3과목.빅데이터 모델링

(Ch_02. 분석기법 적용 - SEC 02. 고급 분석기법-1)

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

01 범주형 자료 분석

- 범주형 자료 분석은 독립변수와 종속변수가 모두 범주형이거나 두 변수 중 하나가 범주형일 때 사용하는 분석 방법이다.
- 범주형 변수는 주어진 데이터의 순서가 없는 명목형 변수와 순서가 있는 순서형 변수로 나뉜다.
- 범주형 자료 분석은 각 집단의 비율 차이를 비교하고자 할 때 사용된다.

변수에 따른 데이터 분석 방법

독립변수	종속변수 (반응변수)	분석 방법
범주형	범주형	분할표 분석, 카이제곱 검정(교차검정), 피셔의 정확검정
범주형	수치형	T-검정, 분산 분석
수치형	범주형	로지스틱 회귀 분석

3. 분석기법 적용 – 고급 분석기법

1) 분할표 분석(Contingency table analysis)

- 분할표 분석은 상대위험도와 승산비를 활용하여 분할표를 만든 뒤 변수간의 상호 관련성을 분석하는 방법이다.
- 범주형 데이터의 개수에 따라 일원(One-way)분할표(1개), 이원(Two-way)분할표(2개), 다원(Multi-way) 분할표(3개 이상)로 나뉜다.

이원 분할표 예

구분	A기종 사용자	B기종 사용자	합계
남자	50	60	110
여자	70	50	120
합계	120	110	230

3. 분석기법 적용 – 고급 분석기법

① 상대위험도(RR : Relative Risk)

▶ 상대위험도는 위험인자에 노출된 A집단의 사건 발생 확률을 위험인자에 노출되지 않은 B집단의 사건 발생 확률로 나눈 값이다.

▶ 상대위험도는 다음과 같이 계산할 수 있다.

$$\text{상대위험도(RR)} = \frac{A\text{집단의 위험률}}{B\text{집단의 위험률}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

A집단과 B집단의 사건 발생 확률 분할표

구분	사건 발생	사건 미발생	합계
A집단 (위험인자에 노출된 경우)	a	b	$a+b$
B집단 (위험인자에 노출되지 않은 경우)	c	d	$c+d$
합계	$a+c$	$b+d$	$a+b+c+d$

상대위험도 결과

상대위험도 값	설명
$RR < 1$	A집단의 사건 발생 확률이 낮음
$RR = 1$	집단과 사건 발생 확률이 연관성이 없음
$RR > 1$	A집단의 사건 발생 확률이 높음

3. 분석기법 적용 – 고급 분석기법

② 승산비(Odds Ratio, 교차비, 오즈비)

▶ 승산비(Odds Ratio)는 특정 사건이 발생할 확률(p)과 그 사건이 발생하지 않을 확률($1-p$)의 비를 의미한다.

$$\frac{\text{사건이 발생할 확률}}{\text{사건이 발생하지 않을 확률}} = \frac{p}{1-p}$$

▶ 승산비는 위험인자에 노출된 A집단의 승산비를 위험인자에 노출되지 않은 B집단의 승산비로 나눈 값이다.

$$\text{승산비}(OR) = \frac{\text{위험인자에 노출되었을 때 } \frac{\text{질병 발생수}}{\text{질병 미발생 수}}}{\text{위험인자에 노출되지 않았을 때 } \frac{\text{질병 발생수}}{\text{질병 미발생 수}}} = \frac{A\text{집단 사건 발생 확률}}{B\text{집단 사건 발생 확률}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

A집단과 B집단의 사건 발생 확률 분할표

구분	사건 발생	사건 미발생	합계
A집단 (위험인자에 노출된 경우)	a	b	$a+b$
B집단 (위험인자에 노출되지 않은 경우)	c	d	$c+d$
합계	$a+c$	$b+d$	$a+b+c+d$

승산비 결과

상대위험도 값	설명
$OR < 1$	A집단의 사건 발생 확률이 낮음
$OR = 1$	집단과 사건 발생 확률이 연관성이 없음
$OR > 1$	A집단의 사건 발생 확률이 높음

3. 분석기법 적용 – 고급 분석기법

③ 상대위험도(RR)와 승산비(OR)의 활용

▶ 상대위험도와 승산비는 주로 의학 분야에서 위험인자와 질환 발생과의 연관성을 확인하기 위한 분석 방법으로 자주 사용된다.

▶ 대표적인 예로 위험인자에 노출된 경우(A집단)와 그렇지 않은 경우(B집단)에서 질환이 발생된 경우와 그렇지 않은 경우를 확률 분할표로 나눈 후, 각 범주에 따른 연관성을 분석할 수 있다.

예시) 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

3. 분석기법 적용 – 고급 분석기법

2) 카이제곱 검정(Chi-Squared Test)

- 카이제곱 검정은 범주형 자료 간의 차이를 분석하는 모수적 통계 방법이다.

  참고) 평균, 표준편차, 분산 등을 통계량(statistic, 표본의 특성을 수치로 나타낸 것)이라고 하고,
  모집단의 모평균, 모표준편차, 모분산 등을 모수(parameter, 모집단의 특성을 수치로 나타낸 것)
  라고 한다.

- 카이제곱 검정은 적합도 검정, 독립성 검정, 동질성 검정으로 분류한다.

모수적 통계방법(parametric method) : 정규성을 갖는다는 모수적 특성을 이용하는 통계적 방법
비모수적 통계방법(nonparametric method) : 정규분포를 따르지 않거나 각 집단 간 10명 미만의 소규모 집단인 경우 자료를 크기 순으로 배열하여 순위를 매기고, 순위의 합을 통해 차이를 비교하는 순위합 검정을 적용하는데 이와 같이 모수의 특성을 사용하지 않는 통계적 방법이며, 데이터 샘플의 크기가 매우 작은 경우에도 사용할 수 있으며, 순위와 부호를 기반으로 하여 이상치의 영향을 받지 않는다.아울러 모집단의 분포에 대한 가정을 필요로 하지 않는다. 데이터가 연속형이 아닌 경우에도 사용이 가능하다.

3. 분석기법 적용 – 고급 분석기법

2) 카이제곱 검정(Chi-Squared Test)

① 적합도 검정(Goodness of Fit Test)

- ▶ 변수가 1개이고, 그 변수가 2개 이상의 범주로 구성되어 있을 때 사용하는 일변량 분석 방법이다.
- ▶ 표본집단의 분포가 주어진 특정 분포를 따르고 있는지 검정하는 방법이다.

② 독립성 검정(Test of Independence)

- ▶ 변수가 두 개 이상의 범주로 분할되어 있을 때 사용되며, 각 범주가 서로 독립적인지 연관성이 있는지 검정하는 방법이다.

③ 동질성 검정(Test of Homogeneity)

- ▶ 하나의 범주형 변수를 기준으로 각 그룹이 특정 요인에 대해 서로 비슷한지 알아보는 방법이다.

3. 분석기법 적용 – 고급 분석기법

3) T-검정(T-Test)

● T-검정은 두 집단의 평균을 비교하는 모수적 통계방법으로 표본이 정규성, 등분산성, 독립성 등을 만족할 경우 사용 가능하다.

● T-검정에는 단일표본 T-검정, 대응표본 T-검정, 독립표본 T-검정이 있다.

① 단일표본 T-검정

▶ 하나로 구성된 모집단의 평균값을 기준값과 비교하고자 할 때 사용하는 분석 방법이다.

예) 전국 고등학교 3학년 학생의 평균 키가 170cm일 때, 대구 한국고등학교 3학년 1반 학생 30명의 평균 키를 측정하여 전국의 평균 키와 비교하는 방법

② 대응표본 T-검정

▶ 동일한 표본의 A시점과 B시점을 비교하고자 할 때 사용하는 분석 방법이다.

예) 동일한 교수법의 효과를 비교하기 위해 동일한 교수법으로 학생들을 지도한 뒤 중간고사와 기말고사 시험 성적을 비교하는 방법

③ 독립표본 T-검정

▶ 독립된 두 집단의 평균 차이를 검정하는 분석 방법이다.

예) 제품 브랜드에 따른 소비자의 만족도 조사

3. 분석기법 적용 – 고급 분석기법

4) 피셔의 정확 검정(Fisher's Exact Exam)

- 피셔의 정확 검정은 표본 수가 적을 때 사용하는 카이제곱 검정 방법이다.
- 피셔의 정확 검정은 범주형 데이터의 기대빈도가 5 미만인 셀이 20%를 넘는 경우 사용한다.

기대빈도(Expected Counts) : 두 변수가 독립일 경우 이론적으로 기대할 수 있는 빈도의 분포

3. 분석기법 적용 – 고급 분석기법

02 다변량 분석(Multivariate analysis)

- 여러 현상이나 사건에 대한 관측치를 개별적으로 분석하지 않고 동시에 분석하는 통계적인 기법이다.
- 각 변수를 개별적이 아닌 동시에 분석하여 여러 변수들 간의 상관성을 고려한다.

1) 상관관계 분석

- 두 변수 사이에 어떠한 선형적 관계를 갖는지 분석하는 기법으로 상관계수(Correlation coefficient,r)를 계산하여 변수들 간의 상관관계를 분석하는 방법이다.
- 상관계수(r)는 -1~1의 범위를 갖는다. 상관계수가 1인 경우 강한 양의 상관관계를 갖고, -1인 경우 강한 음의 상관관계이며, 0인 경우 상관관계가 없음을 의미한다.

2) 다차원척도법(MDS: Multi Dimensional Scaling)

- 다차원척도법은 개체 간의 근접성을 시각화하는 통계기법이다.
- 개체들 사이의 유사성, 비유사성을 측정하여 개체들을 2차원 혹은 3차원 공간상의 점으로 표현하는 분석 방법이다.
- 스트레스 값이 0에 가까우면 적합도가 높고, 1에 가까우면 적합도가 낮다.
- 다차원척도법에서 개체들의 거리를 계산할 때는 유클리드 거리 행렬을 이용한다.

3. 분석기법 적용 – 고급 분석기법

3) 다변량 분산 분석(MANOVA: Multivariate Analysis of Variance)

- 종속변수가 2개 이상일 때 사용하는 분석 방법으로 종속변수(Y) 간의 공분산을 사용하여 다수의 종속변수들에서 집단 간의 차이가 있는지 검정하는 방법이다.
- 종속변수가 벡터의 형태로 주어지기 때문에 모집단의 평균 벡터 사이에 차이가 있는지 여부를 판단하는 것이 중요하다.

4) 주성분 분석(PCA: Principal Component Analysis)

- 데이터 전체 변동을 최대한 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 순이다.

공분산 : 두 개의 확률 변수의 선형관계를 나타내는 값이다. 한 확률 변수의 증감에 따른 다른 확률 변수의 증감의 경향에 대한 측도이다. 쉽게 말해 분산이라는 개념을 확장하여 두 개의 확률 변수의 흩어진 정도를 공분산이라고 하는 것이다.
다중공선성 : 다중 회귀 분석에서 독립(설명)변수들 간의 선형관계가 존재하면 회귀계수의 정확한 추정이 난해해지는 문제
누적기여율 : 주성분을 고유 값의 내림차순으로 정렬하여 상위 개의 주성분으로 설명할 수 있는 정보량의 비율

3. 분석기법 적용 – 고급 분석기법

개념 체크

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 시각화하는 통계기법이다.

● 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

02 다음 중 비모수적 추론에 대한 설명으로 적절하지 않은 것은?

① 데이터 샘플의 크기가 매우 작은 경우에도 사용할 수 있다.

② 순위와 부호를 기반으로 하여 이상치의 영향을 받지 않는다.

③ 모집단의 분포에 대한 가정을 필요로 하지 않는다.

④ 데이터가 연속형 측정값인 경우에만 사용할 수 있다.

비모수적 통계방법(nonparametric method) : 정규분포를 따르지 않거나 각 집단 간 10명 미만의 소규모 집단인 경우 자료를 크기 순으로 배열하여 순위를 매기고, 순위의 합을 통해 차이를 비교하는 순위합 검정을 적용하는데 이와 같이 모수의 특성을 사용하지 않는 통계적 방법이며, 데이터 샘플의 크기가 매우 작은 경우에도 사용할 수 있으며, 순위와 부호를 기반으로 하여 이상치의 영향을 받지 않는다. 아울러, 모집단의 분포에 대한 가정을 필요로 하지 않는다. 데이터가 연속형이 아닌 경우에도 사용이 가능하다.

3. 분석기법 적용 – 고급 분석기법

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 사용하여 분석하는 방법이다.

③ 로짓 모형은 종속변수가 범주형이고, 독립변수가 수치형 또는 범주형일 때 사용하는 분석 방법이다.

④ 카이제곱 검정은 독립변수와 종속변수가 모두 범주형일 때 사용하는 분석 방법이다.

공분산 분석(ANCOVA)은 종속변수가 연속형이고, 독립변수가 범주형일 때 사용하는 분석 방법이다.

변수에 따른 데이터 분석 방법

독립변수	종속변수	분석방법
범주형	범주형	분할표 분석, 카이제곱 검정, 피셔의 정확검정
범주형	수치형	T-검정, 분산 분석
수치형	범주형	로지스틱 회귀 분석

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
합계	80	90	170

① 1/8 ② 8/17

③ 13/17 ④ 13/28

상대위험도(RR : Relative Risk)

● 상대위험도는 위험인자에 노출된 A집단의 사건 발생 확률을 위험인자에 노출되지 않은 B집단의 사건 발생 확률로 나눈 값이다.

$$\text{상대위험도}(RR) = \frac{A\text{집단의 위험률}}{B\text{집단의 위험률}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{10}{10+30}}{\frac{70}{70+60}}$$

$$= \frac{1300}{2800} = \frac{13}{28} \text{ 이 된다.}$$

3. 분석기법 적용 – 고급 분석기법

03 시계열 분석

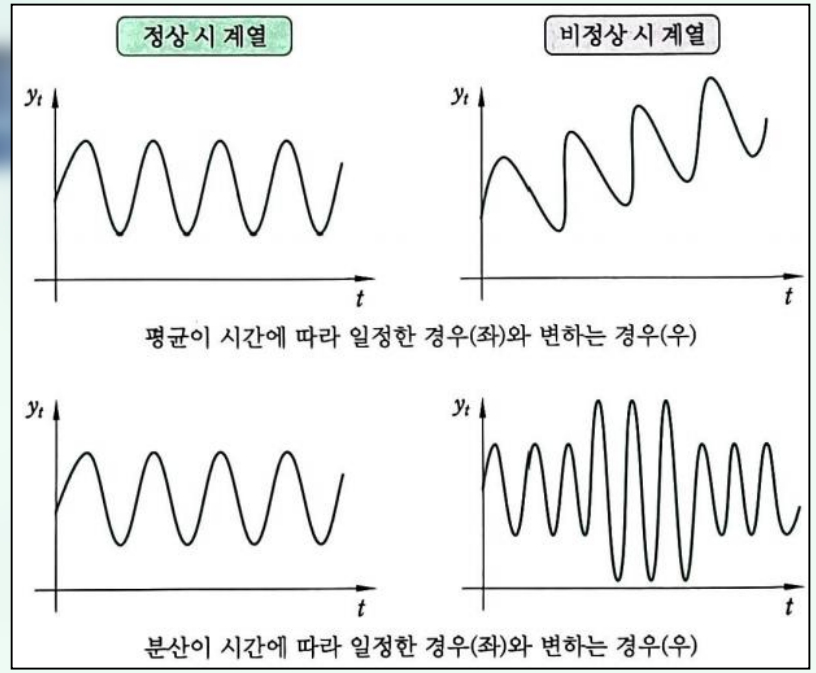
1) 시계열 분석의 정의

- 시계열 분석(Time-series analysis)은 시간의 흐름에 따라 관측된 과거 데이터를 분석하여 미래의 데이터를 예측하는 분석 기법이다.
- 시계열 데이터의 x축은 시간, y축은 관측값을 나타낸다.

3. 분석기법 적용 – 고급 분석기법

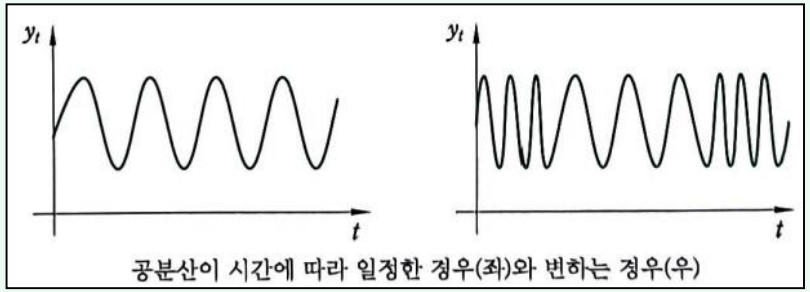
2) 시계열 데이터 정상성(Stationary)과 비정상성(Non-stationary)

- 시계열 데이터는 정상성을 만족해야 한다.
- 정상성은 시점에 상관없이 시계열 특성이 일정한 것을 의미하고, 비정상성은 시점에 따라 시계열 특성이 변하는 것을 의미한다.



평균이 시간에 따라 일정한 경우(좌)와 변하는 경우(우)

분산이 시간에 따라 일정한 경우(좌)와 변하는 경우(우)



공분산이 시간에 따라 일정한 경우(좌)와 변하는 경우(우)

- 정상성의 조건은 평균이 일정하고, 분산이 시점에 의존하지 않으며, 공분산은 시차에만 의존하고 시점에는 의존하지 않는다는 것이다.

3. 분석기법 적용 – 고급 분석기법

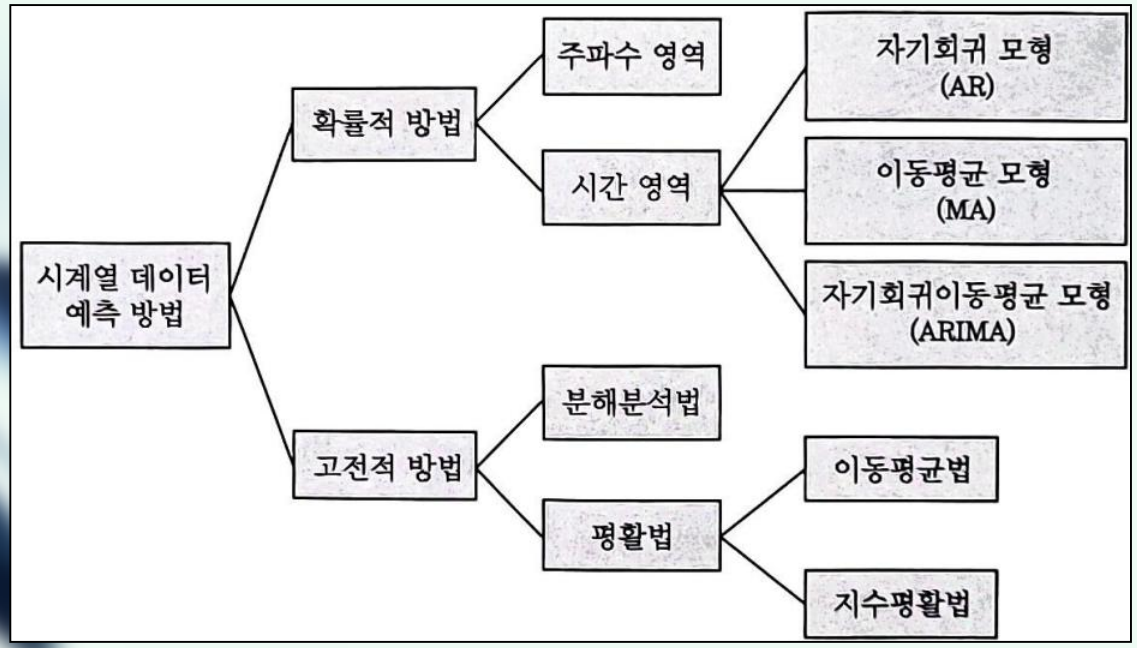
3) 시계열 데이터 예측 방법

- 시계열 데이터의 예측 방법은 확률적 방법과 고전적 방법으로 나뉜다.
- 확률적 방법은 주파수 영역과 시간 영역으로 나뉘고, 시간 영역에는 자기회귀모형, 이동평균 모형, 자기회귀이동평균 모형이 있다.
- 고전적 방법은 분해분석법과 평활법으로 나뉘고, 평활법에는 이동평균법(Moving Average)과 지수평활법(Exponential Smoothing)이 있다.
- 이동평균법은 일정 기간의 관측치를 이용하여 평균을 구하고, 이를 이용해 예측하는 방법으로 장기적인 추세를 쉽게 파악할 수 있다.
- 지수평활법은 일정기간의 평균을 활용하는 이동평균법과 다르게 모든 시계열 데이터를 사용하여 평균을 구하고, 시간의 흐름에 따라 최근 시계열 데이터에 더 많은 가중치를 부여하여 미래를 예측하는 방법이다.

자기회귀 모형(AR 모형, Auto Regressive Model) : 과거의 데이터가 현재의 데이터와 선형적으로 의존하여 영향을 미치는 모형을 의미한다.
이동평균 모형(MA 모형, Moving Average Model) : 시간이 지날수록 관측치의 평균값이 지속적으로 증가하거나 감소하는 모형이다.
자기회귀 이동평균 모형(ARMA 모형, Auto Regressive Moving Average) : 자기 회귀 모형과 이동평균 모형을 합친 모형이다.

3. 분석기법 적용 – 고급 분석기법

3) 시계열 데이터 예측 방법



3. 분석기법 적용 – 고급 분석기법

4) 시계열 데이터 공분산 기법

- 시계열 데이터의 공분산 기법으로는 자기상관(autocorrelation)이 있다.
- 상관계수가 두 변수 사이의 선형 관계의 크기를 측정하는 것과 같이 자기상관은 시계열 데이터의 시차값(logged values) 사이의 선형 관계를 측정한다.
- 자기상관계수(Autocorrelation coefficients)는 동일한 변수(Yt, $Yt - 1$, $Yt - 2$, ...)의 서로 다른 시간 차이(time lag)를 두고 관계를 분석하는 것이다.
- 자기상관함수(ACF : Auto Correlation Function)는 임의의 어떤 신호($p(t)$)와 그 신호를 임의의 시간(t)만큼 지연시킨 신호($p(t+t)$) 사이의 상관관계를 파악할 수 있는 함수이다.
- 데이터에 추세(Trend)가 존재할 때 자기상관함수는 양의 값을 갖는 경향을 보이고, 이러한 자기상관함수 값은 시차가 증가함에 따라 서서히 감소한다.

공분산 : 두 개의 확률 변수의 선형관계를 나타내는 값이다. 한 확률 변수의 증감에 따른 다른 확률 변수의 증감의 경향에 대한 측도이다. 쉽게 말해 분산이라는 개념을 확장하여 두 개의 확률 변수의 흩어진 정도를 공분산이라고 하는 것이다.

3. 분석기법 적용 – 고급 분석기법

5) 시계열 모형

- 자기회귀 모형(AR 모형, Auto Regressive Model) : 과거의 데이터가 현재의 데이터와 선형적으로 의존하여 영향을 미치는 모형을 의미한다.
- 이동평균 모형(MA 모형, Moving Average Model) : 시간이 지날수록 관측치의 평균값이 지속적으로 증가하거나 감소하는 모형이다.
- 자기회귀 이동평균 모형(ARMA 모형, Auto Regressive Moving Average) : 자기 회귀 모형과 이동평균 모형을 합친 모형이다.
- 자기회귀 누적 이동평균 모형(ARIMA 모형) : 자기회귀와 이동평균을 모두 고려하는 모델로 시계열의 비정상성(Non-stationary)을 설명하기 위해 관측치 간의 차분(Differencing)을 사용하는 모형이다.

ARIMA(p, d, q)

p : AR 관련, d : 몇 번 차분했는지, q : MA 관련

- ARIMA(0,0,0) : 백색잡음 모형
- ARIMA(0,1,0) : 확률보행 모형
- ARIMA(p,0,0) : 자기회귀 모형
- ARIMA(0,0,q) : 이동평균 모형

ARIMA 차수에 따른 모형 유형

차분 : 시계열의 수준에서 나타내는 변화를 제거하여 시계열의 평균 변화를 일정하게 만들어 주는 작업
백색잡음 모형 : 시점에 상관없이 평균이 0이고, 분산이 σ^2인 시계열 자료를 말하며, 정상 시계열의 대표적인 예이다.
확률보행 모형 : 데이터가 정상성을 나타내지 않는 모델로 주로 금융이나 경제분야에서 데이터 분석에 사용된다.

3. 분석기법 적용 – 고급 분석기법

5) 시계열 모형

- 시계열 분해 : 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법이다.
- 시계열 분해 구성 요소에는 추세, 계절성, 순환, 불규칙 요인이 있다.

구성 요소	설명
추세 (Trend)	데이터가 장기적으로 증가하거나 감소하는 것으로 추세가 꼭 선형일 필요은 없다.
계절성 (Seasonal)	주, 월, 분기, 반기 단위 등 특정 시간의 주기로 나타나는 패턴
순환 (Cycle)	경기변동과 같이 정치, 경제, 사회적 요인에 의한 변화로 일정 주기가 없는 장기적인 변화 현상
불규칙요인 (Irregular Factor)	설명될 수 없는 요인 또는 돌발적인 요인에 의해 일어나는 변화로 예측 불가능한 임의의 변동

- 시계열 분해 그래프의 관측치를 통해 추세(Trend), 계절성(Seasonal), 잔차(Residual)를 확인할 수 있다.

잔차(residual) : 표본(Sample)으로 추정한 회귀식과 실제 관측값의 차이로 각각의 자료가 직선에 얼마나 잘 맞는지 확인하는 도구

3. 분석기법 적용 – 고급 분석기법

개념 체크

01 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 : MA 관련

● 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 σ^2인 시계열 자료를 의미하며, 정상 시계열의 대표적인 예이다.

확률보행 모형 : 데이터가 정상성을 나타내지 않는 모델로,

02 다음 중 시계열 모형이 아닌 것은?

① 백색잡음 ② 이항분포

③ 자기회귀 ④ 이동평균

● 자기회귀 모형(AR 모형, Auto Regressive Model) : 과거의 데이터가 현재의 데이터와 선형적으로 의존하여 영향을 미치는 모형을 의미한다.

● 이동평균 모형(MA 모형, Moving Average Model) : 시간이 지날수록 관측치의 평균값이 지속적으로 증가하거나 감소하는 모형이다.

● 자기회귀 이동평균 모형(ARMA 모형) : 자기회귀 모형과 이동평균 모형을 합친 모형이다.

● 자기회귀 누적 이동평균 모형(ARIMA 모형) : 자기회귀와 이동평균을 모두 고려하는 모델로 시계열의 비정상성을 설명하기 위해 관측치 간의 차분을 사용하는 모형이다.

3. 분석기법 적용 – 고급 분석기법

03 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 요인**이 있다.

추세(Trend) : 데이터가 장기적으로 증가하거나 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

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

① 차수가 p, d, q인 모델은 ARIMA(p, d, q)로 나타낸다.

② 비정상 시계열을 안정적으로 정상화하기 위해 차수 d는 되도록 크게 설정해야 한다.

③ 차수 d가 0인 경우 ARMA(p, q) 모델을 사용한 것과 동일하다.

④ AR은 자기회귀 모델을 나타내고, MA는 이동평균 모델을 나타낸다.

비정상 시계열을 설명하기 위해서는 단순히 차분의 횟수(차수)를 높이는 것이 아니라 적절한 수치로 설정해야 한다.

3. 분석기법 적용 – 고급 분석기법

05 다음 설명하는 시계열에 대한 명칭은?

주, 월, 분기, 반기 단위 등 특정 시간의 주기로 나타나는 패턴

① 추세 ② 계절

③ 주기 ④ 불규칙

06 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 σ^2 인 시계열 자료를 말하며, 정상 시계열의 대표적인 예이다. 비정상 시계열의 대표적인 예로는 확률 보행(Random Walk)이 있다. 확률 보행은 임의의 방향으로 연속적인 걸음이 나타난다는 의미로 예측이 불가능한 변동이 발생하는 것을 의미한다. 시계열 데이터의 정상성은 시점에 상관없이 시계열 특성이 일정한 것을 의미한다. 정상성의 조건은 평균이 일정하고, 분산이 시점에 의존하지 않으며, 공분산은 시차에만 의존하고 시점에는 의존하지 않는다는 것이다.

3. 분석기법 적용 – 고급 분석기법

07 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)은 대표적인 정상 시계열로서 독립적이고 동일한 분산을 갖는다.

08 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.

시계열 데이터 예측 방법

● 시계열 데이터의 예측 방법은 확률적 방법과 고전적 방법으로 나뉜다.

● 확률적 방법은 주파수 영역과 시간 영역으로 나뉘고, 시간 영역에는 자귀회귀 모형, 이동평균 모형, 자기회귀이동평균 모형이 있다.

● 고전적 방법은 분해분석법과 평활법으로 나뉘고, 평활법은

3. 분석기법 적용 – 고급 분석기법

04 베이지안 기법

1) 조건부 확률

- 조건부 확률은 어떤 사건이 일어난다는 조건에서 다른 사건이 일어날 확률을 의미한다.

사건 A가 조건으로 일어날 때 사건 B가 발생할 확률

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

사건 B가 조건으로 일어날 때 사건 A가 발생할 확률

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

3. 분석기법 적용 – 고급 분석기법

2) 베이즈 정리(Bayes' Theorem)

- 베이즈 정리는 두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 설명하는 확률이론으로 B가 발생할 때, A가 발생할 확률을 의미한다.
- 어떤 사건이 서로 배반(排反)하는 원인이 둘에 의해 일어난다고 할 때 실제 사건이 일어났을 때 이것이 두 원인 중 하나일 확률을 구하는 정리이다.

베이즈 정리 $$P(A_i|B)=\frac{P(A_i \cap B)}{P(B)}=\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

배반 : 두 개의 사건이 동시에 일어날 수 없는 경우

3. 분석기법 적용 – 고급 분석기법

3) 나이브 베이즈 분류(Naive Bayes Classification)

- 나이브 베이즈 분류는 베이즈 정리에 기반한 통계적 분류 방법을 의미하며, 가장 단순한 지도학습(supervised learning) 방법 중 하나이다.
- 사건 A가 발생했을 때, 사건 B가 발생할 확률을 이용하여 사건 B가 발생했을 때, 사건 A가 일어날 확률을 추정하는 기법이다.
- 데이터 산출 속도가 빠르기 때문에 실시간 분류 및 텍스트 분석 분야에서 주로 사용된다.

3. 분석기법 적용 – 고급 분석기법

05 딥러닝 분석

1) 딥러닝의 개념

- 딥러닝(Deep Learning)은 대용량 데이터를 처리하기 위해 인공신경망을 기반으로 구현되는 기계학습 알고리즘이다.
- 딥러닝은 기존 인공신경망 모델의 문제점이었던 기울기 소실 문제를 해결하였고, GPU를 활용한 연산으로 데이터 분석 시간을 단축시켰다.

2) 딥러닝 알고리즘

딥러닝 알고리즘에는 DNN, CNN, RNN, GAN 등이 있다.

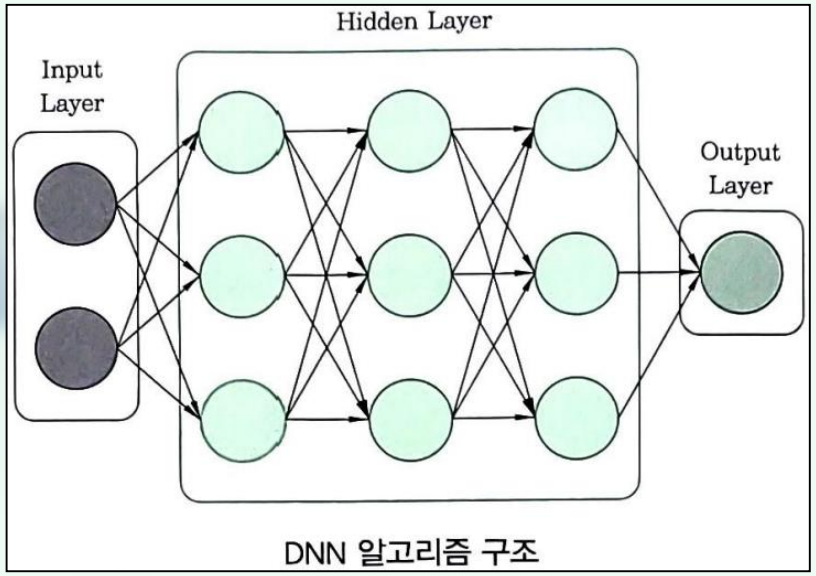
① DNN(Deep Neural Network, 심층신경망) 알고리즘

- ▶ DNN은 은닉층(Hidden Layer)이 하나만 존재하는 ANN(인공신경망)과 다르게 입력층(Input Layer)과 출력층(Output Layer) 사이에 2개 이상의 은닉층이 존재하는 알고리즘을 의미한다.
- ▶ 데이터는 입력층에서 가중치가 곱해져 은닉층으로 이동하고, 은닉층에서 역시 가중치가 곱해져 다음 계층으로 이동한다.
- ▶ 역전파 알고리즘을 통해 출력층에서 은닉층으로, 다시 입력층으로 역순으로 연산을 수행하며 최적의 결과를 도출하게 된다.

3. 분석기법 적용 – 고급 분석기법

① DNN(Deep Neural Network, 심층신경망) 알고리즘

▶ 심층신경망의 가중치(weight) 파라미터는 경사하강법을 통하여 갱신될 수 있다.



DNN 알고리즘 구조

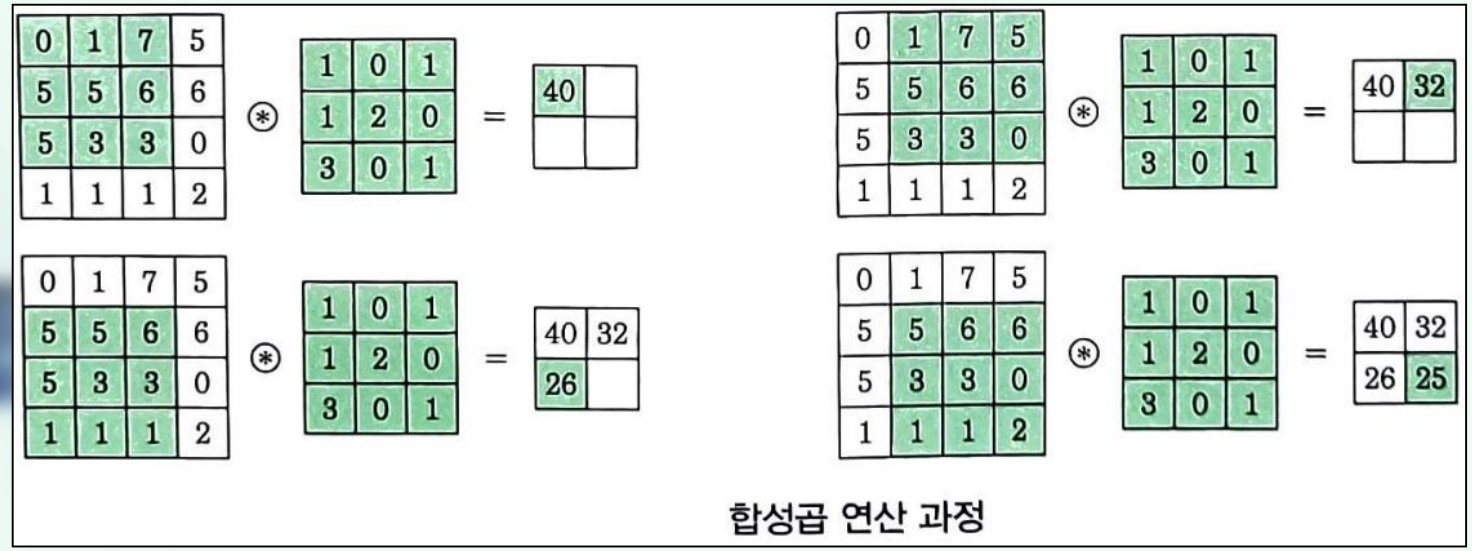
② CNN(Convolution Neural Network, 합성곱 신경망) 알고리즘

▶ CNN은 합성곱(Convolution)과 풀링(Pooling) 과정을 거쳐 데이터를 분석하는 알고리즘으로 주로 시각적 이미지 분석에서 많이 사용된다.

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

3. 분석기법 적용 – 고급 분석기법

② CNN(Convolution Neural Network, 합성곱 신경망) 알고리즘



합성곱 연산 과정

3. 분석기법 적용 – 고급 분석기법

② CNN(Convolution Neural Network, 합성곱 신경망) 알고리즘

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 크기는 다음과 같다.

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

예시) CNN에서 원본 이미지가 5×5에서 스트라이드가 1이고, 필터가 3×3일 때, 피처맵은?

$n = 5,\ p = 0$(사용되지 않았으므로), $s = 1,\ f = 3$

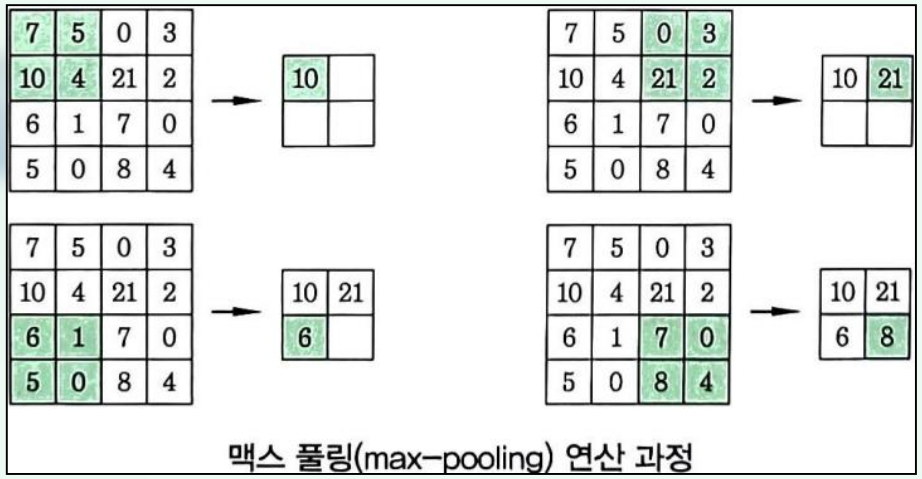
$$= \left(\frac{n+sp-f}{s}+1\right) * \left(\frac{n+sp-f}{s}+1\right)$$

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

3. 분석기법 적용 – 고급 분석기법

② CNN(Convolution Neural Network, 합성곱 신경망) 알고리즘

▶ 풀링(Pooling) : 풀링은 합성곱 과정을 거친 데이터를 요약하는 작업으로 추출한 특징은 유지하면서 데이터의 사이즈를 줄여주는 과정이다. 풀링은 최댓값을 선택하는 max pooling과 평균값을 선택하는 average pooling 방법이 있다.

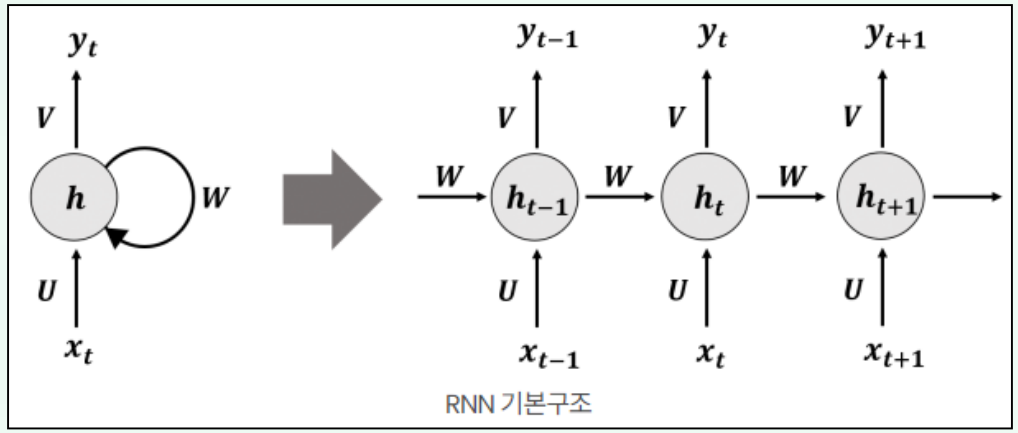


맥스 풀링(max-pooling) 연산 과정

3. 분석기법 적용 – 고급 분석기법

③ RNN(Recurrent Neural Network, 순환신경망) 알고리즘

- ▶ RNN은 언어 데이터, 시계열 데이터 등과 같이 연속적인 데이터 분석에 특화된 알고리즘으로 과거 데이터를 기반으로 현재 데이터를 학습하는 특징이 있다.
- ▶ RNN 알고리즘은 장기 의존성 문제와 기울기 소실 문제가 발생할 수 있기 때문에 이를 보완한 LSTM(Long Short Term Memory, 장단기 메모리) 기법이 개발되었다.
- ▶ LSTM은 망각 게이트, 입력 게이트, 업데이트 게이트, 출력 게이트로 구성되어 불필요한 데이터를 제거하고, 필요한 데이터만 업데이트 하여 출력값으로 활용한다.



RNN 기본구조

LSTM : RNN의 변형 구조로써 게이트 메커니즘을 추가한 모델이라고 할 수 있다.
망각 게이트(Forget Gate) : 불필요한 정보를 삭제하기 위한 게이트이다.
입력 게이트(Input Gate) : 입력으로 들어온 입력 벡터(입력 토큰)에 대한 정보를 기억하는 역할을 한다.
출력 게이트(Output Gate) : 현재 시점의 은닉 벡터를 결정하는 게이트이다.

3. 분석기법 적용 – 고급 분석기법

③ RNN(Recurrent Neural Network, 순환신경망) 알고리즘

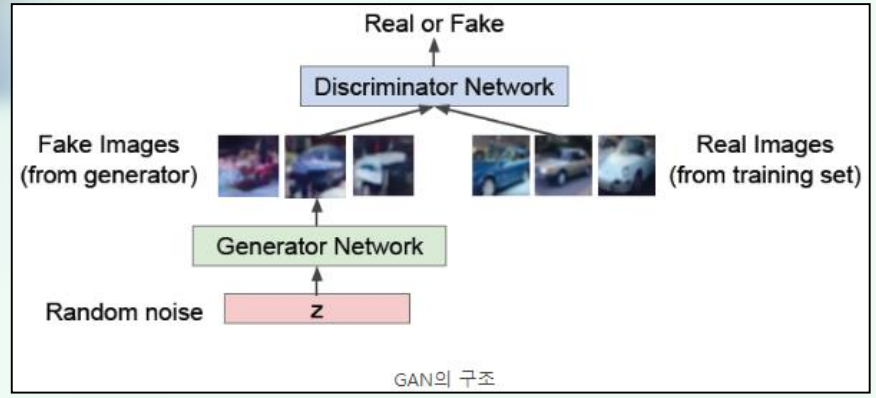
- ▶ LSTM과 유사한 성능을 갖지만 복잡한 구조를 단순화시킨 방법이 게이트 순환 유닛(GRU: Gated Recurrent Unit)이다.
- ▶ GRU는 LSTM의 장기 의존성 문제에 대한 해결책은 유지하면서 은닉 상태를 업데이트하는 계산량을 줄였다.
- ▶ GRU는 업데이트 게이트와 리셋 게이트만 존재한다.

3. 분석기법 적용 – 고급 분석기법

④ GAN(Generative Adversarial Network, 생성적 적대 신경망) 알고리즘

▶ GAN은 진짜와 같은 가짜를 만들어내는 생성자(Generator)와 만들어진 데이터의 진위 여부를 확인하는 판별자(Discriminator)가 대립하여 성능을 개선해 나가는 알고리즘이다.

▶ 진짜와 같은 가짜를 만들어내는 것이 GAN 알고리즘의 목표라고 할 수 있으며, 이는 딥페이크(Deep Fake) 기술로 활용된다.



GAN의 구조

▶ 이렇게 서로 경쟁하면서 학습을 함으로써, generator는 점점 더 실제와 같은 데이터를 생성하게 되고, discriminator는 점점 더 실제와 가짜 데이터를 잘 구별할 수 있게 될 것이다.

딥 페이크 : 인공 지능을 기반으로 한 인간 이미지 합성 기술

3. 분석기법 적용 – 고급 분석기법

개념 체크

01 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.

02 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 문제가 발생할 수 있기 때문에 이를 보완한 LSTM(장단기 메모리)기법이 개발되었다.

① GAN ② DNN

③ RNN ④ CNN

순환 신경망 알고리즘(RNN)에 대한 설명이다.

3. 분석기법 적용 – 고급 분석기법

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

$n = 5, p = 0$(사용되지 않았다), $s = 1, f = 3$

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

04 다음의 각 과제에 대한 분석 방법이 적절하게 연결된 것은?

가. 영화 감상평에 대한 긍정/부정 판단

나. 사원증 대신 얼굴 인식으로 출입 가능한 보안 게이트 설치

다. 사용자가 업로드한 이미지에 대한 설명을 제공하는 앱 개발

라. 공장 로봇이 돌발 상황에 적절하게 대응할 수 있도록 운동능력 훈련

① 가: 순환신경망, 나: 합성곱신경망, 다: 순환신경망+합성곱신경망, 라: 강화학습

② 가: 순환신경망, 나: 강화학습, 다: 합성곱신경망, 라: 순환신경망+합성곱신경망

③ 가: 합성곱신경망, 나: 순환신경망, 다: 강화학습, 라: 순환신경망+합성곱신경망

④ 가: 합성곱신경망, 나: 순환신경망+합성곱신경망, 다: 순환신경망, 라: 강화학습

순환신경망(RNN)은 언어 데이터, 시계열 데이터 등과 같이 연속적인 데이터 분석에 특화된 알고리즘이다.

3. 분석기법 적용 – 고급 분석기법

05 다음 중 심층신경망에 대한 설명으로 적절하지 않은 것은?

① 심층신경망은 입력층과 출력층 사이에 여러 개의 은닉층들로 이루어진 인공신경망이다.

② 심층신경망은 오류역전파 알고리즘으로 학습될 수 있다.

③ 심층신경망의 가중치(weight) 파라미터는 경사하강법을 통하여 갱신될 수 있다.

④ 합성곱신경망은 합성곱 계층으로 일반적인 인공신경망 계층을 대신한다.

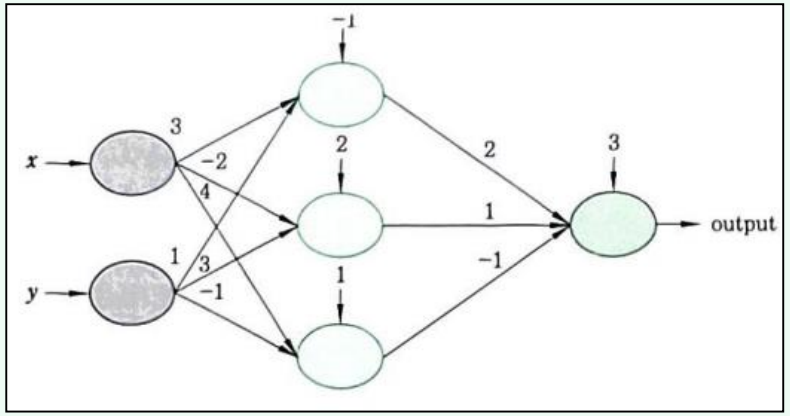
합성곱신경망은 합성곱 계층과 일반적인 인공신경망 계층으로 구성된다.

DNN(Deep Neural Network, 심층 신경망) 알고리즘

● DNN은 은닉층(Hidden Layer)이 하나만 존재하는 ANN(인공신경망)과 다르게 입력층(Input Layer)과 출력층(Output Layer) 사이에 2개 이상의 은닉층이 존재하는 알고리즘을 의미한다.

● 데이터는 입력층에서 가중치가 곱해져 은닉층으로 이동하고, 은닉층에서 역시 가중치가 곱해져 다음 계층으로 이동한다.

06 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. 따라서 다음과 같이 연산을 하면 된다.

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

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

3. 분석기법 적용 – 고급 분석기법

07 다음 중 순환신경망에서 발생하는 기울기 소실(Gradient Vanishing), 기울기 폭발(Gradient Exploding)에 대한 설명으로 적합한 것은?

① RNN은 LSTM(Long Short Term Memory)의 장기의존성 문제 를 보완하기 위한 알고리즘이다.

② 순환신경망은 입력 게이트, 망각 게이트, 출력 게이트로 구성된다.

③ 기울기 클리핑(Gradient Clipping)은 기울기 소실을 막기 위해 기울기 값을 자르는 방법이다.

④ 기울기 소실이란 오차 역전파 과정에서 입력층으로 갈수록 가중치에 따른 결과값의 기울기가 작아져 0에 수렴 하는 문제이다.

기울기 클리핑은 기울기 폭발을 막기 위해 일정 임계값을 넘지 못하도록 기울기 값을 자르는 방법이다. LSTM은 RNN의 장기의존성 문제를 보완하기 위한 알고리즘으로 입력 게이트, 망각 게이트, 업데이트 게이트, 출력 게이트로 구성된다.

3. 분석기법 적용 – 고급 분석기법

06 비정형 데이터 분석

- 비정형 데이터는 이미지, 영상, 문서 데이터와 같이 정형화된 데이터의 구조를 갖지 않는 데이터를 의미 한다.
- 비정형 데이터 분석은 이러한 비정형 데이터를 분석하여 의미 있는 정보를 도출해 내는 분석을 의미한다.
- 비정형 데이터 분석 방법에는 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 사회 연결망 분석이 있다.

1) 텍스트 마이닝(Text Mining)

- 텍스트 마이닝은 텍스트 형태로 이루어진 비정형 데이터를 수집한 뒤, 자연어 처리(NLP)를 통해 유의미한 정보를 추출하는 방법이다.

  [텍스트 마이닝 절차]

  텍스트 수집 → 텍스트 전처리(토큰화, 품사 태깅 등) → 텍스트 의미 추출 → 텍스트 패턴 분석 → 정보 생성

- 텍스트 전처리 기법에는 토큰화(Tokenization), 품사 태깅(POS Tagging), 표제어 추출(Lemmatization), 어간 추출(Stemming), 불용어(Stopword) 처리가 있다.

자연어 처리(NLP: Natural Language Processing) : 사람이 이해할 수 있는 언어를 기계가 이해할 수 있는 언어로 처리 하는 기술
토큰화 : 문서를 토큰(token)이라 불리는 작은 단위로 나누는 기술
품사 태깅 : 형태소(의미를 갖는 가장 작은 말의 단위)의 품사를 태깅하는 기술

3. 분석기법 적용 – 고급 분석기법

1) 텍스트 마이닝(Text Mining)

텍스트 전처리 기법

기법	설명
토큰화	문서를 토큰(token)이라 불리는 작은 단위로 나누는 기술 예) I love to listen K-pop -> I / love / to / listen / K-pop
품사 태깅	형태소(의미를 갖는 가장 작은 말의 단위)의 품사를 태깅하는 기술 예) I / go / to / school -> 인칭 대명사 / 동사 / 전치사 / 명사
표제어 추출	단어들로부터 표제어(단어가 사전에 등재된 형태)를 찾는 기법 예) is gone, is cleaned -> be + p.p
어간 추출	단어에서 접사를 제거하여 어간(용언 사용 시 변하지 않는 부분)을 추출하는 기술 예) 자는, 자고, 자서 -> 자~
불용어 처리	단어에서 조사, 접미사와 같이 의미 분석에 중요도가 낮은 단어를 처리하는 기술

- 텍스트 마이닝 기능에는 정보 추출, 문서 요약, 문서 분류, 문서 군집화가 있다.
- 텍스트 마이닝 기법 중 자연어를 컴퓨터가 이해할 수 있도록 벡터로 만들어주는 것을 벡터화(Vectorize)라고 한다.
- 벡터화 방법에는 Bag of Words, TF-IDF, One-hot encoding, Word Embedding이 있다.

3. 분석기법 적용 – 고급 분석기법

1) 텍스트 마이닝(Text Mining)

① Bag of Words

▶ 가장 단순한 벡터화 방법 중 하나로 문서에서 문법이나 단어의 순서를 무시하고 단순히 단어의 빈도만 고려한 벡터화 방법이다.

② 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 가진다.

$$IDF(\omega) = log\left(\frac{n}{1+df(\omega)}\right)$$

(n : 분류대상이 되는 모든 문서의 수, $df(\omega)$: 단어 ω가 들어있는 문서의 수)

③ One-hot encoding

▶ 표현하고 싶은 데이터를 1값으로, 그렇지 않은 데이터를 0값으로 표현하는 방식이다.

3. 분석기법 적용 – 고급 분석기법

1) 텍스트 마이닝(Text Mining)

④ 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.

예) LSA, Word2Vec, FastText, GloVe

방법	설명
LSA	LSA(Latent Semantic Analysis)는 잠재 의미 분석으로 문서 및 용어와 관련된 일련의 개념을 생성하여 문서 집합과 포함된 용어 사이의 관계를 분석하는 자연어 처리 기술
Word2Vec	워드 임베딩에 기반하여 각 단어 간의 유사도를 벡터화하여 해당 단어의 의미를 수치화할 수 있는 방식
FastText	Facebook의 AI Research lab에서 만든 단어 임베딩 및 텍스트 분류 학습을 위한 라이브러리
GloVe	글로브(Global Vectors for Word Representation)는 카운트 기반과 예측 기반을 모두 사용하는 방법론으로 2014년에 미국 스탠포드 대학에서 개발한 단어 임베딩 방법론

분포가설 : 단어의 의미는 주변 단어에 의해 형성된다.

감사합니다.