3.2 분석기법 적용

개념 1

● 회귀분석

회귀분석이란 독립벼눗와 종속변수 간에 선형적인 관계를 도출해 독립변수가 종속변수에 미치는 영향을 분석하는 방법이다.

회귀분석을 진행하기 전에, 산점도를 이용해서 데이터의 경향성을 먼저 살펴본다.

개념 2

● 회귀분석의 가정

1. 선형성

독립변수와 종속변수가 선형관계이다

2. 독립성

잔차와 독립변수가 서로 독립이어야 한다.

3. 등분산성

독립변수와 무관하게 잔차의 분산이 일정하다.

4. 정규성

오차가 정규분포를 따르는 확률변수이다. 잔차가 정규분포.

5. 비상관성

오차끼리는 독립이다. 잔차끼리 독립

개념 3

● 회귀분석 학습방법

회귀모형은 오차의 제곱합이 최소가 되는 방향으로 학습된다.

개념 4

● 회귀모형의 모형적합성

* SSR : 회귀선에 의해서 설명되는 변동값
SSE : 회귀선에 의해서 설명되지 않는 변동값.

1. 분산분석표 이용

SST,SSR,SSE 등을 이용해서 모형 성능 평가

특히, F-통계량을 이용해서 통계적으로 유의성을 확인한다.

* AIC, BIC ↓ 여야
모형의 설명력 ↑ 이라고 판단한다.

개념 5

● 회귀계수의 유의성 검증

각 회귀계수에 대한 t-통계량을 이용해서 회귀계수의 유의성을 판별할 수 있다.

개념 6

● 모형의 설명력

회귀모형의 설명력은 R^2로 판단할 수 있다. 1에 가까울수록 모형이 데이터를 잘 설명한다고 해석한다.

**참고

결정계수는 회귀계수가 많아질수록 증가하는 성질이 있어, 제대로된 모형 판별을 못할 수 있다.

이를 해결하고자 수정된 결정계수가 사용되며, 이는 회귀계수가 많아져도 항상 증가하지는 않는다.

개념 7

● 로지스틱 회귀분석 개념

앞에서는 회귀모형이 직선이였는데, 로지스틱 회귀모형은 시그모이드 함수 형태인 것을 말한다.

주로, 분류문제와 output 데이터가 0과1사이인 경우에 사용된다.

개념 8

● 오즈(odds)

오즈는 확률 p가 주어졌을 때, 사건이 발생할 확률이 사건이 발생하지 않은 확률의 몇배인지에 대한 지표이다.

$$odds = \frac{p}{1-p}$$

개념 9

● 로짓변환

$$logit(p) = log\frac{p}{1-p} = log(odds)$$

개념 10

● 시그모이드 함수

$$f(x) = \frac{1}{1+e^{-x}}$$

개념 11

● 로지스틱 모형의 통계적 유의성 검증

1. 모형의 통계작 유의성 검증

이탈도(deviance)를 이용해 로지스틱 회귀모형의 유의성을 검증할 수 있다.

이탈도는 모형이 설명되지 못하는 데이터의 정도를 의미하며 이탈도가 적을수록 모형이 통계적으로 유의하다고 판단된다.

2. 계수의 유의성 검증

왈드 검정을 통해 독립변수에 대한 유의성 검증을 할 수 있다.

3. 모형의 설명력

McFadden이 제안한 의사 경정계수와 Cox and Snell이 제안한 결정계수를 통해 모형의 설명력을 판단할 수 있다.

개념 12

● 의사결정 트리

의사결정 나무는 데이터에 존재하는 규칙을 학습하는 알고리즘이다.

크게 회귀 트리와 분류 트리로 구분된다.

개념 13

● 의사결정 트리의 구성요소

1. 뿌리마디

나무구조가 시작되는 마디

2. 자식마디

하나의 마디로부터 분리되어 나간 2개의 마디들

3. 부모마디

자식마디의 상위마디

4. 끝마디

자식마디가 없는 마디

5. 가지

마디와 마디를 잇는 선

6. 깊이

가지를 이루는 마디의 개수

개념 14

● 가지치기 분리 기준

1. 종속변수가 이산형

카이제곱 통계량, 지니 지수, 엔트로피 지수가 사용된다. 불순도를 측정하는 지수들이다.

2. 종속변수가 연속형

ANOVA F-통계량, 분산감소량이 사용된다.

개념 15

● 카이제곱 통계량 구하는 방법

카이제곱 통계량은 각 셀의 $\frac{(\text{기대도수}-\text{실제도수})^2}{\text{기대도수}}$를 모두 더한 값이다.

아래와 같은 표가 있을때, 카이제곱 통계량 계산은 다음과 같다.

$\frac{(33-40)^2}{33}+\frac{(22-15)^2}{22}+\frac{(27-20)^2}{27}+\frac{(18-25)^2}{18}=8{,}2$

구분	Good	Bad	합계
왼쪽마디	40(33)	15(22)	55
오른쪽마디	20(27)	25(18)	45
부모마디	60	40	100

기대도수(왼쪽마디,Good)=$60*55/over100=33$

개념 16

● 지니 지수

Gini index = $1-\sum_i p_i^2$ $P_i = \frac{\text{i원소 개수}}{\text{전체 개수}}$

개념 17

● 엔트로피 지수

Entropy = $-\sum_i p_i log_2 p_i$

개념 18

● 의사 결정트리의 장단점

1.장점

해석하기 쉽고, 이상치에 덜 민감하다. 또한 수학적 가정이 불필요하고 변수 선택이 자동적으로 이루어진다.

2. 단점

분류 기준 값의 경계선 부근의 자료 값에 대해서는 오차가 클 수 있다.

모형이 너무 복잡할 경우, 예측 정확도가 하락하고 해석하기 어렵다.

각 변수의 영향력을 파악하기 힘들다.

개념 19

● 인공신경망

인공신경망은 생물의 신경계를 모방해 예측 및 분류를 하는 머신러닝 알고리즘이다.

개념 20

● 활성함수

인공신경망 노드의 계산 결과는 그 다음 노드로 바로 전달되지 않고 비선형함수를 통과해서 전달되는데, 이러한 비선형함수를 활성함수라고 한다.

개념 21

● 할섬함수 종류

1. Step 함수

$$s(x) = \begin{cases} 0 & x < 0 \\ 1 & x => 0 \end{cases}$$

2. Sigmoid 함수

$$s(x) = \frac{1}{exp(-x)}$$

3. Sign 함수

$$s(x) = \begin{cases} -1 & x < 0 \\ 1 & x => 0 \end{cases}$$

4. tanh 함수

$$s(x) = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)}$$

5. ReLU 함수

$$s(x) = \begin{cases} 0 & x <= 0 \\ x & x > 0 \end{cases}$$

6. Softmax 함수

$$s(x_i) = \frac{exp(x_n)}{\sum_{i=1}^{k} exp(x_i)}, n = 1, ..., k$$

개념 22

● 역전파 알고리즘

역전파 알고리즘은 출력층에서 결정된 결과값의 오차를 출력층에서 입력층으로 역으로 전파하여 오차가 최소가 되게 가중치를 갱신한다.

인공신경망은 역전파 알고리즘을 통해 학습한다.

개념 23

● 단층 퍼셉트론

입력층이 은닉층(hidden layer)를 거치지 않고 출력층과 바로 연결된다. 이 출력값이 정해진 임계값을 넘었을 경우 1 넘지 못하면 0으로 출력한다.

개념 24

● 다층 퍼셉트론

입력층과 출력층 사이에 하나 이상의 은닉층을 두어 비선형적인 데이터도 학습할 수 있게 한 알고리즘이다.

개념 25

● 인공신경망 모형의 장단점

1. 장점

잡음에 민감하게 반응하지 않고, 비선형적인 문제를 분석하는데 유용하다. 패턴인식, 분류, 예측 등의 문제에 효과적이다.

2. 단점

모형이 복잡하면 학습에 오래걸린다. 초기 가중치에 따라 전역해가 아닌 지역해로 수렴할 수도 있다. 해석이 쉽지 않다.

개념 26

● 서포트 벡터 머신(SVM)

서포트 벡터 머신은 분류와 회귀분석에 사용되는 지도학습 알고리즘이다.

1. 마진(Margin)

결정경계에 가까운 데이터들은 서포트 벡터라고 부르는데, 서포트벡터와 결정경계의 거리를 마진이라고 부른다.

SVM은 이러한 마진을 최대화 시키는 방향으로 학습하는 알고리즘이다.

개념 27

● 소프트 마진

기존의 SVM은 마진을 엄격하게 정의하는 방법으로 하드 마진방법으로 불린다.

소프트 마진은 SVM는 마진에 대하여 약간의 오분류를 허용하는 방법이다.

개념 28

● 커널을 이용한 SVM

실제 데이터는 선형으로 분리되지 않는 데이터가 대부분이다. 선형으로 분리되지 않는 데이터를 커널트릭을 통해 저차원에서 고차원으로 매핑하는 SVM방법이다.

개념 29

● 서포트 벡터 머신의 장단점

1. 장점

서포트 벡터만을 이용해 결정경계를 생성하므로 데이터가 희소할 때 효과적이다.

비선형 데이터도 커널 트릭을 이용해 분류할 수 있다.

인공신경망보다 과적합의 위험이 적다.

노이즈의 영향이 적다.

2. 단점

데이터의 크기가 클때, 학습시간이 오래걸린다.

블랙박스 모형이라 해석이 어렵다.

* 기출 정리

계층적 군집 : 군집수 모름 안정함

비계층적 군집 : K-means

DBSCAN => 개체들의 밀도 계산을 기반으로 밀접하게 분포된 개체들끼리 그룹핑

회귀분석 결과 창에서 $|P(X>|t|) = p\text{-value}$ 이다.

서포트 벡터가 여러개 일 수 있다. 속도가 느리다. (SVM)

RBF 커널 = 2차원 → ∞ 차원

↳ 비선형 Data 일때 사용한다.

* 연관성 분석.

$A \rightarrow B$

지지도 = $P(A \cap B)$

신뢰도 = $P(B|A)$

향상도 = $\frac{P(B|A)}{P(B)}$

$$\text{odds ratio} = \frac{\text{관심집단 오즈}}{\text{비교집단 오즈}}$$

* 적합도 검정

df = 범주의 수 − 1

H_0 : 기대되는 빈도와 일치

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

* 독립성 검정.

df = (범주 1의 수 − 1) × (범주 2의 수 − 1)

H_0 : 요인 1과 요인 2는 독립.

↳ 동질성 검정은 독립성과 방법이 동일.