시계열자료분석팀

5팀

김 민

이수린

김동환

서유진

장다연

INDEX

1

2주차 복습

정상 시계열

정상 시계열

추세 또는 계절성이 없고 약정상성을 만족하지만

오차항은 백색잡음이 아닌 시계열 자료

자기상관 존재

정상 시계열

정상 시계열

추세 또는 계절성이 없고 약정상성을 만족하지만

오차항은 백색잡음이 아닌 시계열 자료

자기상관 존재

정상 시계열 가정하는 모형

AR

MA

ARMA

AR 모형

AR 모형

과거 시점의 관측값들과 현재 시점의 오차항으로 설명되는 모델

AR(1) : $X_t = \phi_1 X_{t-1} + Z_t$

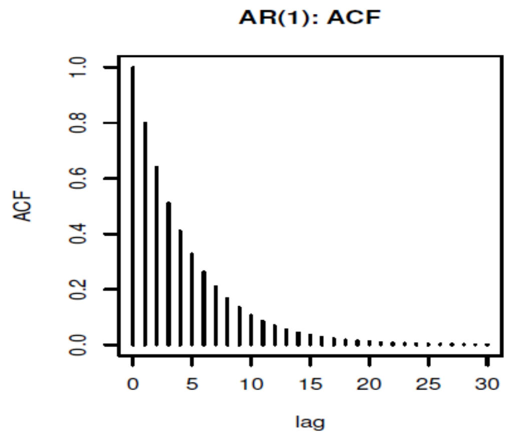
AR 모형

AR 모형

과거 시점의 관측값들과 현재 시점의 오차항으로 설명되는 모델

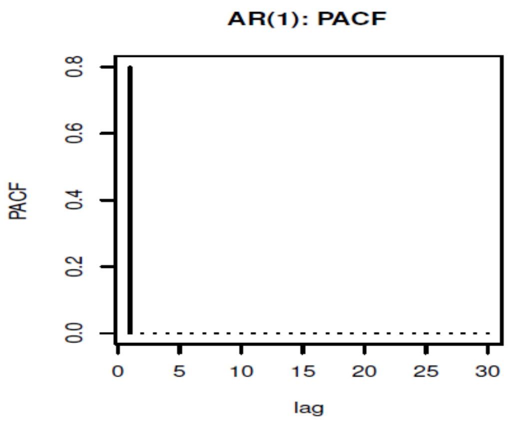
$$\text{AR(1)} : X_t = \phi_1 X_{t-1} + Z_t$$

AR(1) ACF



지수적으로 감소

AR(1) PACF



시차 이후로 절단

MA 모형

MA 모형

과거 시점의 오차항으로 설명되는 모델

MA(1) : $\boldsymbol{X_t = Z_t + \theta_1 Z_{t-1}}$

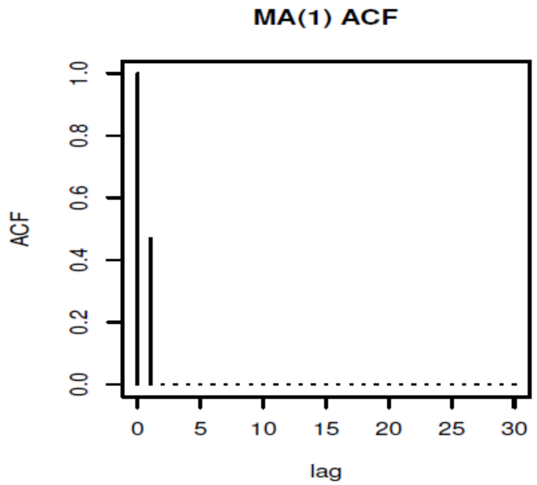
MA 모형

MA 모형

과거 시점의 오차항으로 설명되는 모델

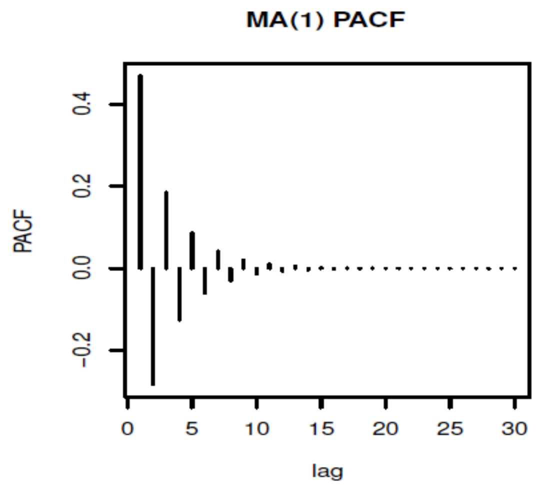
MA(1) :$\boldsymbol{X_t = Z_t + \theta_1 Z_{t-1}}$

MA(1) ACF



시차 이후로 절단

MA(1) PACF



지수적으로 감소

ARMA 모형

ARMA 모형

AR과 MA를 동시에 포함하는 모델

ARMA(1, 1) : $\boldsymbol{X_t - X_{t-1} = Z_t + Z_{t-1}}$

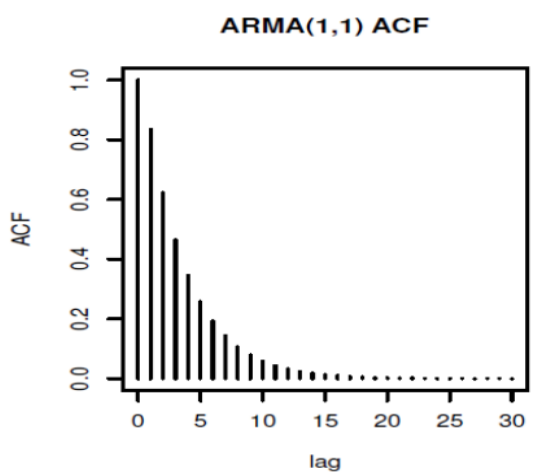
ARMA 모형

ARMA 모형

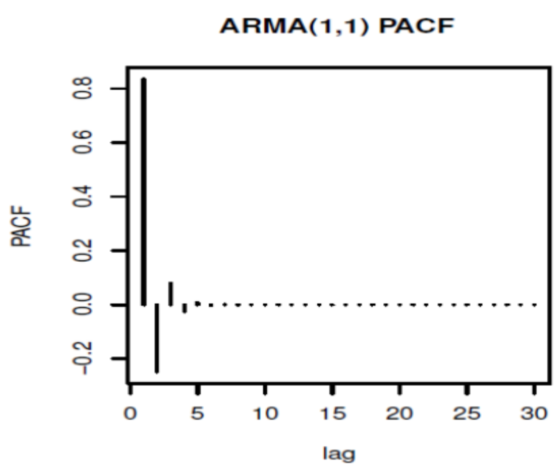
AR과 MA를 동시에 포함하는 모델

ARMA(1, 1) : $\boldsymbol{X_t - X_{t-1} = Z_t + Z_{t-1}}$

ARMA(1,1) ACF



ARMA(1,1) PACF



두 그래프 모두 지수적으로 감소

ARMA 모형

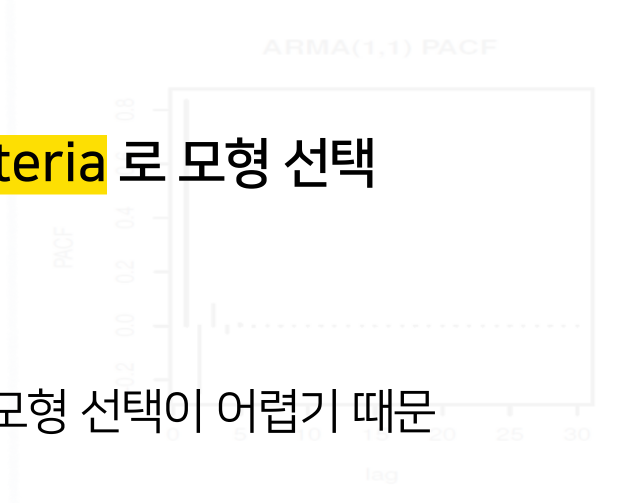
ARMA 모형

AR과 MA를 동시에 포함하는 모델

ARMA(1, 1) : $X_t - X_{t-1} = Z_t + Z_{t-1}$

ARMA(1,1) ACF

ARMA(1,1) PACF



Information criteria 로 모형 선택

AIC, AICC, BIC

ACF, PACF 그래프만으로 모형 선택이 어렵기 때문

두 그래프 모두 지수적으로 감소

1 2주차 복습

AR, MA, ARMA 모형의 조건

정상 시계열 가정하는 모형

AR　　MA　　ARMA

정상성, 가역성을 위한 조건

계수 절댓값이 1보다 작아야 함

= 특성방정식 근의 절댓값이 1보다 커야 함

1 2주차 복습

시계열 모형의 적합

모형 식별 — p,q 차수 결정

모수 추정 — MLE 사용

모형 진단 — 모수, 잔차 검정

예측 — 미래 시점은 현재까지 관측값들의 선형결합임을 가정
참값과 예측값 사이의 오차를 최소화하는 MSPE 최소화

1 2주차 복습

시계열 모형의 적합

대부분의 시계열 데이터는 비정상 시계열

(약정상성 만족 x, 추세나 계절성 존재, 공분산 시점에 의존)

정상화 과정을 거쳐야 함

모형 식별 → p,q 차수 결정

모수 추정 → MLE 사용

모형 진단 → 모수, 잔차 검정

예측 → 미래 시점은 현재까지 관측값들의 선형결합임을 가정

참값과 예측값 사이의 오차를 최소화하는 MSPE 최소화

시계열 모형의 적합

모형 식별 → p,q 차수 결정

모수 추정 →

모형 진단 →

예측 → 미래 시점은 현재까지 관측값들의 선형결합임을 가정

참값과 예측값 사이의 오차를 최소화하는 MSPE 최소화

대부분의 시계열 데이터는 비정상 시계열

(약정상성 만족 x, 추세나 계절성 존재, 공분산 시점에 의존)

정상화 과정을 거쳐야 함

비정상 시계열 모형

사전 정상화 과정 없이, 모델 자체에서 정상화가 이루어짐

2

ARIMA

ARIMA 모형의 정의

ARIMA 모형

추세(polynomial trend) 가 있어 정상성을 만족하지 않는
시계열 자료에 적용 가능한 모형

ARMA에 차분이 결합된 형태

ARIMA 모형의 정의

ARIMA 모형

추세(polynomial trend) 가 있어 정상성을 만족하지 않는
시계열 자료에 적용 가능한 모형

ARMA에 차분이 결합된 형태

구체적으로!

$$\phi(B)X_t = \theta(B)Z_t \quad + \quad \text{d차 차분} \quad = \quad \phi(B)(1-B)^d X_t = \theta(B)Z_t$$

ARMA 모형의 특성 방정식

ARIMA 모형의 특성 방정식

ARIMA 모형의 정의

ARIMA 모형

추세*(polynomial trend)*가 있어 정상성을 만족하지 않는
시계열 자료에 적용 가능한 모형

$$\phi(B)X_t = \theta(B)Z_t \quad + \quad \text{d차 차분} \quad = \quad \phi(B)(1-B)^d X_t = \theta(B)Z_t$$

ARMA 모형의 특성 방정식 ARIMA 모형의 특성 방정식

ARMA에 차분이 결합된 형태

d차 차분한 시계열이 정상 과정 ARMA(p,q)를 따를 때,
해당 시계열에 ARIMA(p,d,q)를 적합할 수 있다.

구체적으로!

$$\phi(B)X_t = \theta(B)Z_t \quad + \quad \text{d차 차분} \quad = \quad \phi(B)(1-B)^d X_t = \theta(B)Z_t$$

ARMA 모형의 특성 방정식 ARIMA 모형의 특성 방정식

ARIMA 모형의 정의

ARIMA 모형

추세*(polynomial trend)*가 있어 정상성을 만족하지 않는 시계열 자료에 적용 가능한 모형

$$\phi(B)X_t = \theta(B)Z_t \quad + \quad \text{d차 차분} \quad = \quad \phi(B)(1-B)^d X_t = \theta(B)Z_t$$

ARMA 모형의 특성 방정식 ARIMA 모형의 특성 방정식

ARMA에 차분이 결합된 형태

d차 차분한 시계열이 정상 과정 ARMA(p,q)를 따를 때,

구체적으로!

해당 시계열에 ARIMA(p,d,q)를 적합할 수 있다.

(AR의 차수, 차분, MA의 차수)

$$\phi(B)X_t = \theta(B)Z_t \quad + \quad \text{d차 차분} \quad = \quad \phi(B)(1-B)^d X_t = \theta(B)Z_t$$

ARMA 모형의 특성 방정식 ARIMA 모형의 특성 방정식

ARIMA 모형의 정의

ARIMA 모형의 특성 방정식

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t$$

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d X_t = \left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

d=0인 경우

$$\phi(B)X_t = \theta(B)Z_t$$

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)X_t = \left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

→ ARMA(p,q)모형과 동일

d=1 이상인 경우 → d차 추세 제거 가능

ARIMA 모형의 정의

ARIMA(p, d, q) 모형을 언제 적용…?

- 주어진 시계열 데이터에 polynomial trend 존재
- 오차가 ARMA(p, q)를 따르는 경우

두 조건 모두 만족시키는 경우에만 적용 가능!

ARIMA 모형의 정의

ARIMA(p, d, q) 모형을 언제 적용…?

- 주어진 시계열 데이터에 polynomial trend 존재
- 오차가 ARMA(p, q)를 따르는 경우

ARIMA 모형의 장점!

1. 사전 정상화 필요 없음

→ 모형 자체에 차분을 통한 추세 제거가 포함되어 있기 때문

2. 선형 과정 만족

2 ARIMA

ARIMA 모형의 적합 절차

Step1) TS plot / 잔차 plot 확인 후 이분산성이 나타나는 경우 등분산성을 만족하도록 변환

Step 2) TS plot과 ACF 그래프를 통해 정상 / 비정상 시계열 여부 판단

ACF의 감소 속도를 통해 판단

정상 시계열

비정상 시계열

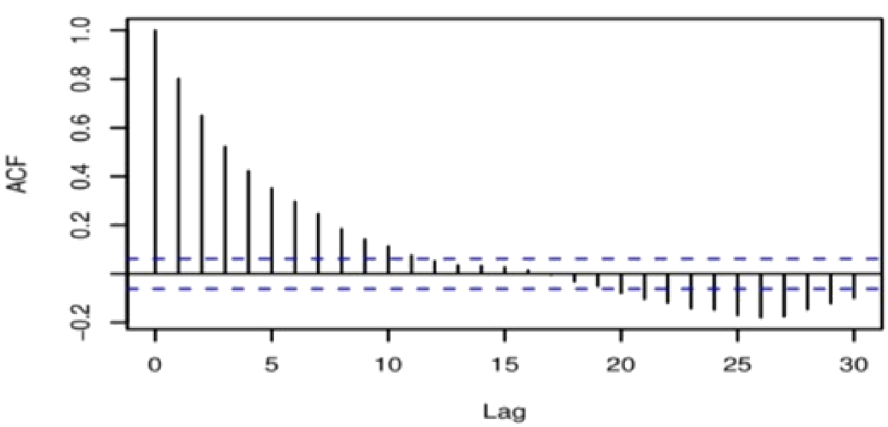
ARIMA 모형의 적합 절차

Step1) TS plot / 잔차 plot 확인 후 이분산성이 나타나는 경우 등분산성을 만족하도록 변환

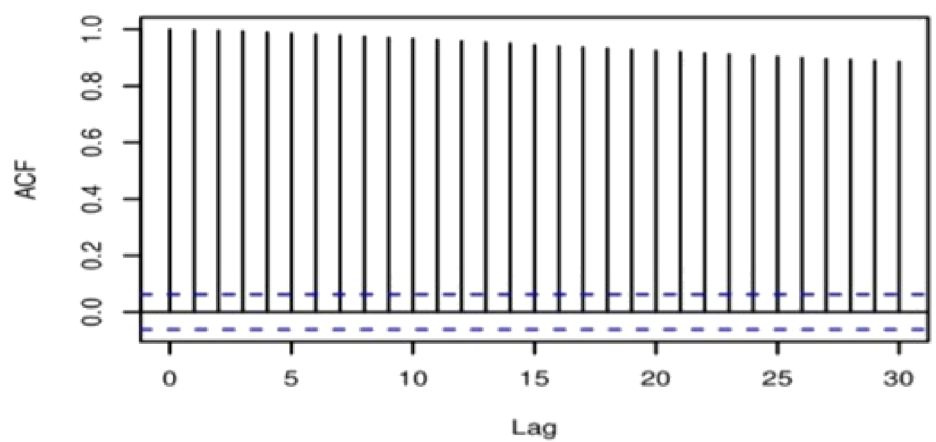
Step 2) TS plot과 ACF 그래프를 통해 정상 / 비정상 시계열 여부 판단

ACF의 감소 속도를 통해 판단

정상 시계열



비정상 시계열



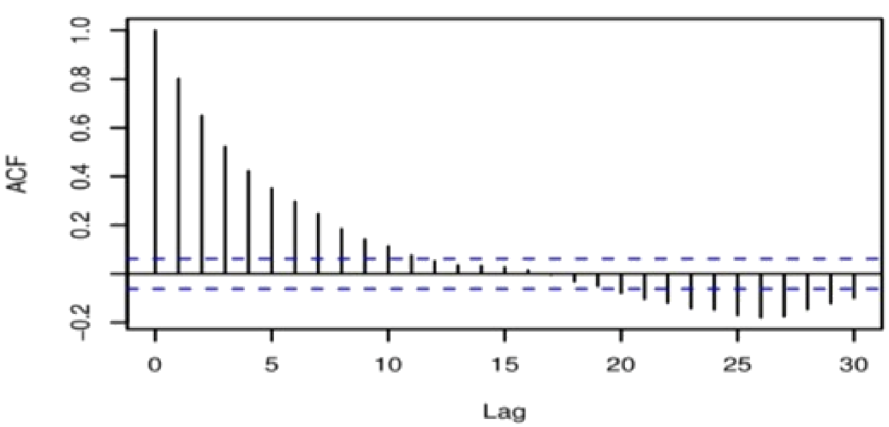
ARIMA 모형의 적합 절차

Step1) TS plot / 잔차 plot 확인 후 이분산성이 나타나는 경우 등분산성을 만족하도록 변환

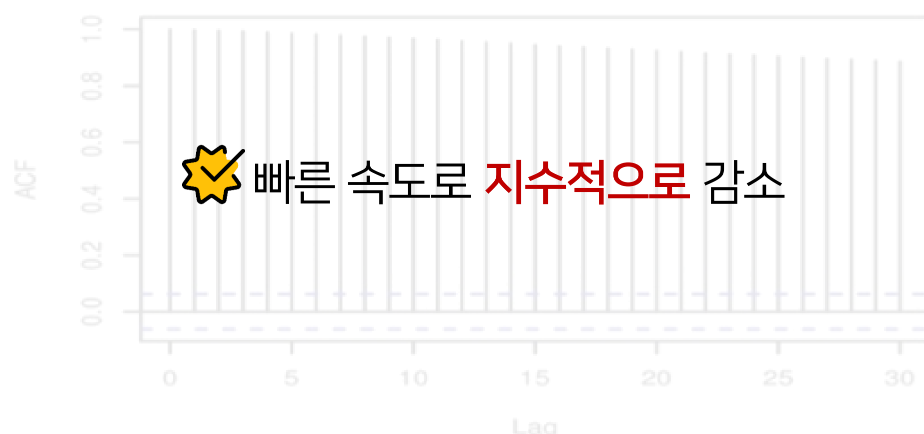
Step 2) TS plot과 ACF 그래프를 통해 정상 / 비정상 시계열 여부 판단

ACF의 감소 속도를 통해 판단

정상 시계열



비정상 시계열



2 ARIMA

ARIMA 모형의 적합 절차

Step1) TS plot / 잔차 plot 확인 후 이분산성이 나타나는 경우 등분산성을 만족하도록 변환

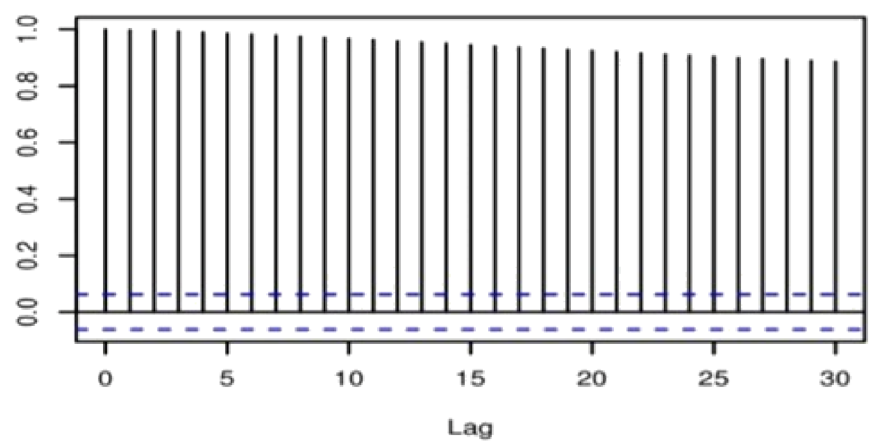
Step 2) TS plot과 ACF 그래프를 통해 정상 / 비정상 시계열 여부 판단

ACF의 감소 속도를 통해 판단

정상 시계열

비정상 시계열

느린 속도로 선형적으로 감소



ARIMA 모형의 적합 절차

Step 3) TS plot에서 추세가 관측되는 경우! → 추세에 맞춰 d차 차분 적용

ARIMA(p, d, q)의 모수 d의 차수 결정

과대차분의 위험이 있어 1차 or 2차 차분을 대부분 적용

과대차분이란 ?

차분을 필요 이상으로 진행하는 것

→ 이미 정상화가 되었음에서 불구하고 차분 적용

복잡한 ACF, 분산이 커지는 문제 발생

ARIMA 모형의 적합 절차

Step 3) TS plot에서 추세가 관측되는 경우! → 추세에 맞춰 d차 차분 적용

ARIMA(p, d, q)의 모수 d의 차수 결정

과대차분의 위험이 있어 1차 or 2차 차분을 대부분 적용

과대차분이란

차분을 필요 이상으로 진행하는 것

→ 이미 정상화가 되었음에도 불구하고 차분 적용

복잡한 ACF, 분산이 커지는 문제 발생

ARIMA 모형의 적합 절차

과대차분의 위험이 있어 1차 or 2차 차분을 대부분 적용

차분을 필요 이상으로 진행하는 것

→ 이미 정상화가 되었음에도 불구하고 차분 적용

정리

비정상 시계열에

1. **1차, 2차 차분** 적용
2. **ACF, PACF** 그래프 확인

지수적으로 감소하는 그래프가 나타나면

적절한 적합이라고 판단 가능

2 ARIMA

ARIMA 모형의 적합 절차

다 했나요…?

체크!

- [ ] p,q의 차수 결정
- [ ] 모수 추정
- [ ] 모형 진단

Step 4) 최종 모형으로 예측 진행

ARIMA 모형의 적합 절차

정상 시계열 모형

ex) 회귀, 평활, 차분

자세한 내용은 시계열팀 클린업 2주차 참고…

비정상 시계열에 대한 정상화 적용 후

정상 시계열 모형 적용

입력값 : 정상화를 거친 오차항

결과값 : 오차항에 대한 예측값

$$X_t = m_t + s_t + Y_t$$

추세 및 계절성을 추정하여 추가로 더해주어야 함

ARIMA 모형의 적합 절차

정상 시계열 모형

ex) 회귀, 평활, 차분

자세한 내용은 시계열팀 클린업 2주차 참고…

비정상 시계열에 대한 정상화 적용 후

정상 시계열 모형 적용

입력값 : 정상화를 거친 오차항

결과값 : 오차항에 대한 예측값

$$X_t = m_t + s_t + Y_t$$

추세 및 계절성을 추정하여 추가로 더해주어야 함

ARIMA 모형의 적합 절차

정상 시계열 모형

비정상 시계열에 대한 정상화 적용 후

정상 시계열 모형 적용

입력값 : 정상화를 거친 오차항

결과값 : 오차항에 대한 예측값

추세 및 계절성을 추정하여 추가로 더해주어야 함

비정상 시계열 모형

연산 과정에서 차분을 통한 정상화 포함

입력값 : 전처리를 거치지 않은 원본 시계열

결과값 : 원본 시계열에 대한 예측값

추세 및 계절성을 더해주는 과정 X

ARIMA 모형의 적합 절차

정상 시계열 모형

비정상 시계열에 대한 정상화 적용 후

정상 시계열 모형 적용

입력값 : 정상화를 거친 오차항

결과값 : 오차항에 대한 예측값

추세 및 계절성을 추정하여 추가로 더해주어야 함

연산 과정에 정상화가

포함되기 때문!

비정상 시계열 모형

연산 과정에서 차분을 통한 정상화 포함

입력값 : 전처리를 거치지 않은 원본 시계열

결과값 : 원본 시계열에 대한 예측값

추세 및 계절성을 더해주는 과정 X

3

SARIMA

SARIMA 모형의 정의

SARIMA 모형

추세와 계절성이 있어 정상성을 만족하지 않는
비정상 시계열 자료에 적용 가능한 모형

ARIMA
VS
SARIMA

공통점 연산 과정 내에 차분을 통한 정상화 포함

ARIMA는 추세만 있는 경우에만 적용!

차이점 추세 뿐만 아니라 계절성이 존재하는 비정상 시계열에 적용 가능!

SARIMA 모형의 정의

시계열분석팀 클린업 1주차 …

비정상 시계열 : $X_t = s_t + Y_t$

계절 성분

결정적*(deterministic)* 계절 성분,

모든 주기에 있어서 계절 성분이 동일함을 가정

실제 상황에서는…

주기마다 계절성 다를 수 있음

SARIMA 모형의 정의

시계열분석팀 클린업 1주차 …

비정상 시계열 : $X_t = s_t + Y_t$

SARIMA 모형

계절성 사이의 상관관계에 대해서 모델링하는 확률적 접근법

결정적(deterministic) 계절성분,

계절성이 주기마다 동일하지 않지만, 각각의 계절성 사이 연관성 존재

모든 주기에 있어서 계절성분이 동일함을 가정

실제 상황에서는…

주기마다 계절성 다를 수 있음

SARIMA 모형의 정의

	January	February	…	December
Year 1	Y_1	Y_2	…	Y_{12}
Year 2	Y_{13}	Y_{14}	…	Y_{24}
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	…	$Y_{12+12(r-1)}$

EXAMPLE : 주기가 s=12인 시계열 데이터

⋮

확률적 계절성분을 고려하는 모델

SARIMA 모형의 정의

EXAMPLE : 주기가 s=12인 시계열 데이터

동일한 월에 속하는 시계열끼리의 상관관계

	January	February	…	December
Year 1	Y_1	Y_2	…	Y_{12}
Year 2	Y_{13}	Y_{14}	…	Y_{24}
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	…	$Y_{12+12(r-1)}$

ex) 1월에 속하는 $Y_1, Y_{13}, Y_{25}, \ldots, Y_{1+12(r-1)}$를 하나의 열로 고정

각 열을 고정

1월 ~ 12월까지의 각 월이 동일한 ARMA(P,Q) 모형을 따른다고 가정

⋮

= 각 월은 ARMA(P,Q)를 따르는 계절성을 갖는다

3 SARIMA

SARIMA 모형의 정의

EXAMPLE : 주기가 s=12인 시계열 데이터

동일한 월에 속하는 시계열끼리의 상관관계

	January	February	…	December
Year 1	Y_1	Y_2	…	Y_{12}
Year 2	Y_{13}	Y_{14}	…	Y_{24}
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	…	$Y_{12+12(r-1)}$

ex) 1월에 속하는 $Y_1, Y_{13}, Y_{25}, \ldots, Y_{1+12(r-1)}$를 하나의 열로 고정

각 열을 고정

1월 ~ 12월까지의 각 월이 동일한 ARMA(P,Q) 모형을 따른다고 가정

매 주기마다 계절성이 동일하지는 않지만

Year 1의 계절 성분과 Year 2의 계절 성분의 상관관계가 모델에 의해 설명 가능

= 각 월은 ARMA(P,Q)를 따르는 계절성을 갖는다

SARIMA 모형의 정의

EXAMPLE : 주기가 s=12인 시계열 데이터

각 월끼리의 상관관계

	January	February	…	December
Year 1	Y_1	Y_2	…	Y_{12}
Year 2	Y_{13}	Y_{14}	…	Y_{24}
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	…	$Y_{12+12(r-1)}$

Ex) Year 1에 속하는 $Y_1, Y_2, Y_3, \ldots, Y_{12}$*를 하나의 행으로 고정*

각 행을 고정

동일한 연도에 속한 연속된 값들에 대해 ARMA(p, q) 모형을 따른다고 가정

⋮

연속된 시계열끼리의 상관관계

SARIMA 모형의 정의

EXAMPLE : 주기가 s=12인 시계열 데이터

각 열과 행의 상관관계

	January	February	…	December
Year 1	Y_1	Y_2	…	Y_{12}
Year 2	Y_{13}	Y_{14}	…	[illegible]
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	…	$Y_{12+12(r-1)}$

~ ARMA(P, Q)

~ ARMA(p, q)

열 : ARMA(P, Q) 모형을 통해 각 주기의 계절 성분 표현

행 : ARMA(p, q) 모형을 통해 주기 내에서의 연속된 시계열끼리의 상관관계 표현

SARIMA 모형의 정의

열 : ARMA(P, Q) 모형을 통해 각 주기의 계절 성분 표현

행 : ARMA(p, q) 모형을 통해 주기 내에서의 연속된 시계열끼리의 상관관계 표현

수식으로 확인해보자…

$$\Phi(B^{12})Y_t = \Theta(B^{12})\phi^{-1}(B)\theta(B)Z_t$$

$$\phi(\mathrm{B})\Phi(\mathrm{B}^{12})\mathrm{Y}_\mathrm{t} = \theta(B)\Theta(B^{12})Z_t, \qquad Z_t \sim WN(0,\sigma^2)$$

12시점 전과의 상관관계, 주기마다의 계절 성분

바로 전 시점과의 상관관계로 연속된 시계열끼리의 상관관계

SARIMA 모형의 정의

열 : ARMA(P, Q) 모형을 통해 각 주기의 계절 성분 표현

행 : ARMA(p, q) 모형을 통해 주기 내에서의 연속된 시계열끼리의 상관관계 표현

수식으로 확인해보자…

$$\Phi(B^{12})Y_t = \Theta(B^{12})\phi^{-1}(B)\theta(B)Z_t$$

$$\phi(\mathrm{B})\Phi(\mathrm{B}^{12})\mathrm{Y}_\mathrm{t} = \theta(B)\Theta(B^{12})Z_t, \qquad Z_t \sim WN(0, \sigma^2)$$

12시점 전과의 상관관계, 주기마다의 계절 성분

Seasonal ARMA 모델 : SARMA(p, q)(P, Q)

3 SARIMA

SARIMA 모형의 정의

Seasonal ARMA 모델
SARMA(p, q)(P, Q) + 차분 = SARIMA(p, d, q)(P, D, Q)

$$Y_t = (1-B)^d (1-B^{12})^D X_t$$

$$\phi(B)\Phi(B^{12})(1-B)^d(1-B^{12})^D X_t = \theta(B)\Theta(B^{12})Z_t, \qquad Z_t \sim WN(0, \sigma^2)$$

전체 시계열의 추세(Y_t)에 대한 d차 차분

계절 성분이 갖는 D차의 추세에 대해 lag-12차분을 d번 적용

3 SARIMA

SARIMA 모형의 정의

정리를 해보자…

SARIMA(p, d, q)(P, D, Q)

연속된 시계열에 대한 모델링

p : AR의 차수

d : 전체 시계열이 갖는 추세 제거

q : MA의 차수

SARIMA 모형의 정의

정리를 해보자…

SARIMA(p, d, q)(P, D, Q)

연속된 시계열에 대한 모델링

p : AR의 차수

d : 전체 시계열이 갖는 추세 제거

q : MA의 차수

계절 성분에 대한 모델링

P : AR의 차수

D : 계절 성분이 갖는 추세 제거

Q : MA의 차수

SARIMA 모형의 정의

정리를 해보자…

SARIMA(p, d, q)(P, D, Q)

모든 모수(p, d, q, P, D, Q)를 사용할 필요 없이,

필요에 따라 **다양한 조합으로 모델링** 가능!

ex) SARIMA(0, 0, 0)(1, 0, 0) , SARIMA(1, 1, 0)(1, 0, 0)

연속된 시계열에 대한 모델링

p : AR의 차수
d : 전체 시계열이 갖는 추세 제거
q : MA의 차수

계절 성분에 대한 모델링

P : AR의 차수
D : 계절 성분이 갖는 추세 제거
Q : MA의 차수

SARIMA 모형 적합 절차

Step 1) 등분산성 조건

TS plot 또는 잔차를 통해 이분산성이 나타나는지 확인

→ 변환을 통해 등분산성 만족시키기

로그, 제곱근, Box-Cox

SARIMA 모형 적합 절차

Step 1) 등분산성 조건

TS plot 또는 잔차를 통해 이분산성이 나타나는지 확인

⟶ 변환을 통해 등분산성 만족시키기

Step 2) 차분 또는 계절 차분 여부 결정

1차 또는 2차 차분을 적용해 ACF 그래프 변화를 확인

⟶ d와 D의 차수 결정

3 SARIMA

SARIMA 모형 적합 절차

Step 1) 등분산성 조건

ACF 그래프가 느리게 감소 → 정상화 필요

ACF 그래프가 느리게 감소, 규칙적으로 구불거림 → 계절성분 존재

Step 2) 차분 또는 계절 차분 여부 결정

1차 또는 2차 차분을 적용해 ACF 그래프 변화를 확인

→ d와 D의 차수 결정

SARIMA 모형 적합 절차

Step 3) P, Q와 p, q 차수 결정

AR 또는 MA만 사용되는 경우 (SAR, SMA 모형)

ACF, PACF 그래프가 절단되는 차수 확인해 결정

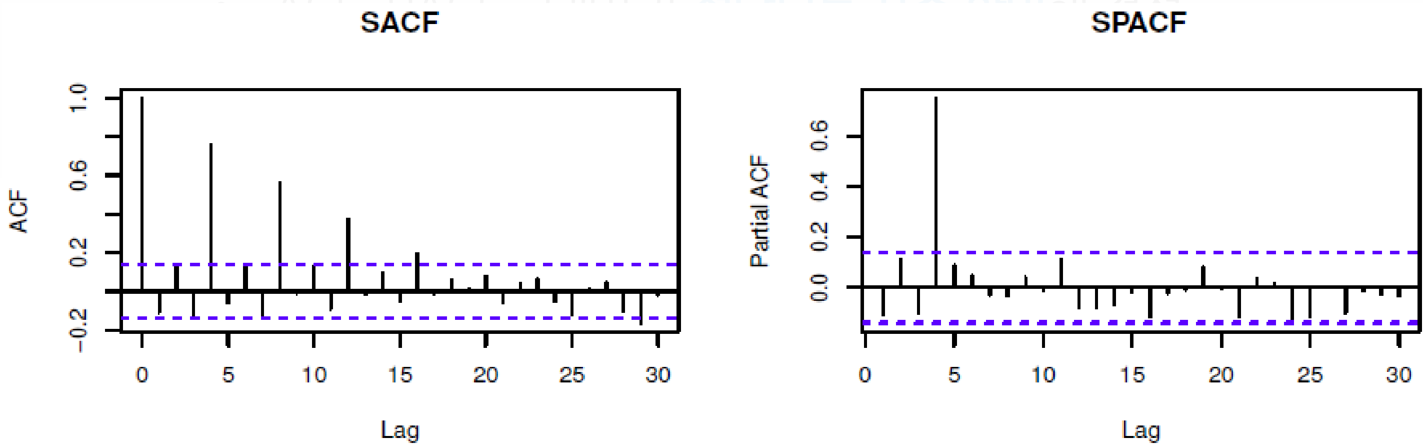
계절성분의 모수 (P, Q)를 결정할 때는 주기마다의 그래프 확인

SARIMA 모형 적합 절차

Step 3) P, Q와 p, q 차수 결정

Ex) AR 또는 MA만 사용되는

주기가 4인 시계열 데이터



4시점 간격으로 지수적으로 감소

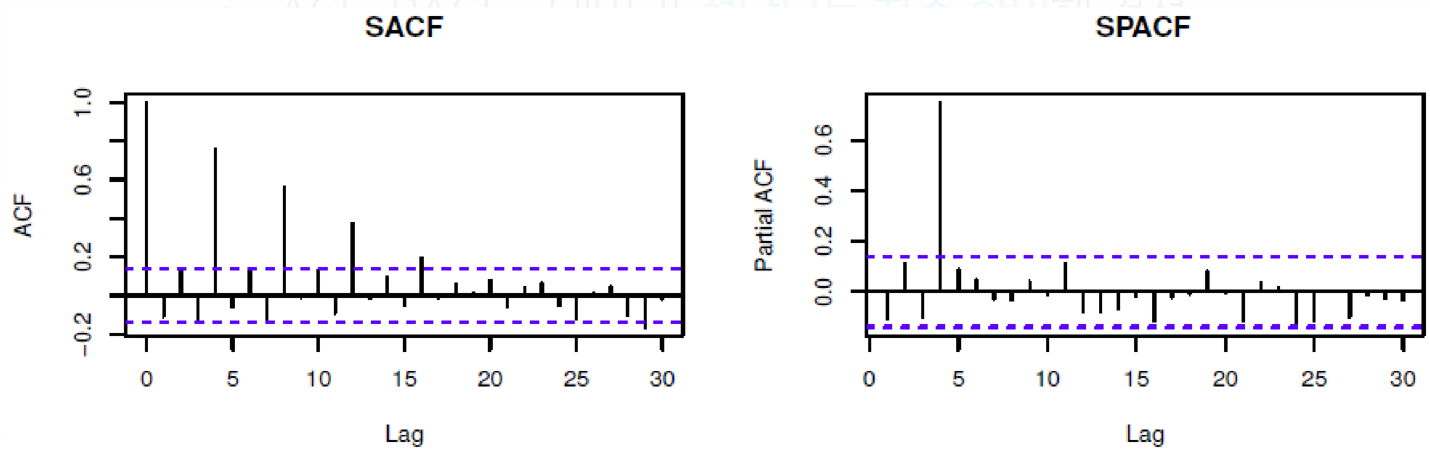
4 이후로 절단됨

SARIMA 모형 적합 절차

Step 3) P, Q와 p, q 차수 결정

Ex) AR 또는 MA만 사용되는

주기가 4인 시계열 데이터



4시점 간격으로 지수적으로 감소

4 이후로 절단됨

SAR(1) 모형 적합

SARIMA 모형 적합 절차

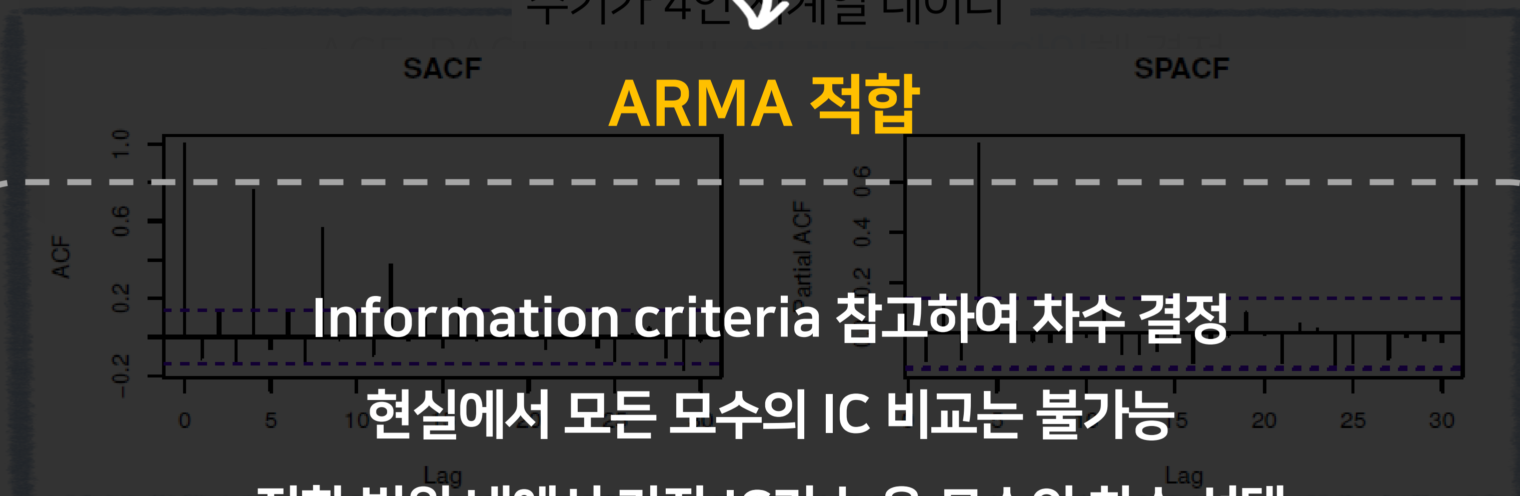
ACF와 PACF 모두에서 절단되는 부분이 보이지 않는 경우

Step 3) P, Q와 p, q 차수 결정

Ex) AR 또는 MA만 사용되는 경우

주기가 4인 시계열 데이터

ARMA 적합



Information criteria 참고하여 차수 결정

현실에서 모든 모수의 IC 비교는 불가능

정한 범위 내에서 가장 IC가 높은 모수의 차수 선택

4시점 간격으로 지수적으로 감소

4 이후로 절단됨

SAR(1) 모형 적합

SARIMA 모형 적합 절차

Step 4) 모수 추정

R 커맨드를 통해 MLE 추정

arima(data, order=c(p,d,q), seasonal=list(order=(P,D,Q), period=s))

SARIMA 모형 적합 절차

Step 4) 모수 추정

R 커맨드를 통해 MLE 추정

```
arima(data, order=c(p,d,q), seasonal=list(order=(P,D,Q), period=s))
```

Step 5) 예측

R 커맨드를 통해 예측

```
forecast(model, n.ahead=h)
```

몇 시점 이후 값을 예측할 것인지 설정

자세한 내용은 2주차 클린업 참고!

4

이분산 시계열 모형

이분산 시계열 모형

환율, 주가 등 금융 관련 시계열 데이터에서는

불확실성을 나타내는 분산 움직임을 다룸

4 이분산 시계열 모형

이분산 시계열 모형

환율, 주가 등 금융 관련 시계열 데이터에서는

불확실성을 나타내는 분산 움직임을 다룸

경제학의 변동성 → 통계학의 '조건부 분산'

미래의 변동이 현재까지의 상황에 의존

수익의 통계적 특성

수익율에 대한 정의

Simple return

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Log return

$$r_t = log\, P_t - log\, P_{t-1} = log\,(1 + R_t) \approx R_t$$

수익의 통계적 특성

수익률에 대한 정의

Simple return

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Log return

$$r_t = log\, P_t - log\, P_{t-1} = log\,(1 + R_t) \approx R_t$$

Simple return에 근사

덧셈으로 표현되어 계산 편리

Log를 통한 분산 안정화 효과

수익의 통계적 특성

수익율에 대한 정의

Simple return

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

앞으로 다루는 모든 수익율은 log return

Log return

$$r_t = log\ P_t - log\ P_{t-1} = log\ (1 + R_t) \approx R_t$$

Simple return에 근사

덧셈으로 표현되어 계산 편리

Log를 통한 분산 안정화 효과

수익의 통계적 특성

1) 평균에 대칭
2) r_t 자체는 상관관계 크지 않음 (ACF가 거의 백색잡음에 가까움)
3) $|r_t|$ 또는 ${r_t}^2$은 강한 상관관계 존재
4) QQ plot의 양 끝이 정규분포에서 크게 벗어나는 heavy-tailed distribution
5) Local variance는 시간에 의존하여 발생하는 조건부 이분산성

4 이분산 시계열 모형

수익의 통계적 특성

1) 평균에 대칭

2) r_t 자체는 상관관계 크지 않음 (ACF가 거의 백색잡음에 가까움)

3) $|r_t|$ 또는 $r_t{}^2$은 강한 상관관계 존재

4) QQ plot의 양 끝이 정규분포에서 크게 벗어나는 heavy-tailed distribution

5) Local variance는 시간에 의존하여 발생하는 조건부 이분산성

$|r_t|$ 또는 $r_t{}^2$는 변동(분산)을 의미

수익율의 분산은 강한 상관관계를 가짐

변동성이 시점에 의존하는 조건부 이분산성

4 이분산 시계열 모형

ARCH

ARCH (Auto-Regressive Conditional Heteroscedasticity)

수익율에 대해 모델링하는 모형

$$r_t = \sigma_t \varepsilon_t \sim N(0, {\sigma_t}^2)$$

$$\sigma_t{}^2 = Var(r_t | F_{t-1}) = a_0 + a_1 r_{t-1}{}^2$$

t-1 시점까지의 모든 정보의 집합

$r_t | F_{t-1}$ t-1까지의 모든 정보를 다 알고 있을 때의 t 시점의 수익율

ARCH(1) ${\sigma_t}^2$에 대해 AR(1) 구조를 가정한 것

4 이분산 시계열 모형

ARCH

ARCH (Auto-Regressive Conditional Heteroscedasticity)

수익율에 대해 모델링하는 모형

$$r_t = \sigma_t \varepsilon_t \sim N(0, {\sigma_t}^2)$$

$$ {\sigma_t}^2 = Var(r_t | F_{t-1}) = a_0 + a_1 {r_{t-1}}^2 $$

${\sigma_t}^2$이 ${r_{t-1}}^2$에 의존, ${r_t}^2$도 ${r_{t-1}}^2$에 의존함

ARCH(1) ${\sigma_t}^2$에 대해 AR(1) 구조를 가정한 것

ARCH

ARCH(m) 모형

$\sigma_t{}^2$에 대해 AR(m) 구조를 가정한 것

$$r_t = \sigma_t \varepsilon_t \sim N(0, \sigma_t{}^2)$$

$$\sigma_t{}^2 = a_0 + a_1 r_{t-1}{}^2 + \cdots + a_m r_{t-m}{}^2 \approx r_t{}^2$$

GARCH

GARCH (Generalized Auto-Regressive Conditional Heteroscedasticity)

수익율에 대해 모델링하는 모형

${\sigma_t}^2$에 대해 ARMA(p, q) 구조 가정

GARCH

GARCH (Generalized Auto-Regressive Conditional Heteroscedasticity)

수익율에 대해 모델링하는 모형

${\sigma_t}^2$에 대해 ARMA(p, q) 구조 가정

${r_t}^2$은 ${\sigma_t}^2$에 근사, ${\eta_t}^2 = {r_t}^2 - {\sigma_t}^2$

GARCH(1, 1)

$$r_t = \sigma_t \varepsilon_t \sim N(0, {\sigma_t}^2)$$

$${\sigma_t}^2 = \alpha_0 + \alpha_1 {r_{t-1}}^2 + \beta_1 {\sigma_{t-1}}^2$$

$${r_t}^2 = \alpha_0 + (\alpha_1 + \beta_1){r_{t-1}}^2 + \eta_t - \beta_1 \eta_{t-1}$$

AR(1) MA(1)

4 이분산 시계열 모형

GARCH

GARCH (Generalized Auto-Regressive Conditional Heteroscedasticity)

수익율에 대해 모델링하는 모형

${\sigma_t}^2$에 대해 ARMA(p, q) 구조 가정

${r_t}^2$은 ${\sigma_t}^2$에 근사, ${\eta_t}^2 = {r_t}^2 - {\sigma_t}^2$

GARCH(p, q)

$$r_t = \sigma_t \varepsilon_t \sim N(0, {\sigma_t}^2)$$

$${\sigma_t}^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i {r_{t-i}}^2 + \sum_{j=1}^{q} \beta_i {\sigma_{t-j}}^2$$

5

시계열 데이터 with ML/DL

Intro

전통적인 통계적 모델들..

AR,MA,ARMA,ARIMA…

좋은 성능을 보이고 있지만,
ML/DL을 통한 접근도 좋은 성능을 보임.

ML/DL을 이용한 시계열 분석

시계열 데이터의 전처리

시계열데이터에서 결측치를

단순히 제거하려 한다면 여러 문제가 발생함

타임스탬프가 일정해지지 않음
평균과 분산에 왜곡이 생김..etc

일반적인 경우 평균 또는 최빈값으로 결측치를 대체하는데

시계열 데이터의 경우 시계열의 특성을 반영한 방법을 이용해야 함

결측치 보간법

LOCF(Last Observation Carried Forward)

직전에 관측된 값으로 결측치 보간

NOCB(Next Observation Carried Backward)

직후에 관측된 값으로 결측치 보간

Moving Average / Moving Median

Window 내에서의 평균 또는 중앙값으로 결측치 보간

결측치 보간법

LOCF

NOCB

Moving Average / Moving Median

Window 내에서의 평균 또는 중앙값으로 결측치 보간

패턴이 급변하는 경우

결측치 전, 후로 패턴이 급격히 변화하는 구간에서는
해당 보간법들로는 충분하지 못함

예를 들어, 주가 예측 시 주가가 급격하게 상승/하락하는 구간에서
해당 보간 방법들은 평균과 분산에 왜곡을 일으킬 가능성이 큼

다른 방법 필요

결측치 보간법

선형 / 비선형 보간법

결측치를 중심으로 window를 설정 후 구간 내에 함수를 적합하고 이를 통해 결측치 보간

스플라인 보간법(Spline Interpolation) (데마팀 2주차 클린업 참고!)

결측치를 중심으로 한 window의 전체 값을 한 번에 적합하지 않고, 이를 소구간으로 분할 하여 보간

모델링을 통한 결측값 예측

결측치가 너무 많은 경우 각 결측치마다 이전까지의 데이터로 모델링하고, 그 값을 예측하여 보간

노이즈 처리 (Denoising)

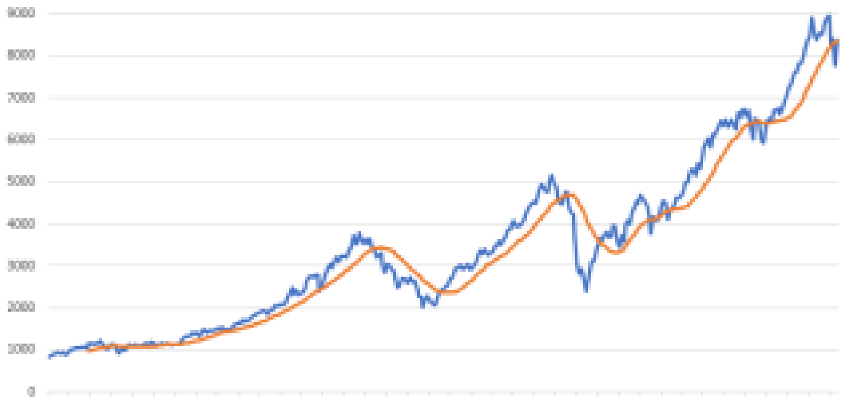
노이즈

의도하지 않은 데이터의 왜곡을 불러오는 모든 것을 의미

Example

평균적인 대중교통 이용량에 대해 분석하고자 할 때,
유명 가수의 콘서트로 인해 이용량이 급격하게 상승한 것은
관점에 따라 노이즈로 볼 수 있음

노이즈 처리 (Denoising)

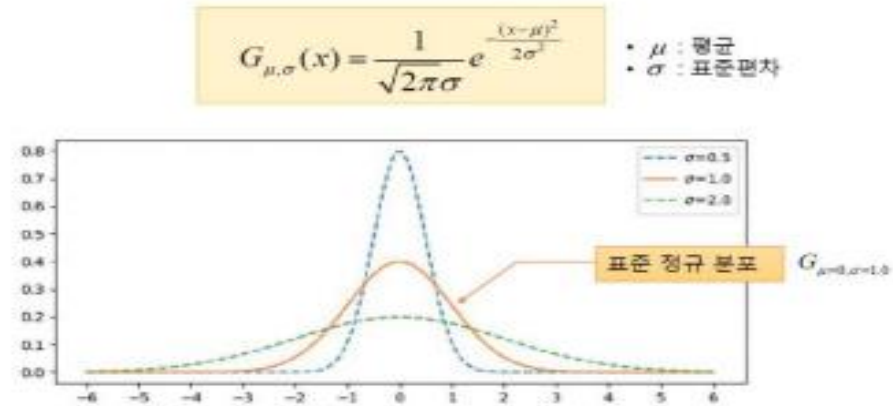


평활(Smoothing)

이동평균평활법 혹은 지수평활법 을 통해 denoising 가능

노이즈가 많이 발생하는 데이터에서는 노이즈 자체가 평균에 반영되기 때문에 적절하지 않고, 노이즈가 적은 데이터에 효과적

노이즈 처리 (Denoising)



가우시안 필터(Gaussian Filter)

일종의 평활 방법으로, 가우시안 필터를 적용하여 denoising 가능
중심에서 멀어질수록 작은 가중치를 부여해서 이미지처리에서 부드러움 효과 혹은 잡음제거에 사용되나, 시계열 데이터에도 적용 가능함

시계열 데이터의 교차검증

교차검증(Cross Validation)

과적합을 방지하기 위한 방법 중 하나로

데이터를 train,validation set으로 나눠서 모델을 학습,검증

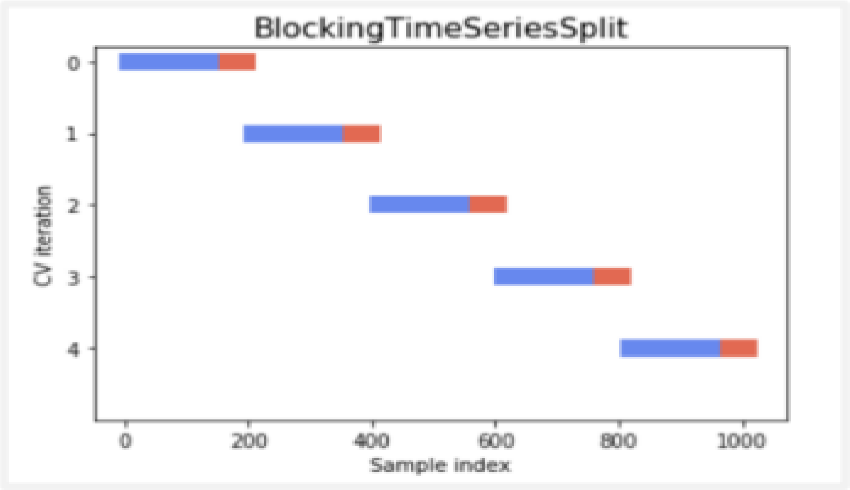
일반적으로 사용하는 CV방법들은 시계열 데이터에서

매우 중요한 **시간의 순서를 고려하지 않기 때문**에 사용 할 수 없음

Blocked Time Series CV

Time Series CV

시계열 데이터의 교차검증



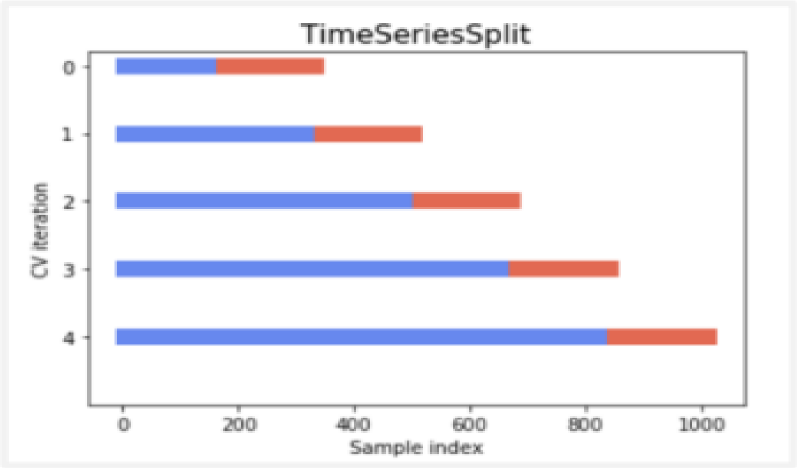
Blocked Time Series CV

Rolling window CV 라고도 불림

동일한 사이즈의 window를 옆으로 이동시키며

window 내에서 일정한 비율로 train과 validation을 분할함

시계열 데이터의 교차검증



Time Series CV

Expanding window CV 라고도 불림

Window를 계속해서 누적시켜

이전 단계의 train과 validation을 다음 단계의 train set으로 활용

5 시계열 데이터 with ML/DL

시계열 데이터와 딥러닝 기법

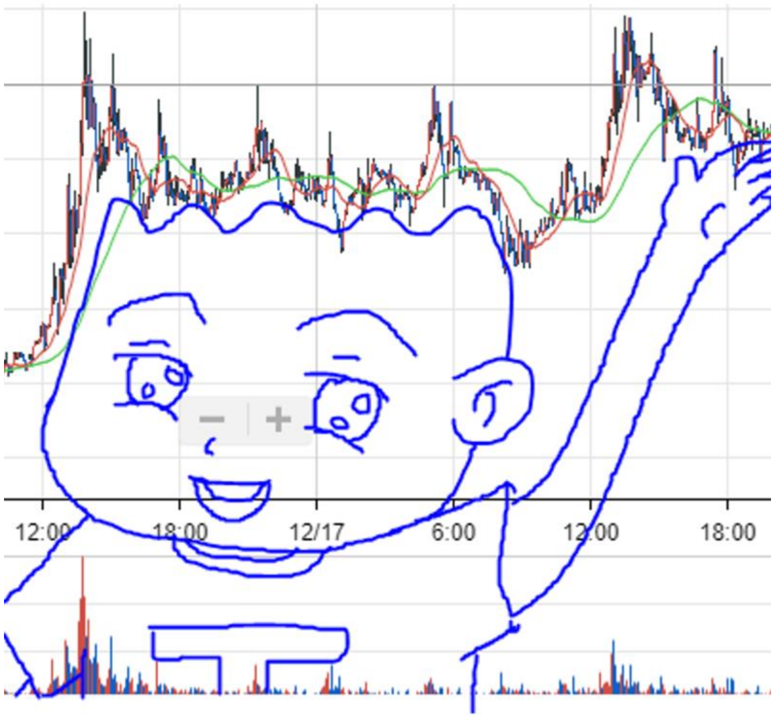
시계열 데이터 중 최근

가장 이목이 쏠리는 분야는 바로 **주식**

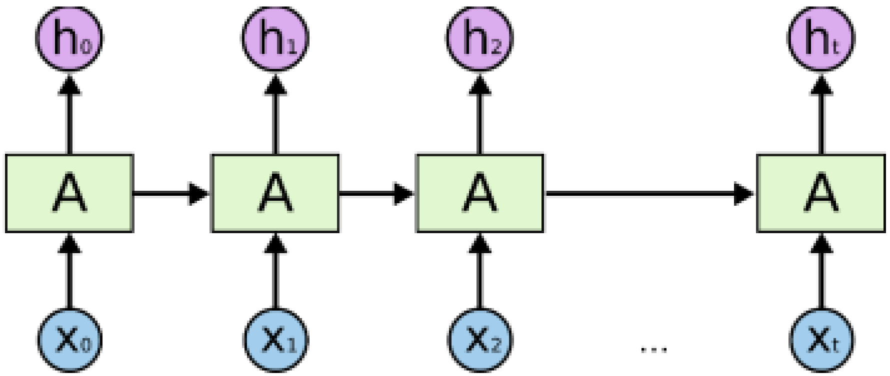
특히 딥러닝 기법을 활용한 주가 예측

연구가 자주 보임

딥러닝 기법을 활용한 시계열 데이터 분석



LSTM



LSTM(Long Short-Term Memory Network)

LSTM은 RNN에서 파생되었고(A부분이 RNN보다 더 복잡해짐)

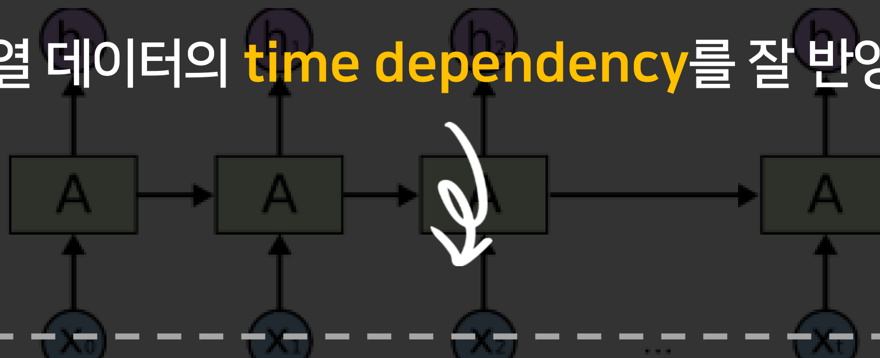
RNN의 장기 의존성 문제를 개선한 모델

(구체적인 내용은 딥러닝팀 3주차 클린업 참고!)

LSTM

RNN 기반 모델

LSTM은 시계열 데이터의 time dependency를 잘 반영 할 수 있는 모델



RNN 기반 모델은 이전의 학습 결과가 다음 학습 단계에 영향을 주기 때문에

종속적인 데이터를 처리하는데 **효과적**임

LSTM(Long Short-Term Memory Network)

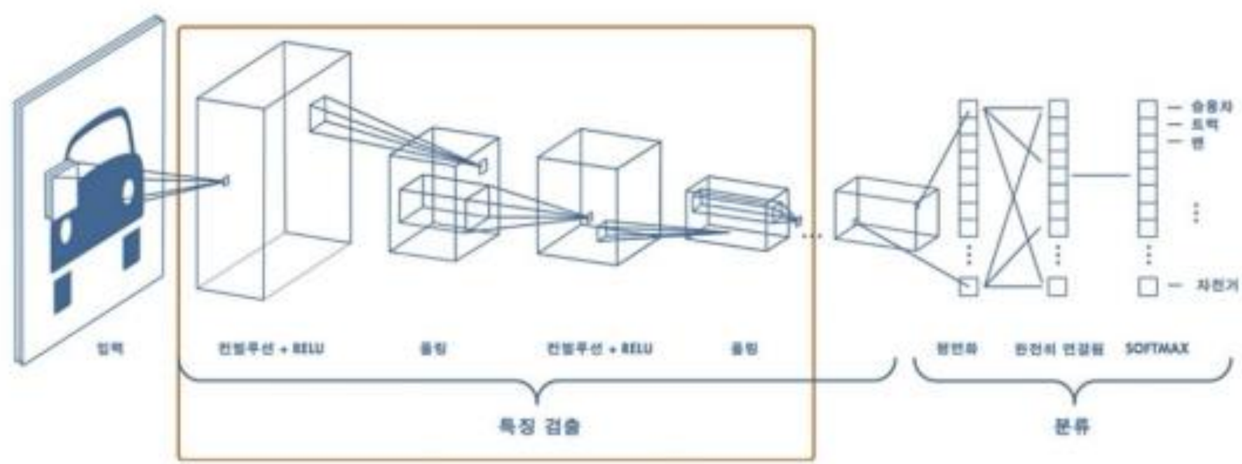
LSTM은 RNN에서 파생되었고 (A부분이 RNN보다 더 복잡해짐)

RNN의 장기 의존성 문제를 개선한 모델

randomwalk와 같이 데이터 간의 **종속성이 불분명할 경우** 효과적이지 못함

(구체적인 내용은 딥러닝팀 3주차 클린업 참고!)

CNN



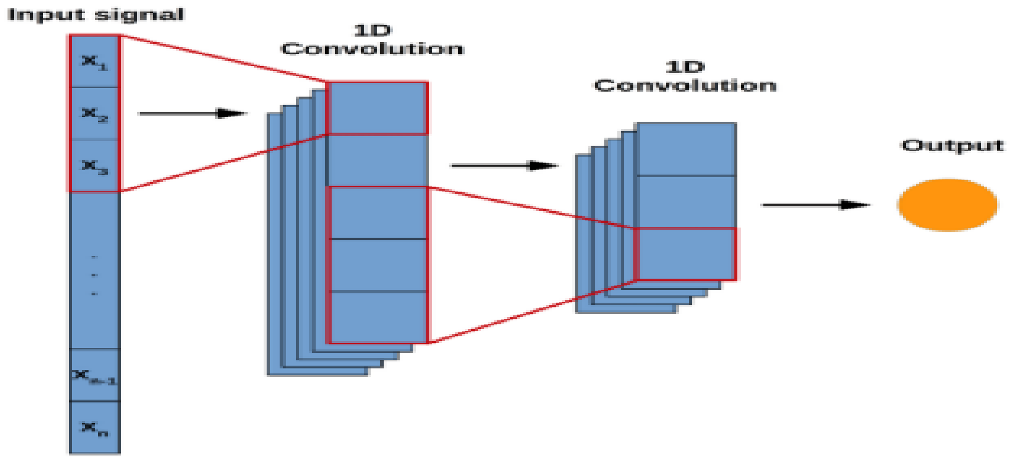
CNN(Convolutional Neural Network)

CNN은 이미지 처리에서 많이 사용되며

이미지의 특징을 추출하여 학습하는 모델임

(구체적인 내용은 딥러닝 2주차 클린업 참고!)

CNN

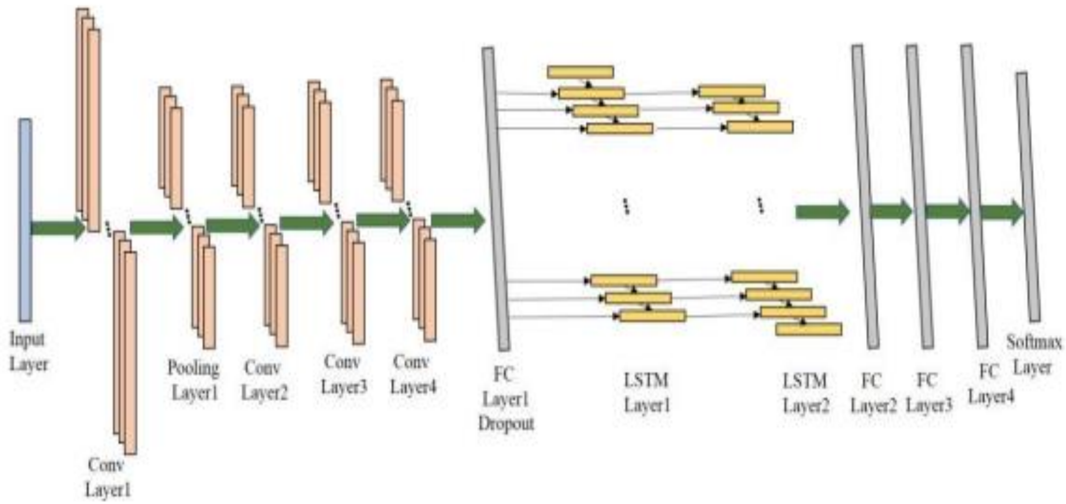


시계열 데이터에 CNN을 사용하면 순간순간의 값 보다는 전체적인 그림에 주목함

시계열 데이터의 전체적인 추세 등을 파악해서

전반적인 패턴을 파악하고 특징을 추출하여 다음 시점을 예측함

CNN+LSTM



CNN+LSTM

CNN을 이용해서 피처맵을 추출하고

Flatten하여 LSTM에서 학습 후, FC layer를 거쳐 결과값 도출

감사합니다!