채워 넣기...

개념 5

● 연관성 분석의 장단점

1. 장점

수많은 상품 간에 존재하는 유의미한 구매 패턴을 발견할 수 있다.

자료구조와 계산 과정이 간단하다.

2. 단점

품목의 수가 증가함에 따라 필요한 계산량이 기하급수적으로 증가한다.

연관성 분석을 통해 발견되는 규칙의 수가 많아 유의미한 규칙을 찾기 힘들다.

개념 6

● 군집분석

관측치들의 유사성에 기초해 전체 데이터를 몇 개의 집단으로 나누는 분석기법이다.

대표적인 거리측도로는 유클리드 거리, 맨해튼 거리 등이 있고 유사성 측도로는 코사인 거리와 상관계수가 있다.

개념 7

● 거리측도

1. 변수가 연속형인 경우

유클리드 거리 : $d(x,y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$

맨해튼 거리 : $d(x,y) = |x_1 - y_1| + ... + |x_n - y_n|$

2. 변수가 범주형인 경우

단순 일치 거리 : $d(i,j) = \frac{m}{t}$

이외에도 자카드 거리, 해밍거리 등이 사용된다.

개념 8

● 계층적 군집

1. 계층적 군집의 개념

개별 관측치 간의 거리를 계산해서 가장 가까운 관측치부터 결합해감으로써 계층적 트리 구조를 형성하고 이를 통해 군집화를 수행하는 방법이다.

2. 군집 간 거리

단일 연결법 : 각 군집에 속하는 임의의 개체 사이의 거리 중에서 가장 작은 값을 거리로 정의

완전 연결법 : 각 군집에 속하는 임의의 개체 사이의 거리 중에서 가장 큰 값을 거리고 정의

평균 연결법 : 모든 가능한 관측치 쌍 사이의 평균 거리를 이용하는 방법

중심 연결법 : 각 군집의 중심적 사이의 거리를 이용하는 방법

와드 연결법 : 군집의 평균과 각 관측치 사이의 오차 제곱 합의 크기를 이용하는 방법

개념 9

● 비계층적 군집

1. K-means 군집

step1 : 주어진 군집의 개수만큼 초기값을 선택

step2 : 초기값과 각 개체 간의 거리를 계산하고 개체별로 가장 가까운 초기값에 데이터를 할당

step3 : 각 군집 내 개체들의 평균을 구해 군집의 중심점을 업데이트

step4 : 중심점 변화가 일정 수준 이하가 될 때까지 위의 과정을 반복

2. DBSCAN

DBSCAN은 밀도기반 군집분석의 한 방법으로 밀집한 정도를 바탕으로 군집한다.

3. 가우시안 혼합 모델

데이터가 *k*개의 정규분포로부터 생성됐다고 가정하는 모델로 데이터로부터 모수와 가중치를 추정하는 대표적인 모수적 군집 방법이다.

개념 10

● EM 알고리즘

1. E-step

E 단계에서는 모수를 임의의 값으로 설정하고 이 모수가 정확하는 가정하에 잠재변수 Z의 기대치를 추정한다.

2. M-step

M단계에서는 E단계에서 추정한 잠재변수 Z가 정확하는 가정하에 모수를 추정한다.

개념 11

● 범주형 자료분석 개념

범주형 자료분석은 독립변수와 종속변수가 모두 범주형 데이터이거나 둘 중 하나가 범주형 데이터일 때 사용하는 분석방법이다.

개념 12

● 상대적 위험도

상대적 위험도는 코호트 연구에서 주로 사용하는 방법으로 위험인자에 노출됐을 때 질병이 발생할 확률과 위험인자에 노출되지 않았을 때 질병이 발생할 확률의 비로 표현된다.

코호트 연구 : 특정인자가 질별발생에 영향을 미치는지를 확인하는 연구 방법

개념 13

● 카이제곱 검정

카이제곱 검정은 범주형 자료 간의 차이를 보여주는 분석방법이고, 적합성검정, 동질성 검정, 독립성 검정의 세가지 형태로 나뉜다.

개념 14

● 적합도 검정

변수가 1개이고 그 변수가 2개 이상의 범주로 구성되어 있을때 사용하는 일변량 분석방법이다.

아래와 같은 데이터가 있을때, 가설은 다음과 같다.

$H_0 : p_A = 0{,}1, P_B = 0{,}3, P_C = 0{,}2, P_D = 0{,}4$ vs $H_1 : H_0$ is not

구분	A	B	C	D	합
관측빈도	5	40	15	39	100
기대빈도	10	30	20	40	100

풀이)

카이제곱 통계량 = $\frac{(10-6)^2}{10} + \frac{(30-40)^2}{30} + \frac{(20-15)^2}{20} + \frac{(40-39)^2}{40} = 6{,}21$

자유도 = 4-1 =3

자유도에 대한 유의수준 0.05 카이제곱 분위수 = 7.81

따라서, 귀무가설을 채택한다.

개념 15

● 독립성검정

독립성 검정은 변수가 2개 이상의 범주로 분할되어 있을 때 사용하는 방법으로,각 범주가 종속변수에 영향을 주는지를 확인하는 검정 방법이다.

아래와 같은 데이터가 있을때, 가설은 다음과 같다.

H_0 :국적에 따른 최종 학력은 차이가 없다. H_1 : 국적에 따른 최종 학력은 차이가 있다.

구분	중졸	고졸	대졸	합계
한국	17(25.2)	33(26.6)	20(18.2)	70
중국	37(28.8)	24(30.4)	19(20.8)	80
합계	54	57	39	150

풀이)

카이제곱 통계량 = $\frac{(17-25,2)^2}{25,2} + \frac{(33-26,6)^2}{26,6} + \frac{(20-18,2)^2}{18,2} + \frac{(37-28,8)^2}{28,8} + \frac{(24-30,4)^2}{30,4} + \frac{(19-20,8)^2}{20,8} = 8,23$

자유도 = (2-1)(3-1)=2

자유도에 대한 유의수준 0.05 카이제곱 분위수 = 5.99

따라서, 귀무가설을 기각한다.

개념 16

● 동질성 검정

동질성 검정은 각 부모집단으로 추출된 관측치들이 각 범주 내에서 서로 균일한 값을 가지는가를 검정하는 방법이다.

아래와 같은 데이터가 있을때, 가설은 다음과 같다.

$H_0 : P_{A_j} = P_{B_j}$ vs H_1 : 식단에 따라 다이어트 효과가 차이가 있다.

구분	효과적	보통	효과없음	합계
식단 A	50(33.3)	30(36.7)	20(30)	100
식단 B	50(66.7)	80(73.3)	70(60)	200
합계	100	110	90	300

풀이)

카이제곱 통계량 = $\frac{(50-33,3)^2}{33,3} + \frac{(30-36,7)^2}{36,7} + \frac{(20-30)^2}{30} + \frac{(50-66,7)^2}{66,7} + \frac{(80-73,3)^2}{73,3} + \frac{(70-60)^2}{60} = 19,3$

자유도 = $(2-1)(3-1) = 2$

자유도에 대한 유의수준 0.05 카이제곱 통계량 = 5.99

귀무가설을 기각한다.

개념 17

● 인공신경망-퍼셉트론

1. 퍼셉트론의 개념

퍼셉트론은 인간의 신경망에 있는 뉴런의 모델을 모방하여 입력츠, 출력층으로 구성한 인공신경망 모델이다.

2. 퍼셉트론의 구성요소

퍼셉트론의 구조는 입력값, 가중치, 순 입력함수, 활성화함수, 예측값으로 되어 있다.

개념 18

● 다층 퍼셉트론

1. 다층 퍼셉트론

퍼셉트론 모형의 입력층과 출력층 사이에 하나 이상의 은닉층을 가지는 모형이다.

2. 문제점

과대적합, 기울기 소실의 문제들이 있다.

개념 19

● 단일표본 T-test

한 집단의 평균이 모집단의 평균과 같은지 검정하는 것을 단일표본 t-test라고 한다.

T-통계량 $= \frac{\bar{X}-\mu}{s/\sqrt{n}}$ 을 이용해서 검정을 진행한다.

개념 20

● 대응표본 T-test

대용포본 t-test란 동일한 집단의 처치 전후 차이를 알아보기 위해 사용하는 검정 방법이다.

두 그룹의 데이터를 대응하는 것이 아니라 한 그룹의 처치 전 데이터와 처치 후 데이터를 대응하는 방법이다.

T-통계량 $= \frac{\bar{d}}{s/\sqrt{n}}$

$\bar{d}$= 두 집단 차이의 평균

s = 두 집단 차이의 표본표준편차

개념 21

● 독립표본 T-test

독립표본 t-test는 데이터가 서로 다른 모집단에서 추출된 경우 사용할 수 있는 분석 방법으로 독립된 두 집단의 평균의 차이를 검정한다.

분석을 본격적으로 시행하기 전에, 두 집단의 분산에 대하여 등분산성 검정을 먼저 시행한다.

두 집단이 등분산성을 만족할 때, 아래의 통계량을 이용한다.

T-통계량 = $\frac{\bar{X}_A-\bar{Y}_B}{\sqrt{s_p^2(\frac{1}{n_A}+\frac{1}{n_B})}}$ A : 집단 A의 평균 , $\bar{Y}_B$: 집단 B의 평균

s_p^2 : 통합 분산 추정량

$s_p^2 = \frac{(n_A-1)s_A^2-(n_B-1)s_B^2}{n_A+n_B-2}$

개념 22

● 다변량분산분석

ANOVA가 종속변수가 하나고 범주가 2개 이상일때 각 범주 간의 평균을 비교하는 방법이라면, MANOVA는 2개 이상의 종속변수가 주어졌을 때, 각 범주 간의 평균 벡터의 차이를 비교하는 분석 방법이다.

다변량분산분석을 실행하기 위해서는 독립변수가 범주형 데이터이고, 종속변수가 수치형 데이터여야 한다.

개념 23

● 요인분석

요인분석은 변수 간에 존재하는 상호연관성을 바탕으로 데이터를 적은 수의 요인으로 압축 및 요약해 그룹화 하는 방법이다.

1. 요인분석의 조건

변수가 연속형 데이터 형태 여야 한다.
관측치들은 서로 독립, 각 변수는 다변량 정규분포를 따라야 한다.
변수별로 분산은 모두 동일해야 한다.
표본의 수는 최소한 50이상이여야 한다.

2. 요인분석의 목적

여러개의 변수를 몇개의 요인으로 묶어 자료를 요약하고, 변수의 차원을 축소시킨다.
변수들 내에 존재하는 상호 독립적인 특성을 발견한다.
불필요한 변수를 제거한다.

3. 요인분석의 과정

분석할 변수를 선택한다.
상관관계 행렬을 통해 요인분석이 가능한지 확인한다.
요인추출 방법을 결정하고 이를 통해 요인을 추출한다.
요인의 회전을 통해 해석하기 쉽게 분산을 재분배한다.
요인을 해석한다.
요인점수를 산출한다.

개념 24

● 요인분석

4. 요인추출 방법

요인추출 방법은 크게 주성분 분석에 의한 추출과 공통요인분석에 의한 추출로 구분된다.

주성분 분석은 데이터의 총분산을 사용하는 방법이고, 공통요인분석은 공통분산만을 사용하는 방법이다.

5. 요인의 수 결정 방법

요인의 수를 결정할 때는 최소고윳값을 사용하는 방법이 가장 많이 사용된다.

6. 요인의 회전

요인분석의 각 요인은 단순히 자료를 축약하는 과정에서 변수들의 상관관계에 따라 추출되었기 때문에 그 자체로만으로는 의미 있는 정보를 얻기 어렵기에, 해석하기쉽게 요인을 회전하는 방법을 이용한다.

개념 25

● 판별분석

판별분석은 두 개 이상의 모집단으로부터 추출된 표본들을 분석해 각 표본이 어느 모집단에서 추출된 것인지를 예측하는 분석방법으로 몇 개의 알려진 그룹으로부터 그룹을 구별할 수 있는 판별함수를 도출하고, 도출된 판별함수를 통해 새로운 데이터를 판별하여 분류하는 작업을 수행한다.

절차

1. 판별분석에 사용할 변수를 선정한다.

2. 각 개체를 분류하는 데 사용할 판별함수를 도출한다.

3. 도출된 판별함수의 정확도를 파악한다.

4. 판별함수를 이용해 새로운 데이터가 속할 집단을 예측한다.

개념 26

● 다차원척도법(MDS)

다차원 척도법은 다변량 데이터에 내재된 특성 및 구조를통해 개체 간의 유사성을 측정하고, 이를 원래의 차원보다 낮은 차원의 공간에 점으로 표현하는 분석방법이다.

MDS는 개체의 실제 거리와 모형에 의해 추정된 거리사이의 적합도를 측정하기 위해 stress라는 척도를 사용한다.

$$\text{stress}=\sqrt{\frac{\sum(\text{실제거리}-\text{추정거리})^2}{\sum \text{실제거리}^2}}$$

개념 27

● 시계열 분석

시계열 데이터는 시간에 따라 관측된 데이터를 의미한다.

정상성 조건

1. 시계열의 평균이 시간에 따라서 일정하다.

2. 분산이 시점에 의존하지 않고 일정하다.

3. 시점 간의 공분산이 특정 시점에 의존하지 않고 오직 시차에만 의존한다.

시계열의 구성요소

시계열은 추세요인, 계전요인, 순환요인, 불규칙요인으로 구성되며 이들 요인이 복잡하게 혼합되어 하나의 시계열 데이터를 구성한다.

개념 28

● 시계열 분석 기법

1. 이동평균법

이동평균법은 시계열 데이터에서 일정 기간별로 자료를 묶어 평균을 구하는 방법이다.

2. 지수평활법

이동평균법은 장기적인 추세를 파악하는 것에는 효과적이나 m기간에 따라 평균의 수가 감소하는 단점이 있다. 지수평활법은 이러한 문제점을 해결하기 위해 사용하는 방법으로 최근 자료가 과거 자료보다 예측에 효과적이라는 가정하에 최근 데이터일수록 큰 가중치를 부여한다.

3. 가법모형

가법모형은 시계열 데이터가 네 종류의 시계열 구성요소의 합으로 구성된다고 가정하는 것이다.

4. 승법모형

승법모형은 시계열 데이터가 네 종류의 시계열 구성요소의 곱으로 구성된다고 가정하는 것이다.

개념 29

● 시계열 모형

1. 자기회귀모형

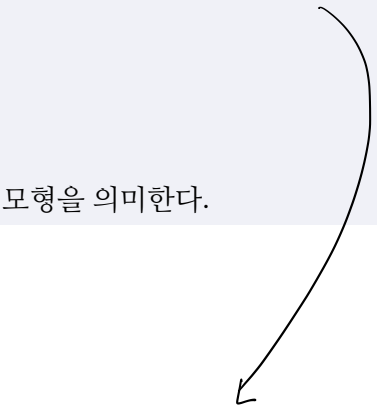
자기회귀모형은 변수들의 자기상관성을 기반으로 한 시계열 모형으로 현시점의 자료를 p 시점 전의 과거 자료를 통해 설명할 수 있는 모델이다.

2. 이동평균모형

이동평균모형은 이동평균 과정으로 현재 데이터가 과거 백색잡음의 선형 가중합으로 구성된다는 모형이다.

3. 자기회귀누적이동평균모형

ARIMA모형은 앞에서의 AR모형과 MA모형을 합친 모형을 의미한다.



시간이 지날수록 관측치의 평균값이
지속적으로 증가하거나 감소할때 사용.

정상성 가정 필요 X

* 기출 정리

주성분 분석 = Data 간 높은 상관관계가 있을 때, 상관관계를 제거

자료의 변동을 최대한 보존 (제거 반대)

변동 폭이 큰 축을 선택.

공분산 행렬을 사용하여 고유값이 1 보다 큰 주성분 개수 이용.

다차원 척도법 : 스트레스 ≈ 0 적합도 좋고

≈ 1 적합도 나쁨.

ARIMA : AR, MA, ARMA 로 가능.

MA : 현시점의 자료를 유한개의 백색잡음 선형결합으로 표현. , 항상 정상성 만족.

필기체 인식 ⇒ RNN