3과목.빅데이터 모델링

(Ch_02. 분석기법 적용 - SEC 02. 고급 분석기법-2)

빅데이터 분석 기사(3과목. 빅데이터 모델링)

CHAPTER 1. 분석 모형 설계

CHAPTER 2. 분석기법 적용

분석기법 적용

분석기법 적용 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 분석기법
2. **고급 분석기법**

3. 분석기법 적용 – 고급 분석기법

06 비정형 데이터 분석

2) 오피니언 마이닝(Opinion Mining)

- 오피니언 마이닝은 웹사이트와 소셜미디어에서 특정 주제에 대한 여론이나 정보(게시글 등)를 수집하여 분석한 뒤 정보를 도출하는 빅데이터 처리 기법이다.
- 오피니언 마이닝은 특정 서비스에 대한 소비자의 의견이 긍정적 혹은 부정적인지를 분석하고, 그 원인을 도출하는 것을 목적으로 하며, 이를 통해 대중의 관심과 여론이 어떻게 변화하는지 파악할 수 있다.

[오피니언 마이닝 분석 절차]

데이터 수집 및 전처리 → 데이터 분류(긍정, 부정) → 요약 및 시각화

3) 웹 마이닝(Web Mining)

- 웹 마이닝은 웹 자원으로부터 의미 있는 정보를 추출하기 위한 데이터 마이닝 기법이다.
- 웹 마이닝의 유형은 웹 구조 마이닝, 웹 내용 마이닝, 웹 사용 마이닝으로 구분된다.

3. 분석기법 적용 – 고급 분석기법

3) 웹 마이닝(Web Mining)

웹 마이닝 유형

유형	설명
웹 구조 마이닝	웹 사이트와 웹 페이지의 구조적 요약 정보를 얻기 위한 기법
웹 내용 마이닝	실제 웹 사이트의 내용 중 의미있는 내용을 추출하는 기법
웹 사용 마이닝	웹 사용자의 패턴을 분석하는 기법

4) 사회 연결망 분석(SNA: Social Network Analysis)

- 사회 연결망 분석은 사회 연결망 데이터를 활용하여 사회 연결망과 사회 구조 등을 사회과학적으로 분석하는 방법이다.
- 네트워크는 노드(Node)와 엣지(Edge)를 기반으로 사회적 관계를 구조화한 것으로 노드는 사회를 구성하는 개체를 의미하고, 엣지는 개체 간의 관계를 의미한다.

[사회 연결망 분석 절차]

데이터 수집 → 데이터 분석 → 데이터 시각화

3. 분석기법 적용 – 고급 분석기법

4) 사회 연결망 분석(SNA: Social Network Analysis)

● 사회 연결망 분석의 주요 속성으로는 응집력, 구조적 등위성, 명성, 범위, 중개가 있다.

사회 연결망 분석의 주요 속성

속성	설명
응집력(Cohesion)	개체들 간의 연결된 정도
구조적 등위성 (Equivalenca)	한 네트워크의 구조적 지위와 그 역할이 동일한 개체들 간의 관계
명성(Prominence)	네트워크 내에서 책임을 갖는 개체 확인
범위(Range)	개체의 네트워크 규모
중개(Brokerage)	다른 네트워크와 연결해주는 정도

사회 연결망 분석 측정 지표

측정 지표	설명
연결 중심성	한 노드가 얼마나 많은 노드와 관계를 맺고 있는지 측정하는 방식
군집 중심성	각 노드 간의 거리를 바탕으로 중심성을 측정하는 방식
근접 중심성	네트워크 내에서 특정 노드가 다른 노드들 사이에 위치하는 정도
위세 중심성 (=아이겐벡터 중심성)	자신의 연결 정도 중심성으로부터 발생하는 영향력과, 자신과 연결된 타인의 영향력을 합하여 결정하는 방법

3. 분석기법 적용 – 고급 분석기법

개념 체크

01 다음 중 비정형 데이터에 대한 설명으로 옳지 않은 것은?

① 사진은 RGB 방식으로 저장한다.

② 웹 문서 데이터는 크롤링 기술을 활용하여 수집한다.

③ 오디오 데이터는 토큰화하여 저장한다.

④ 텍스트 데이터는 자연어 처리하여 유의미한 정보를 추출할 수 있다.

오디오 데이터는 시간에 따른 진폭(Amplitude) 형태로 저장한다. 토큰화하여 저장하는 것은 텍스트 데이터다.

비정형 데이터 분석

● 비정형 데이터는 이미지, 영상, 문서 데이터와 같이 정형화된 데이터의 구조를 갖지 않는 데이터를 의미한다.

● 비정형 데이터 분석은 이러한 비정형 데이터를 분석하여 의미 있는 정보를 도출해 내는 분석을 의미한다.

02 사회관계망 분석(Social Network Analysis)에서 중심성 분석으로 적절하지 않은 것은?

① 근접 중심성 ② 매개 중심성

③ 아이젠벡터 중심성 ④ 포괄 중심성

사회관계망 분석(사회연결망 분석)에서 중심성 측정 지표는 연결 중심성, 근접 중심성, 매개 중심성, 위세 중심성(아이겐벡터 중심성) 이 있다.

사회 연결망 분석 지표

연결 중심성 : 한 노드가 얼마나 많은 노드와 관계를 맺고 있는지 측정하는 방식

근접 중심성 : 각 노드 간의 거리를 바탕으로 중심성을 측정하는 방식

매개 중심성 : 네트워크 내에서 특정 노드가 다른 노드들 사이에 위치하는 정도

위세 중심성(=아이겐벡터 중심성) : 자신의 연결 정도 중심성으로부터 발생하는 영향력과 자신과 연결된 타인의 영향력을 합하여 결정하는 방법

3. 분석기법 적용 – 고급 분석기법

03 다음 중 텍스트 마이닝에서 문장을 2개 이상의 단어로 분리하는 방법은?

① 토픽 모델링 ② TF-IDF

③ N-gram ④ Dendrogram

N-gram : 통계학 기반 언어 모델 중 하나로 n개의 연속된 단어 나열을 의미한다.

토픽 모델링 : 기계학습 및 자연어 처리 분야에서 추상적인 주제를 발견하기 위한 통계적인 모델 중 하나로 텍스트 본문의 숨겨진 의미 구조를 발견하기 위해 사용되는 모델

TF-IDF : 정보 검색과 텍스트 마이닝에서 이용하는 가중치 로 여러 문서로 이루어진 문서군이 있을 때, 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다.

덴드로그램(Dendrogram) : 계층 군집을 트리 구조로 시각화한 도구이다.

04 다음 중 텍스트 마이닝의 텍스트 벡터화 방법이 아닌 것은?

① TF-IDF ② Word Embedding

③ Word2Vec ④ POS Tagging

품사 태깅(POS Tagging) – 텍스트 전처리 기법

형태소(의미를 갖는 가장 작은 말은 단위)의 품사를 태깅하는 기술이다.

TF-IDF(Term Frequency-Inverse Document Frequency)

● 특정 단어가 문서 내에 등장하는 빈도(TF, 단어 빈도)와 그 단어가 문서 전체 집합에서 등장하는 빈도(IDF, 역문서 빈도)를 고려하여 벡터화하는 방법이다.

● 자주 사용된 단어라도 많은 문서에 등장하는 단어의 경우 IDF가 낮아지기 때문에 TF-IDF의 벡터화 결과 작은 값을 지닌다.

Word Embedding

● 분포가설(Distributional hypothesis) 개념을 바탕으로 의미를 포함하는 단어 벡터로 바꾸는 기법

● 분포가설에 의해 비슷한 분포를 가진 단어의 주변 단어 역시 비슷한 의미를 가질 것이라고 가정한다.

3. 분석기법 적용 – 고급 분석기법

07 앙상블 분석

● 앙상블(Ensemble)은 '조화, 통일'을 나타내는 프랑스어를 의미하고, 앙상블 분석은 여러 모형을 종합하여 최종적인 의사결정을 도출하는 분석 방법을 의미한다.

● 앙상블 분석은 분석 결과의 성능을 향상시키기 위해 다수의 모형에서 출력된 결과를 종합하여 하나의 최종 결과를 도출하는 분석 방법이다.

● 앙상블 분석 방법에는 배깅, 부스팅, 랜덤 포레스트, 보팅, 스태킹이 있다.

1) 배깅(Bagging, Bootstrap Aggregation)

● 부트스트랩(Bootstrap) 샘플링으로 추출한 여러 개의 표본에 각각 모형을 병렬적으로 학습하고, 추출된 결과를 집계(aggregation)하는 기법이다.

● 사이즈가 작거나 결측값이 있는 경우에 유리하고, 성능 향상에 효과적이다.

부트스트랩(Bootstrap) : 주어진 데이터에서 동일한 크기의 표본을 랜덤 복원 추출로 뽑은 데이터를 의미함
복원 추출(Sampling with Replacement) : 한 번 뽑은 표본을 모집단에 다시 넣고 다른 표본을 추출하는 방식

3. 분석기법 적용 – 고급 분석기법

2) 랜덤 포레스트(Random Forest)

- 의사결정나무 기반 앙상블 알고리즘으로 모든 속성(feature)들에서 임의로 일부를 선택하고, 그 중 정보 획득량이 가장 높은 것을 기준으로 데이터를 분할한다.
- 분류기를 여러 개 사용할수록 성능이 좋아지고, 예측편향을 줄이고, 과대 적합을 피할 수 있으며, 이상치의 영향을 적게 받는다.

3) 보팅(Voting)

- 여러 개의 분석 모형 결과를 조합하는 방법이다.
- 직접투표와 간접투표가 있다.

직접투표 (Hard Voting)	많이 선택된 클래스를 최종 결과로 예측한다.
간접투표 (Soft Voting)	각 모형의 클래스 확률값을 평균내어 확률이 가장 높은 클래스를 최종 결과로 예측하는 방법이다.

3. 분석기법 적용 – 고급 분석기법

4) 부스팅(Boosting)

- 예측력이 약한 모형들을 결합하여 예측력이 강한 모형을 만드는 알고리즘으로 분류가 잘못된 데이터에 가중치를 적용하여 표본을 추출하는 기법이다.
- 대용량 데이터 분석에 유리하고, 높은 계산 복잡도를 가진다.
- 알고리즘 : AdaBoost, GBM, XGBoost

알고리즘	설명
AdaBoost (Adaptive Boosting)	초기 모형을 약한 모형으로 설정하고, 매 과정마다 가중치를 적용하여 이전 모형의 약점을 보완하는 새로운 모형을 적합(fitting)하여 최종 모델을 생성한다.
GBM (Gradient Boosting Machine)	AdaBoost와 유사하나 가중치 업데이트 시에 경사하강법을 사용하는 알고리즘으로 과적합의 위험이 있다.
XGBoost (Extreme Gradient Boosting)	GBM의 단점인 과적합을 방지하기 위해 파라미터가 추가되어 병렬 학습이 가능한 알고리즘으로 회귀, 분류 문제에서 모두 사용 가능하다.

경사하강법(Gradient Descent) : 반복 수행을 통해 오류를 최소화할 수 있도록 가중치 업데이트 값을 도출하는 기법

3. 분석기법 적용 – 고급 분석기법

5) 스태킹(Staking)

- 여러 분석 모형의 예측값을 최종 모델의 학습 데이터로 사용하는 예측방법이다.
- 기본 스태킹 모델의 경우 과적합 발생 위험이 있어서 CV세트 기반의 스태킹 모델을 사용한다.

CV세트 기반의 스태킹 : 과적합 개선을 위해 최종 메타 모델(개별 모델의 예측된 데이터셋을 기반으로 학습하고 예측하는 방식)을 위한 데이터셋을 만들 때 교차검증(Cross validation) 기반으로 예측된 결과 데이터셋을 활용하는 방식
교차 검증(cross-validation) : 하나의 문제 또는 사건, 주장 같은 것을 서로 다른 시각에서 또는 여러 가지 자료를 토대로 정확성을 높이기 위해 행해지는 가장 기본적인 검사 방법이다.

3. 분석기법 적용 – 고급 분석기법

개념 체크

01 다음 중 예측력이 약한 모형을 연결하여 강한 모형을 만드는 기법으로 오분류된 데이터에 가중치를 주어 표본을 추출하지만 과적합의 위험이 있는 앙상블 기법은?

① 배깅 - AdaBoost

② 배깅 - 랜덤 포레스트

③ 부스팅 - 랜덤 포레스트

④ 부스팅 – GBM

예측력이 약한 모형을 연결하여 강한 모형을 만드는 기법으로 오분류된 데이터에 가중치를 주어 표본을 추출하지만, 과적합의 위험이 있는 앙상블 기법은 부스팅 기법의 GBM(Gradient Boosting Machine)에 대한 설명이다.

배깅(Bagging, Bootstrap Aggregation)

● 부트스트랩(Bootstrap) 샘플링으로 추출한 여러 개의 표본에 각각 모형을 병렬적으로 학습하고, 추출된 결과를 집계(Aggregation)하는 기법이다.

● 사이즈가 작거나 결측값이 있는 경우에 유리하고, 성능 향상에 효과적이다.

02 다음 중 랜덤 포레스트에 대한 설명으로 옳지 않은 것은?

① 분류기를 여러 개 쓸수록 성능이 좋아진다.

② 모델에 사용되는 모델의 개수가 많을수록 모델의 정확도가 높아진다.

③ 여러 개의 의사결정 트리가 모여서 랜덤 포레스트 구조 가 된다.

④ 이상치의 영향을 적게 받는다.

모델의 개수가 많아질 경우 과적합이 발생할 수 있다.

3. 분석기법 적용 – 고급 분석기법

03 다음 중 샘플링 데이터의 가중치를 조정하여 모델을 연속적으로 학습하여 오차를 줄이는 방법은?

① 배깅 ② 스태킹

③ 랜덤 포레스트④ 부스팅

스태킹(Staking)

● 여러 분석 모형의 예측값을 최종 모델의 학습 데이터로 사용하는 예측 방법이다.

● 기본 스태킹 모델의 경우 과적합 발생 위험이 있어서 CV세트 기반의 스태킹 모델을 사용한다.

04 다음 중 배깅에 대한 설명으로 옳지 않은 것은?

① 편향(Bias)이 낮은 과소적합(Underfit) 모델에 효과적 이다.

② 편향(Bias)이 높은 과대적합(Overfit) 모델에 효과적 이다.

③ 가중치를 활용하여 약 분류기를 강 분류기로 만드는 방법이다.

④ 훈련 데이터에서 다수의 부트스트랩 자료를 생성하고, 각 부트스트랩 자료를 결합하여 최종 예측 모형을 만드는 알고리즘이다.

가중치를 활용하여 약 분류기를 강 분류기로 만드는 알고리즘은 부스팅이다.

3. 분석기법 적용 – 고급 분석기법

05 다음 중 앙상블 분석에 대한 설명으로 옳지 않은 것은?

① 앙상블 분석 방법에는 배깅, 부스팅, 랜덤 포레스트, 보팅, 스태킹이 있다.

② 배깅(Bagging)은 데이터 사이즈가 크거나 결측값이 없는 경우에 사용하기 유리하다.

③ 부스팅 (Boosting)의 알고리즘에는 AdaBoost, GBM, XGBoost이 있다.

④ 간접투표(Soft Voting)는 각 모형의 클래스 확률값을 평균 내어 확률이 가장 높은 클래스를 최종 결과로 예측 하는 방법이다.

배깅은 사이즈가 작거나 결측값이 있는 경우에 사용하기 유리하고, 성능 향상에 효과적인 특징이 있다.

06 다음 중 랜덤 포레스트에 대한 설명으로 옳지 않은 것은?

① 랜덤 포레스트는 의사결정나무 기반 앙상블 알고리즘 이다.

② 이상치의 영향을 적게 받는다.

③ 분류기를 여러 개 사용할수록 예측편향이 줄어든다.

④ 랜덤 포레스트 모형에서는 모든 변수(Feature)를 학습 시킨다.

랜덤 포레스트는 의사결정나무 기반 앙상블 알고리즘으로 모든 속성(feature)들에서 임의로 일부를 선택하고, 그 중 정보 획득량이 가장 높은 것을 기준으로 데이터를 분할한다.

3. 분석기법 적용 – 고급 분석기법

08 비모수 통계

- 비모수 통계는 통계학에서 모수에 대한 가정을 전제로 하지 않고, 모집단의 형태에 관계없이 주어진 데이터 에서 직접 확률을 계산하여 통계학적 검정을 하는 분석 방법이다.
- 비모수통계 분석에는 빈도, 부호, 순위 등의 통계량이 사용되고, 이상값에 대한 영향이 적다.
- 비모수통계 검정 방법에는 부호 검정, 윌콕슨-부호 순위 검정, 만-위트니 U 검정, 윌콕슨 순위 합 검정, 크루스칼 왈리스 검정, 런 검정이 있다.

1) 부호 검정(Sign Test)

- 중앙값과의 차이를 부호(+, -)로 전환한 후, 검정한 뒤 부호만 사용하여 두 집단의 분포가 동일한지 검정하는 방법이다.
- 분포의 연속성, 독립성을 가정한다.

2) 윌콕슨-부호 순위 검정(Wilcoxon Singed Rank Test)

- 중앙값과의 차이를 부호뿐만 아니라 상대적인 크기도 고려하여 검정하는 방법이다.
- 분포의 연속성, 독립성, 대칭성을 가정한다.
- 윌콕슨 부호 순위 검정은 단일 표본 검정 기법이다.
- 윌콕슨 순위 합 검정은 이변수 검정 기법이다.

3. 분석기법 적용 – 고급 분석기법

3) 만-위트니 U 검정(Mann-Whitney U Test)

- 두 집단이 순위 척도 자료를 가진 집단이거나, 집단의 표본 수가 비교적 작을 때 두 집단의 차이를 분석하는 검정 방법이다.
- 분포의 연속성, 독립성, 대칭성을 가정한다..

4) 크루스칼왈리스 검정(Kruskal-Wallis Test)

- 세 집단 이상의 분포를 비교하는 검정 방법으로 집단별 평균이 아닌 중위수가 같은지 검정하는 방법이다.
- 분포의 중앙값은 다르나 동일한 형태의 분포를 가지는 것을 가정한다.

5) 런 검정(Run Test)

- 일련의 연속적인 관측값이 임의적으로 나타난 것인지 검정하는 방법으로, 런(Run)은 관측된 데이터에서 한 종류의 부호가 시작되고 끝나는 단위를 의미한다.

3. 분석기법 적용 – 고급 분석기법

개념 체크

01 다음 중 윌콕슨 부호 순위 검정과 윌콕슨 순위 합 검정 에 대한 설명으로 옳지 않은 것은?

① 윌콕슨 순위 합 검정은 모수 분포를 가정한 방법이다.

② 윌콕슨 부호 순위 검정은 단일 표본 검정 기법이다.

③ 윌콕슨 순위 합 검정은 이변수 검정 기법이다.

④ 윌콕슨 순위 합 검정은 자료의 분포에 대한 대칭성 가정이 필요하다.

윌콕슨 순위 합 검정은 비모수적 통계 방법이다.

윌콕슨 부호 순위 검정

● 윌콕슨 부호 순위 검정은 단일 표본에서 중위수에 대한 검정에 사용되며, 또한 대응되는 두 표본의 중위수의 차이 검정에 사용된다.

● 윌콕슨 부호 순위 검정은 일변량 검정이다.

● 주로 30개 이하의 작은 샘플일 때 사용한다.

●차이의 부호뿐만 아니라 차이의 상대적인 크기도 고려한 검정 방법이다.

● 자료의 분포가 연속적이고 독립적인 분포에서 나온

02 다음 중 비모수통계 검정으로 옳지 않은 것은?

① 피어슨 상관계수

② 부호 검정

③ 윌콕슨 부호 순위 검정

④ 만-위트니 검정

피어슨 상관계수는 모수통계 검정 방법이다.

비모수통계 검정 방법에는 부호 검정, 윌콕슨 부호 순위 검정, 만-위트니 U 검정, 윌콕슨 순위합 검정, 크루스칼 왈리스 검정, 런 검정이 있다.

부호 검정

● 중앙값과의 차이를 부호(+, -)로 전환한 후, 검정한 뒤 부호만 사용하여 두 집단의 분포가 동일한지 검정하는 방법

● 분포의 연속성, 독립성을 가정한다.

만-위트니 U 검정

● 두 집단이 순위 척도 자료를 가진 집단이거나, 집단의 표본의 수가 비교적 작을 때 두 집단의 차이를 분석하는 검정 방법이다.

● 분포의 연속성, 독립성, 대칭성을 가정한다.

3. 분석기법 적용 – 고급 분석기법

예상 문제

01 다음 중 다차원척도법에 대한 설명으로 옳지 않은 것은?

① 개체들 사이의 유사성, 비유사성을 측정하여 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사이의 집단화를 시각적으로 표현하는 방법이다.

② 공분산행렬을 사용하여 고윳값이 1보다 큰 주성분의 개수를 이용한다.

③ 스트레스 값이 0에 가까울수록 적합도가 좋다.

④ 유클리드 거리와 유사도를 이용하여 개체 간의 거리를 구한다.

다차원척도법(MDS; Multi Dimensional Scaling)

● 다차원척도법은 개체 간의 근접성을 시각화 하는 통계 기법이다.

● 개체들 사이의 유사성, 비유사성을 측정하여, 개체들을 2차원 혹은 3차원 공간상의 점으로 표현하는 분석방법이다.

● 스트레스 값이 0에 가까우면 적합도가 높고, 1에 가까우면 적합도가 낮다.

● 다차원척도법에서 개체들의 거리를 계산할 때는 유클리드

03 다음과 같이 주어진 표에 대한 해석으로 옳은 것은?

약	조기 암 환자		말기 암 환자		전체 암 환자	
	생존	사망	생존	사망	생존	사망
A	14	8	6	12	20	20
B	7	3	9	21	16	24

(생존율 : 생존/(생존+사망), 사망률 : 100-생존율)

① 조기 암 환자 생존율은 A약이 더 높다.

② A약과 B약의 전체 암 환자 생존율의 차이는 25%이다.

③ 조기, 말기 암 환자 모두에게 A약의 효과가 더욱 높았다.

④ A약의 전체 암 환자 생존율은 50%이다.

위의 주어진 데이터의 생존율을 계산을 해보자.

	조기	말기	전체
A	14/22=63.6%	6/18=33.3%	20/40=50%
B	7/10=70%	9/30=30%	16/40=40%

04 다음 중 독립변수와 종속변수의 유형에 따른 분석 방법으로 적합하지 않은 것은?

① 공분산 분석(ANCOVA)은 종속변수가 범주형, 독립변수가 연속형인 분석 방법이다.

② T-검정은 종속변수가 수치형이고, 2개 범주의 독립변수 를

3. 분석기법 적용 – 고급 분석기법

05 다음 중 기존 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터들을 저차원 공간으로 변환하는 분석 방법은?

① 상관분석 ② 회귀 분석

③ 주성분 분석 ④ 분산 분석

주성분 분석(PCA; Principal Component Analysis)

● 데이터 전체 변동을 최대한 유지, 보존해 주는 주성분을 생성하는 차원 축소 방법이다

● 주성분 분석의 목적은 차원 축소와 다중공선성 해결이다.

● 누적기여율이 85%이상이면 주성분의 수로 결정할 수 있다.

● 주성분 분석의 절차는 축 생성 -> 생성된 축에 데이터 투영 -> 차원 축소 순으로 이루어진다.

06 다음과 같은 이원 분할표를 기준으로 상대위험도(RR)를 계산하면 얼마인가?

구분	질환 발생	질환 미발생	합계
음주	10	30	40
비 음주	70	60	130

07 다음 자기회귀 누적 이동평균 모형(ARIMA)에 대한 명칭 중 틀린 것은?

① ARIMA(0,0,0) : 다중잡음 모형

② ARIMA(0,1,0) : 확률보행 모형

③ ARIMA(p,0,0) : 자기회귀 모형

④ ARIMA(0,0,q) : 이동평균 모형

자기회귀 누적 이동평균 모형에서 ARIMA(0,0,0)는 백색잡음 모형이다.

ARIMA(p, d, q)

p : AR 관련, d = 몇 번 차분했는지, q : MA관련

ARIMA(0,0,0) : 백색잡음 모형

ARIMA(0,1,0) : 확률잡음 모형

ARIMA(p,0,0) : 자기회귀 모형

ARIMA(0,0,q) : 이동평균 모형

백색잡음 모형 : 시점에 상관없이 평균이 0이고, 분산이 시그마 제곱인 시계열 자료를 의미하며, 정상 시계열의 대표적인 예이다.

확률보행 모형 : 데이터가 정상성을 나타내지 않는 모델로

3. 분석기법 적용 – 고급 분석기법

09 다음 중 시계열 분해 요소가 아닌 것은?

① 추세 요인

② 계절 요인

③ 순환 요인

④ 공통 요인

시계열 분해 구성 요소에는 추세, 계절성, 순환, 불규칙 요인이 있다.

추세(Trend) : 데이터가 장기적으로 증가 혹은 감소하는 것으로 추세가 꼭 선형일 필요는 없다.

계절성(Seasonal) : 주, 월, 분기, 반기 등 특정 시간의 주기로 나타나는 패턴

순환(Cycle) : 경기변동과 같이 정치, 경제, 사회적 요인에 의한 변화로 일정 주기가 없는 장기적인 변화 현상

불규칙 요인(Irregular Factor) : 설명될 수 없는 요인 또는 돌발적인 요인에 의해 일어나는 변화로 예측이 불가능한 임의의 변동

10 비정상 시계열에 대한 시계열 모델로서 자기회귀누적 이동평균 모형(ARIMA)에 대한 설명으로 적절하지 않은 것은?

11 다음 설명하는 시계열에 대한 명칭은?

주, 월, 분기, 반기 단위 등 특정 시간의 주기로 나타나는 패턴

① 추세 ② 계절

③ 주기 ④ 불규칙

12 시계열 분석은 정상성을 만족해야 한다. 정상성은 시점 에 상관없이 시계열의 특성이 일정하다는 것을 의미한다. 다음 중 비정상 시계열에 대한 설명이 아닌 것은?

① 평균이 일정하지 않다.

② 분산이 시점에 의존한다.

③ 백색잡음 과정은 대표적인 비정상 시계열이다.

④ 공분산은 시차와 시점에 의존한다.

백색잡음은 시점에 상관없이 평균이 0이고, 분산이 시그마 제곱인 시계열 자료를 의미하고, 정상 시계열의 대표적인 예이다.

비정상 시계열의 대표적인 예로는 확률 보행(Random Walk)이 있다. 확률 보행은 임의의 방향으로 연속적인 걸음이

3. 분석기법 적용 – 고급 분석기법

13 다음 중 ARIMA에 대한 설명으로 옳지 않은 것은?

① ARMA의 일반화 형태이다.

② 일간, 주간, 월간 단위로 예측이 가능하다.

③ AR 모델은 변수의 과거 값을 사용한다.

④ 백색잡음은 독립적이지 않다.

백색잡음 모형 ARIMA(0,0,0)는 대표적인 정상 시계열로써 독립적이고 동일한 분산을 갖는다.

14 다음 중 시계열 데이터 예측 방법에 대한 설명으로 옳지 않은 것은?

① 시계열 데이터 예측 방법은 확률적 방법과 고전적 방법으로 나뉜다.

② 지수평활법은 과거 값에 가중치를 두고, 최근 값에 적은 비중을 두는 방법이다.

③ 이동평균법은 일정 기간의 관측치를 이용하여 평균을 구하고, 이를 이용해 예측하는 방법이다.

④ 확률적 방법은 주파수 영역과 시간 영역으로 나뉜다.

지수평활법은 최근 값에 많은 가중치를 두어 미래를 예측하는 방법이다.

15 다음은 어떤 알고리즘을 설명한 것인가?

합성곱과 풀링 과정을 거쳐 데이터를 분석하는 알고리즘으로 주로 시각적 이미지 분석에서 많이 사용된다.

합성곱은 원본 이미지로부터 특징을 추출하는 과정으로 필터를 활용하여 유사한 이미지 영역을 강조하는 특성 맵(Feature Map)을 출력한다.

풀링은 합성곱 과정을 거친 데이터를 요약하는 작업으로, 추출한 특징은 유지하면서 데이터의 사이즈를 줄여주는 과정이다.

① CNN ② RNN

③ DNN ④ KNN

합성곱 신경망인 CNN 알고리즘에 대한 설명이다.

16 다음과 같은 특징을 갖는 알고리즘은?

언어 데이터, 시계열 데이터 등과 같이 연속적인 데이터 분석에 특화된 알고리즘으로 과거 데이터를 기반으로 현재 데이터를 학습하는 특징이 있다.

이 알고리즘은 장기 의존성 문제와 기울기 소실 문제

3. 분석기법 적용 – 고급 분석기법

17 CNN 알고리즘에서 입력층 원본 이미지가 5×5에서 Stride가 1이고 필터가 3×3일 때, Feature Map은 얼마인가?

① (1,1) ② (2,2)

③ (3,3) ④ (4,4)

CNN Feature Map 계산

스트라이드(지정된 간격으로 필터를 순회하는 간격)가 적용되었을 때, 원본 이미지의 크기가 n * n, 스트라이드가 s, 패딩이 p, 필터가 f * f 일 때, 피처맵의 크기를 구하는 공식은 다음과 같다.

Feature Map = $\left(\frac{n+sp-f}{s}+1, \frac{n+sp-f}{s}+1\right) = \left(\frac{n+sp-f}{s}+1\right) * \left(\frac{n+sp-f}{s}+1\right)$

$$n = 5, p = 0, s = 1, f = 3$$

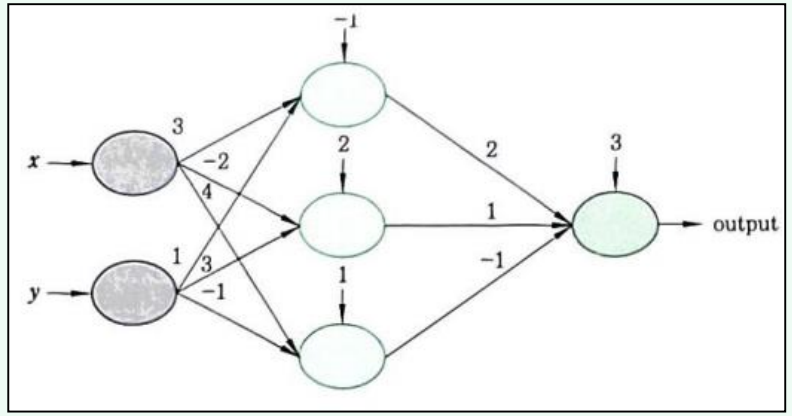
$= \left(\frac{5+0-3}{1}+1, \frac{5+0-3}{1}+1\right) = (3, 3)$이 된다.

18 다음의 각 과제에 대한 분석 방법이 적절하게 연결된 것은?

가. 영화 감상평에 대한 긍정/부정 판단

나. 사원증 대신 얼굴 인식으로 출입 가능한 보안

19 다음의 신경망에서 활성화 함수로 항등 함수를 사용한다고 한다. 입력값이 (x=1, y=2)일 때 출력값은 얼마인가?



① 12 ② 13

③ 14 ④ 15

항등 함수(Identity Function)는 입력값을 그대로 출력해 주는 함수이다.

(1 * 3 + 2 * 1 – 1) * 2 + (1 * -2 + 2 * 3 + 2) * 1 +

(1 * 4 + 2 * -1 + 1) * -1 + 3 = 14가 된다.

20 다음 중 윌콕슨 부호 순위 검정과 윌콕슨 순위 합 검정 에 대한 설명으로 옳지 않은 것은?

① 윌콕슨 순위 합 검정은 모수 분포를 가정한 방법이다.

② 윌콕슨 부호 순위 검정은 단일 표본 검정 기법이다.

3. 분석기법 적용 과목 마무리 문제

마무리 문제

01 다음은 어떤 분석 방법에 대한 설명인가?

하나 이상의 독립변수(X)가 종속변수(Y)에 끼치는 영향을 분석하는 통계기법

독립변수와 종속변수는 선형적인 관계를 갖고, 독립변수를 통해 종속변수를 예측

예) 눈이 올 때 교통사고 발생 확률 분석

① 시계열 분석 ② 회귀 분석

③ 분산 분석 ④ 다중 분석

시계열 분석 : 시간의 흐름에 따라 관측된 과거 데이터를 분석하여 미래의 데이터를 예측하는 분석 기법으로 기온 예측, 가격 예측 등에 사용된다.

분산 분석(ANOVA, Analysis of Variance)

서로 다른 집단의 평균에서 분산값(총 평균과 각 집단 간의 평균 차이에 의해 생긴 분산)을 비교하여 집단 간의 통계학적 차이를 확인하는 방법이다.

02. 다음 중 회귀 분석 가정에 속하지 않는 것은?

① 분산성 ② 선형성

03. 다음 수식이 설명하는 회귀 분석 유형은?

$$Y=aX_1+bX_2+cX_1^2+\cdots+dX_2^2+eX_1X_2+f$$

(독립변수가 2개이고, 2차 함수인 경우)

① 비선형회귀 ② 곡선회귀

③ 단순선형회귀 ④ 다항회귀

회귀 분석 유형

단순 선형회귀 : 독립변수(X)가 1개이고 종속변수와의 관계가 직선인 경우

다중 선형회귀 : 독립변수가 k개이고 종속변수와의 관계가 선형인 경우(1차 함수)

다항 회귀 : 독립변수와 종속변수와의 관계가 1차 함수 이상인 경우

곡선 회귀 : 독립변수가 1개이며, 종속변수와의 관계가 2차 곡선이거나 3차 곡선인 경우

비선형 회귀 : 회귀식의 모양이 미지의 모수들의 선형관계 로 이루어져 있지 않은 경우

04 다음 중 단순선형회귀 분석에 대한 설명으로 틀린 것은?

① 독립변수와 종속변수가 각각 두 개씩 존재하고, 오차항 이

3. 분석기법 적용 과목 마무리 문제

05 다음 설명에 해당하는 명칭은?

설명변수들 사이에 선형관계가 존재하게 되면
회귀계수의 정확한 추정이 어려워지는 것을 의미한다.
이 경우 문제가 있는 변수를 제거하거나 주성분 회귀,
지능형 회귀 모형을 적용하여 문제를 해결할 수 있다.

① 기울기 소실문제 ② 과소적합

③ 다중공선성 ④ 과대적합

기울기 소실(Gradient Vanishing)

역전파(Backprogation) 과정에서 입력층으로 갈수록 기울기가 점점 작아지는 현상

06 다음 중 설명이 틀린 것은?

① AIC는 실제 데이터의 분포와 모형이 예측하는 분포 간의 차이를 나타내는 방법이다.

② BIC는 표본의 크기가 커질수록 복잡한 모형을 더욱 강하게 제한할 수 있다.

③ AIC의 수식은 $-2\ln(L)+2p$이다.

④ 모형의 복잡도에 패널티(벌점)를 적용하는 방법으로 AIC 방법, BIC 방법, DIC 방법이 있다.

07 다음 중 로지스틱 회귀 분석에 대한 설명으로 옳은 것은?

① 로지스틱 회귀 분석은 어떤 사건이 발생할지에 대한 직접적인 예측이 아닌 그 사건이 발생할 확률을 예측하는 방법이다.

② 로지스틱 회귀 분석은 독립변수와 종속변수가 모두 수치형일 때 사용 가능하다.

③ 로지스틱 회귀는 정규분포를 따른다.

④ 로지스틱 회귀 수식은 $Y=\frac{1}{1+e^{X}}$ 이다.

로지스틱 회귀 분석은 독립변수가 수치형이고,
반응변수(종속변수)가 범주형일 때 사용되는 분석 모형이다.
로지스틱 회귀는 이항 분포에 따른다.

로지스틱 회귀 수식은 $Y=\frac{1}{1+e^{-X}}$ 이다.

08 당분 섭취에 따른 당뇨병 발생 결과가 다음과 같다고 할 때 당분 섭취에 따른 당뇨병 발생률에 대한 승산비(Odds)는 얼마인가?

구분	당뇨병 발생	당뇨병 미발생
당분 섭취	80	9

3. 분석기법 적용 과목 마무리 문제

09 다음 중 의사결정나무의 구성 요소가 아닌 것은?

① 연결 마디 ② 뿌리 마디

③ 중간 마디 ④ 부모 마디

의사결정나무의 구성 요소는 부모 마디, 자식 마디, 뿌리 마디, 끝 마디, 중간 마디, 가지, 깊이가 있다.

부모 마디(Parent Node) : 자식 마디의 상위 마디

자식 마디(Child Node) : 하나의 마디로부터 분리되어 있는 2개 이상의 마디

뿌리 마디(Root Node) : 전체 데이터로 시작점이 되는 마디

끝 마디(Terminal Node) : 자식 마디가 없는 가장 하위 마디(=잎 노드)

중간 마디(Internal Node) : 부모 마디와 자식 마디가 모두 있는 마디

가지(Branch) : 마디를 이어주는 연결선

깊이(Depth) : 뿌리 마디에서 끝 마디까지 가지를 이루는 마디의 수

10 다음 중 의사결정나무에 대한 설명으로 옳지 않은 것은?

① 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측

11 다음 중 ReLU 함수의 뉴런이 죽는 Dying ReLU 현상을 해결한 활성화 함수는?

① Sigmoid 함수

② tanh 함수

③ ReLU 함수

④ Leaky ReLU 함수

ReLU 함수의 뉴런이 죽는 Dying ReLU 현상을 해결한 활성화 함수는 Leaky ReLU 함수이다.

시그모이드 함수(Sigmoid Function)

● 로지스틱 회귀 함수에 로짓 변환을 한 형태이다.

● 기울기 소실 문제의 원인이 된다.

tanh(Hyperbolic Tangent Function)

● 시그모이드 함수의 확장된 형태이다.

● 시그모이드보다 학습 속도가 빠르다.

ReLU 함수(ReLU Function)

● 양수 입력 시 어떠한 값의 변형 없이 입력값 그대로 출력하고, 음수 입력 시 항상 0 값을 리턴하는 함수이다.

● 시그모이드 함수의 기울기 소실 문제를 해결한다.

3. 분석기법 적용 과목 마무리 문제

13 다음 수식이 의미하는 거리는?

$$d(i,j)=\sqrt{\sum_{f=1}^{n}(x_{if}-x_{jf})^2}$$

① 민코프스키 거리

② 마할라노비스 거리

③ 유클리드 거리

④ 맨하탄 거리

위의 수식은 유클리드 거리 공식이다.

연속형 변수 거리

수학적 거리

유클리드 거리 : 두 점간 차를 제곱하여 모두 더한 값의 양의 제곱근

맨하탄 거리 : 두 점간 차의 절댓값을 합한 값

민코프스키 거리 : m차원 민코프스키 공간에서의 거리

m = 1 일 때, 맨하탄 거리 동일

m = 2 일 때, 유클리드 거리와 같음

통계적 거리

표준화 거리 : 변수의 측정 단위를 표준화한 거리, D라는 표본

15 다음과 같은 두 집단의 질병 발생 확률 분할표에서 상대위험도(RR)는 얼마인가?

구분	질병 발생	질병 미발생	합계
흡연	10	30	40
비흡연	40	20	60
합계	50	50	100

① 2/3 ② 3/5

③ 3/8 ④ 5/6

상대위험도(RR) $= \frac{A\text{집단의 위험률}}{B\text{집단의 위험률}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{10}{40}}{\frac{40}{60}} = \frac{600}{1600} = \frac{3}{8}$

16 다음은 어떤 분석 방법을 설명한 것인가?

부트스트랩(Bootstrap) 샘플링으로 추출한 여러 개의 표본에 각각 모형을 병렬적으로 학습하고 추출된 결과를 집계(aggregation)하는 기법이다.
사이즈가 작거나 결측값이 있는 경우에 유리하고, 성능 향상에 효과적이다.

① GBM ② 스태킹

③ 부스팅 ④ 배깅

3. 분석기법 적용 과목 마무리 문제

17 다음은 어떤 분석 방법을 설명하는 것인가?

주어진 데이터를 K개의 군집으로 묶는 알고리즘으로 군집 수를 K개만큼 초깃값으로 지정하고, 각 객체를 가까운 초깃값에 할당하여 군집을 형성하는 방법이다.

각 군집의 평균을 재계산하여 초깃값을 갱신하는 과정을 반복하여 K개의 최종군집을 형성한다.

① CNN ② KNN

③ RNN ④ K-means clustering

CNN(합성곱 신경망) 알고리즘

● CNN은 합성곱(Convolution)과 풀링(Pooling)과정을 거쳐 데이터를 분석하는 알고리즘으로 주로 시각적 이미지 분석에 많이 사용한다.

KNN : K-최근접 이웃 알고리즘으로 지도 학습 분석 모형에 속한다.

RNN(순환신경망) 알고리즘

● RNN은 언어 데이터, 시계열 데이터 등과 같이 연속적인 데이터 분석에 특화된 알고리즘으로 과거 데이터를 기반으로

19 다음 중 설명이 옳지 않은 것은?

① 서포트 벡터 머신은 하드 마진 SVM과 소프트 마진 SVM으로 나뉜다.

② 대부분의 경우 하드 마진 SVM을 사용한다.

③ 하드 마진 SVM은 마진의 안쪽 또는 바깥쪽에 잘못 분류 된 데이터가 포함되는 것을 허용하지 않는 모델이다.

④ 소프트 마진 SVM은 마진의 안쪽 또는 바깥쪽에 잘못 분류된 데이터가 포함되는 것을 허용하는 모델이다.

대부분의 경우 소프트 마진 SVM을 사용한다.

20 과일 판매에 대한 데이터가 다음과 같을 때 [오렌지, 사과 → 자몽]에 대한 신뢰도와 지지도는 얼마 인가?

[오렌지, 사과, 딸기]
[오렌지, 사과, 자몽]
[오렌지, 바나나]
[사과, 바나나, 딸기]
[오렌지, 사과, 바나나, 자몽]

60%

② 신뢰도=66.6%, 지지도=40%

③ 신뢰도=33.3%, 지지도=40%

감사합니다.