

시계열자료분석팀

5팀

장다연

심현구

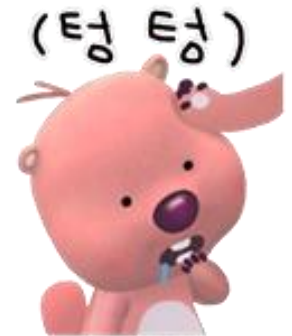
윤세인

이동기

천예원

1

지난 주차 흐름 정리 및 복습

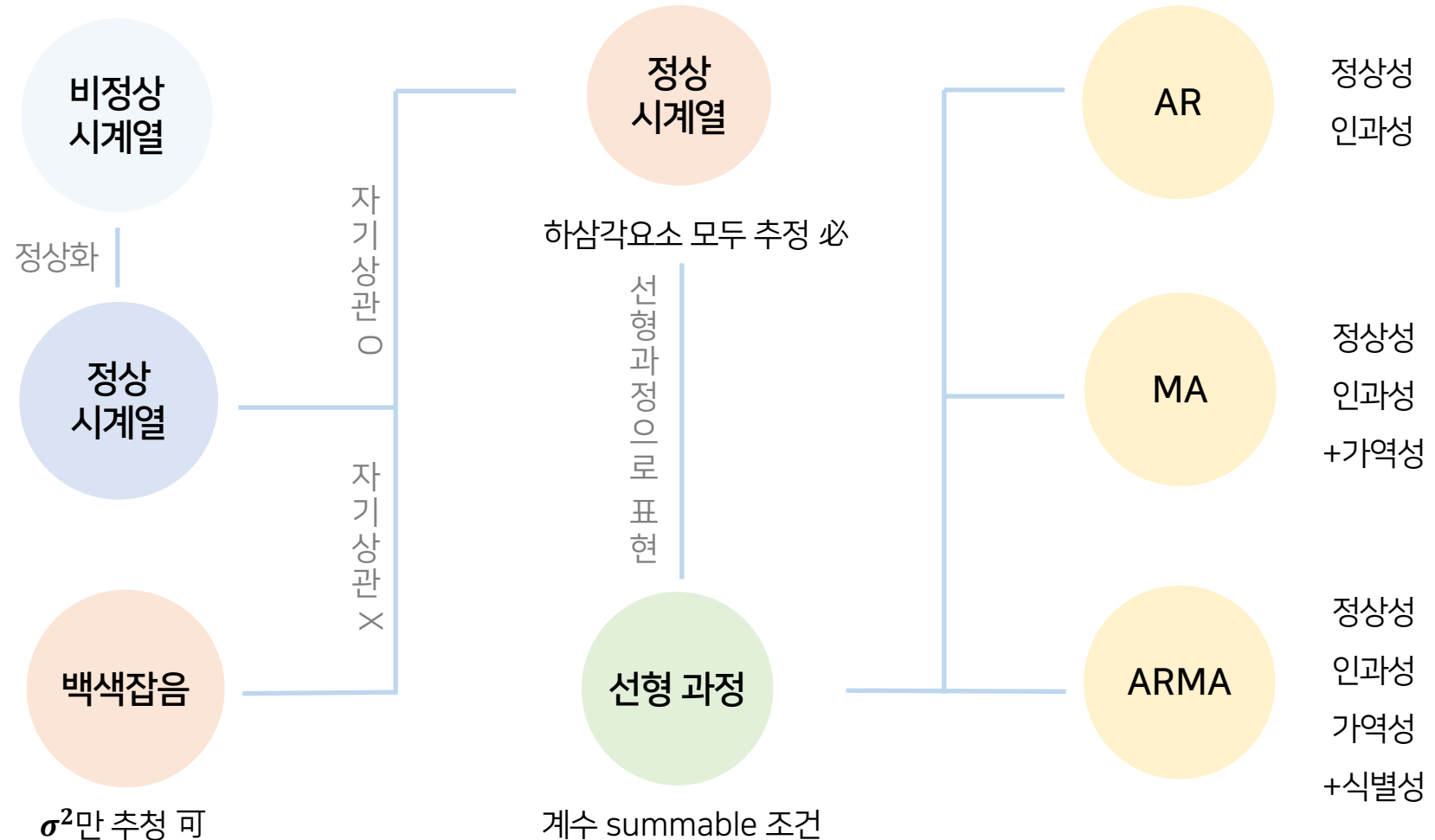


기억이 안나요...

1

지난 주차 흐름 정리 및 복습

시계열 자료 분석 흐름정리



1

지난 주차 흐름 정리 및 복습

정상시계열모형 적합 절차

STEP 1) 모형 식별

AR & MA는 ACF 또는 PACF를 확인해 차수 결정
ARMA는 최소 IC(Information Criteria)를 가지는 모형 선택

	AR(p)	MA(q)	ARMA(p,q)
ACF	지수적으로 감소	q+1 시점부터 절단	지수적으로 감소
PACF	p+1 시점부터 절단	지수적으로 감소	지수적으로 감소

STEP 2) 모수 추정

MLE / LSE / MME 등의 추정량으로 모수 추정

정상시계열모형 적합 절차

STEP 3) 모형 진단

모수에 대한 검정	정상성, 가역성, 식별성 등 만족 여부 확인
	모수 $\neq 0$ 인지 확인
잔차에 대한 검정	추세, 계절성, 이상치 확인
	백색잡음 여부 확인
	정규성 만족 여부 확인

STEP 4) 예측

가지고 있는 데이터의 선형결합을 활용해 미래 예측



MSPE(Mean Squared Prediction Error) 최소화

2

ARIMA

모형의 정의

ARIMA (자기회귀누적이동평균과정)

“d차 차분의 결과 오차항 Y_t 가 ARMA 모형을 만족하는 경우
원본 시계열 데이터가 ARIMA 모형을 따른다”

차분과 ARMA 모형이 결합된 모형

d차 차분

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

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

AR 모형의 특성방정식

MA 모형의 특성방정식

모형의 정의

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

데이터가 **polynomial trend**를 가지는 경우
오차가 **ARMA(p,q)**를 따르는 경우 \longrightarrow 모두 만족하는 경우 ARIMA 사용!



ARIMA

차분(Differencing)과 ARMA모형의 결합인데
ARDMA 모형이 아닌 **ARIMA 모형**인 이유

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

$$Y_t = (1 - B)X_t = X_t - X_{t-1}$$

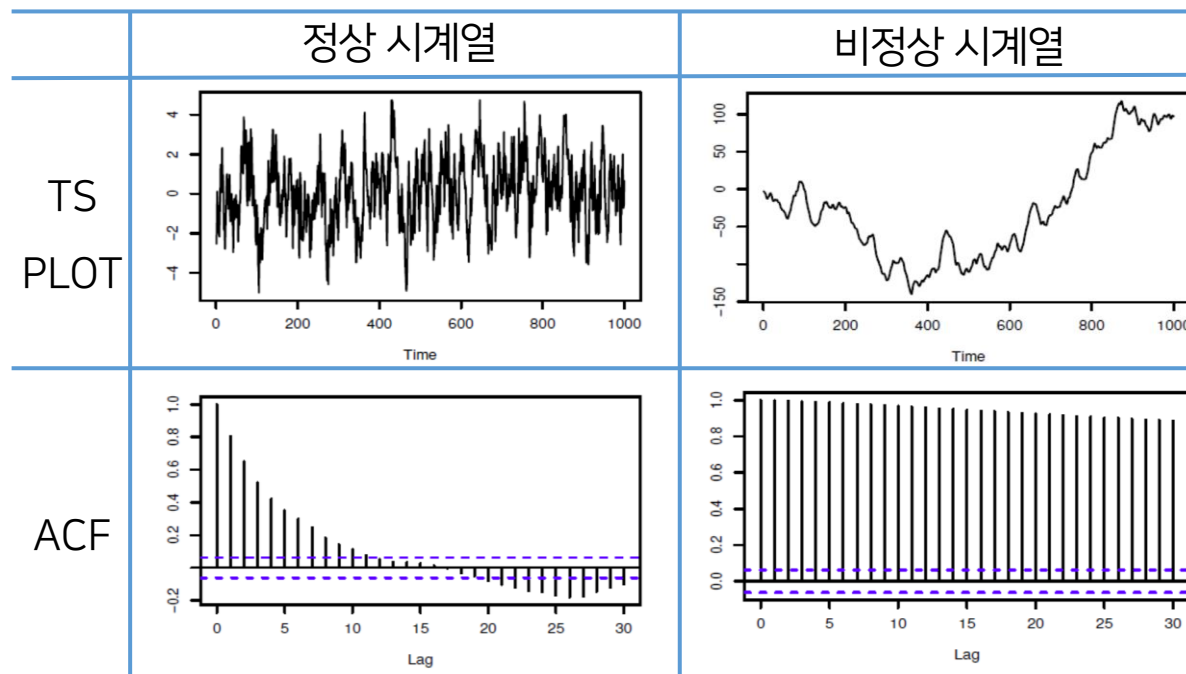
$$X_t = X_{t-1} + Y_t = (X_{t-2} + Y_{t-1}) + Y_t = \cdots = X_0 + \sum_{j=1}^t Y_j$$

X_t 가 Y_j 의 **누적합**으로 표현

ARIMA의 I는 **누적(Integration)**을 의미

모형의 적합 절차

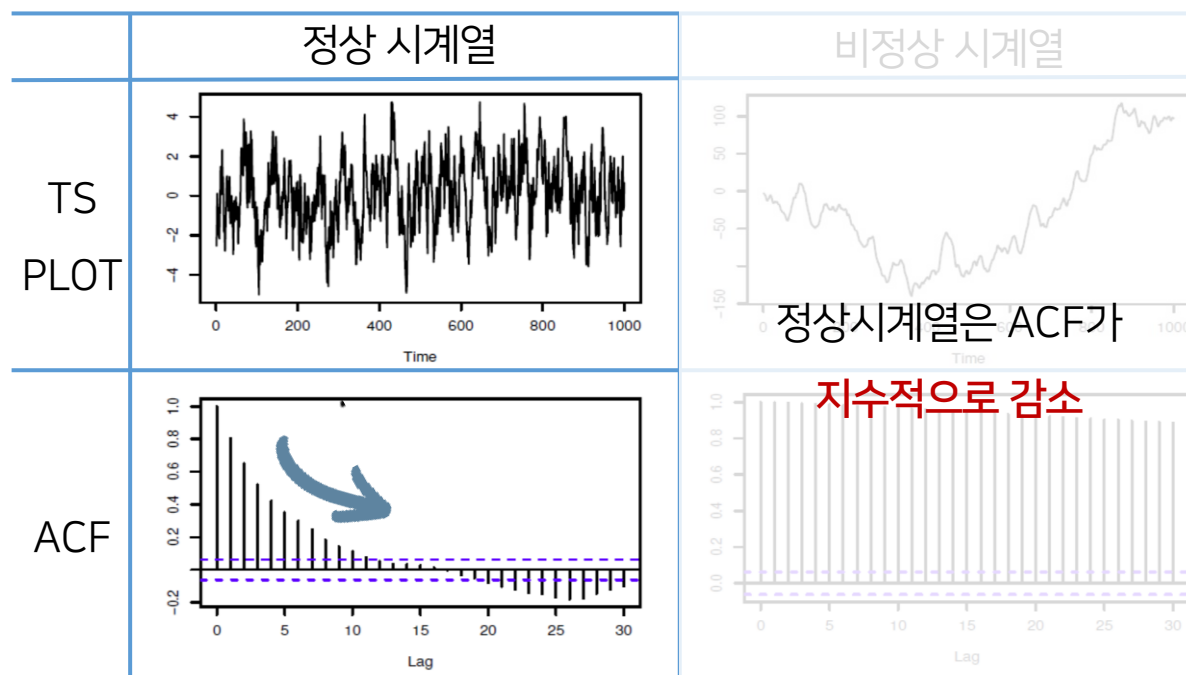
Step 1) TS plot과 ACF 그래프를 통해 정상/비정상 시계열 여부를 판단



ACF의 감소 속도를 통해 판단

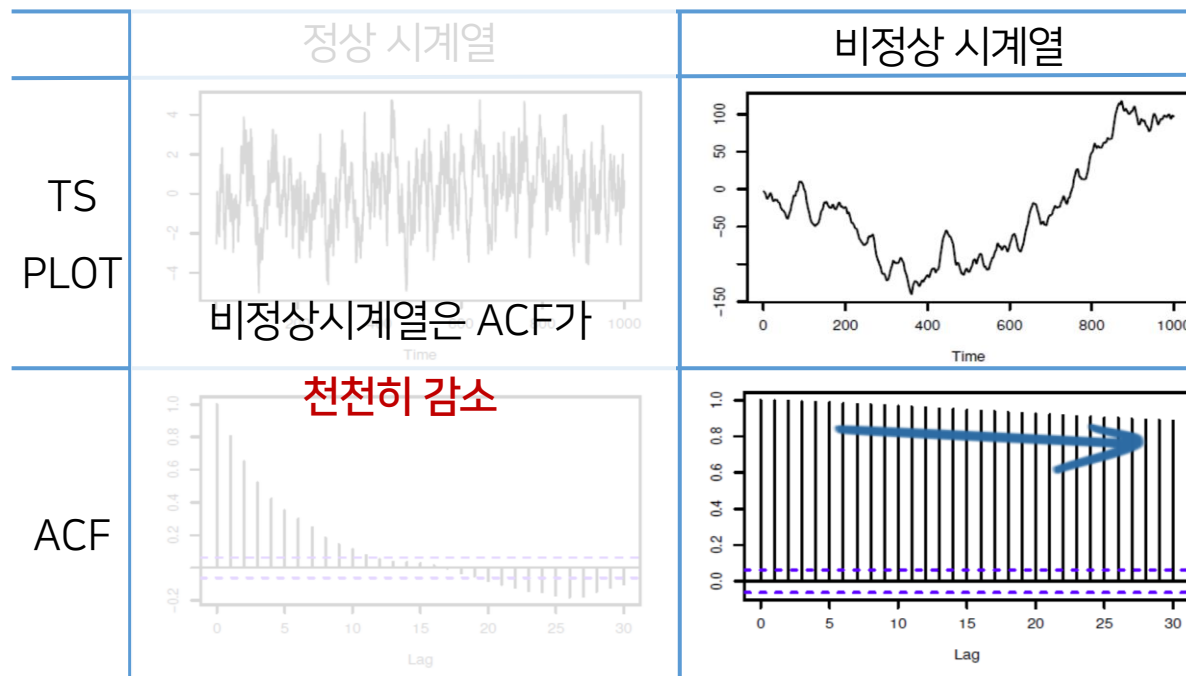
모형의 적합 절차

Step 1) TS plot과 ACF 그래프를 통해 정상/비정상 시계열 여부를 판단



모형의 적합 절차

Step 1) TS plot과 ACF 그래프를 통해 정상/비정상 시계열 여부를 판단



모형의 적합 절차

Step 2) 추세가 관측된다면 차분을 통해 정상화

⋮

ARIMA(p,d,q)의 d의 차수 결정
차수 결정 과정에서 **과대 차분**에 주의

과대차분

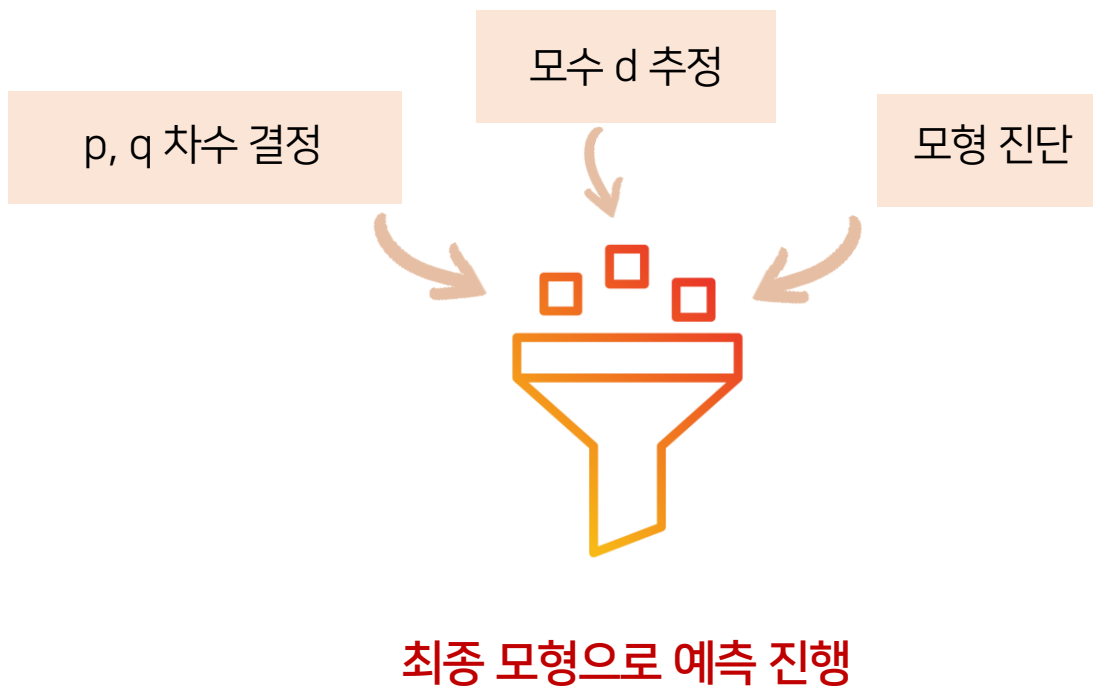
필요이상으로 차분을 진행하는 것으로, 정상성에는 문제가 없지만,
ACF가 복잡해지거나 분산이 커지며, 불필요한 상관관계를 생성

→ 적절한 차수로 차분

대부분의 데이터는 1,2차 차분만으로 정상성 만족

모형의 적합 절차

Step 3) 모형 적합 절차를 마친 모형을 통한 예측



3

SARIMA

그녀의 한마디...



모형의 정의



$$X_t = S_t + Y_t$$

1주차에서 배운 '계절성' 개념은
계절성분이 모든 주기에서 동일함을 가정함!
(deterministic)

그러나 현실의 시계열 데이터에서는 모든 계절성분이 결정적이지 않으며,

다른 성분들과의 상관관계가 존재할 수 있음!



주기의 변화에 따라 **계절성이 변화**할 수 있음을 반영한 **SARIMA 모형**

모형의 정의

SARIMA (Seasonal + ARIMA)

추세와 계절성이 모두 존재하는 비정상 시계열에 적용 가능한 모형



계절성 사이의 상관관계에 대해 모델링하는 **확률적 접근법**으로 이해 가능!

ARIMA

추세만 존재하는
시계열 데이터

VS

SARIMA

추세와 계절성이 모두 존재하는
시계열 데이터

예시를 통해 알아보는 SARIMA

	1월	2월	...	12월
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_{12r}

주기가 12 ($s=12$)인 **확률적 계절성분**을 가정하는 시계열 데이터를 통해 SARIMA모형이 계절성을 다루는 방법을 알아보시다 !

3

SARIMA

예시를 통해 알아보는 SARIMA

Step 1) 데이터의 각 열을 계절성분으로 정의

	1월	2월	...	12월
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_{12r}



1월부터 12월까지의 모든 열은 동일한 ARMA(P, Q) 모델을 따른다고 가정
 즉, 각 월은 ARMA(P, Q)를 따르는 계절성을 가짐!

3

SARIMA

예시를 통해 알아보는 SARIMA

1월부터 12월까지의 모든 열은 동일한 ARMA(P, Q) 모델을 따른다고 가정
즉, 각 월은 ARMA(P, Q)를 따르는 계절성을 가짐!



$$Y_{j+12t} - \Phi_1 Y_{j+12(t-1)} - \dots - \Phi_P Y_{j+12(t-P)} = U_{j+12t} + \Theta_1 U_{j+12(t-1)} + \dots + \Theta_Q U_{j+12(t-Q)}$$

(where $t = 0, 1, \dots, r$ $U_t \sim \text{WN}(0, \sigma_U^2)$)

$$\Phi(B^{12})Y_t = \Theta(B^{12})U_t$$

$j \in [1, 12]$ 는 month에 해당하며, j 가 달라져도 ϕ 와 θ 는 동일하다고 가정


3

SARIMA

예시를 통해 알아보는 SARIMA

Step 2) 데이터의 각 행을 고정하여 상관관계 파악

	1월	2월	...	12월
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_{12r}



동일한 연도에 속한 연속된 값들이 $ARMA(p, q)$ 를 따른다고 가정
 즉, 연속된 시계열 데이터 사이의 상관관계를 파악!

예시를 통해 알아보는 SARIMA

$$\Phi(B^{12})Y_t = \Theta(B^{12})U_t$$

Step 1에서 구한 위 식에서 Y_t 는 U_t 에 의해 표현되고 있으므로
 Y_t 와 Y_j 의 correlation은 U_t 와 U_j 에 대해 모델링 하여 구할 수 있음!

⋮

U_t 를 ARMA(p, q)를 따르는 시계열이라고 가정한다면,

$$\phi(B)U_t = \theta(B)Z_t, Z_t \sim WN(0, \sigma^2)$$

$\{Y_1, \dots, Y_{12}\}$ 와 같이 행에 존재하는 correlation을 반영하기 위해 U_t 를 백색잡음이 아닌 ARMA(p, q)로 가정!

3

SARIMA

예시를 통해 알아보는 SARIMA

	1월	2월	...	12월
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_{12r}

동일한 월에 속하는 시계열끼리의 상관관계 : ARMA(P, Q)

한 주기 내에서 연속된 시계열끼리의 상관관계 : ARMA(p, q)

예시를 통해 알아보는 SARIMA

Step 3) Step1과 Step2를 수식으로 표현

동일한 월에 속하는 시계열끼리의 상관관계 : $\text{ARMA}(P, Q)$

한 주기 내에서 연속된 시계열끼리의 상관관계 : $\text{ARMA}(p, q)$

$$\begin{aligned}\Phi(B^{12})Y_t &= \Theta(B^{12})\phi^{-1}(B)\theta(B)Z_t \\ \phi(B)\Phi(B^{12})Y_t &= \theta(B)\Theta(B^{12})Z_t, \quad Z_t \sim \text{WN}(0, \sigma^2)\end{aligned}$$

여기까지의 과정이 Seasonal ARMA로, $\text{SARMA}(p, q)X(P, Q)$ 이다.

예시를 통해 알아보는 SARIMA

Step 4) SARMA에 차분을 더해 SARIMA 완성

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

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

$$(Z_t \sim \text{WN}(0, \sigma^2))$$

 $(1-B)^d$

전체 시계열 추세에 대해 d차 차분

 $(1-B^{12})^D$

계절성분에 대해 lag-12차분을 D번 적용

d차 차분과 lag-d 차분의 표기 차이 1주차 클린업 참고!

예시를 통해 알아보는 SARIMA

Step 4) SARMA에 차분을 더해 SARIMA 완성

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

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

$$(Z_t \sim \text{WN}(0, \sigma^2))$$

(p, d, q) : 연속된 시계열

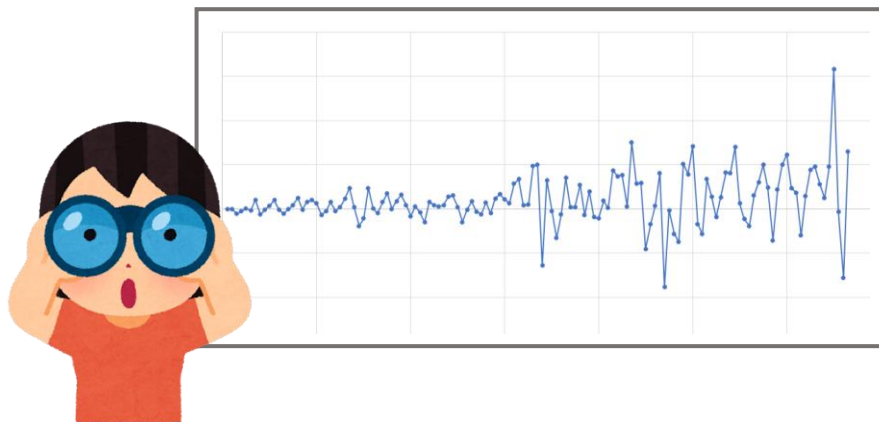
p	AR의 차수
d	전체 시계열 추세에 대한 차분
q	MA의 차수

(P, D, Q) : 계절성분

P	AR의 차수
D	계절성분에 대한 차분
Q	MA의 차수

모형 적합 절차

Step 1) 이분산성이 나타나는 경우 분산 안정화



TS plot과 잔차를 확인하여 **이분산성**이 드러나는 경우
분산 안정화를 진행한다 (log변환, Box-Cox 변환 등 ...)

1주차 클린업 참고!

모형 적합 절차

Step 2) d차 차분 / lag-d차분 진행 여부 검정

ACF그래프의 개형을 통해 d와 D의 차수를 결정

그래프가 **느리게 감소**한다면**전체 차분**을 통한 정상화가 필요! (d)그래프가 **느리게 감소**하면서
규칙적으로 구불거리는 형태라면**계절 차분**이 필요! (D)

과대차분 되지 않도록 주의!



모형 적합 절차

Step 3) p, q 와 P, Q 차수 결정

p, q 는 ARMA와 동일하게 ACF, PACF를 통해 구할 수 있음

P, Q 는 계절성분의 모수이므로 **주기**마다의 그래프를 확인해 보아야 함!



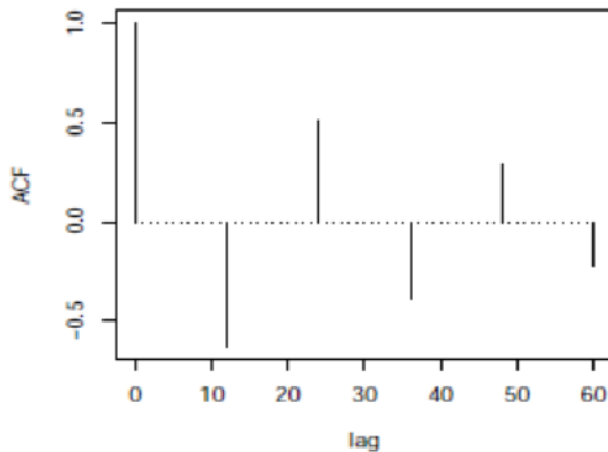
예시를 통해 확인해봅시다!



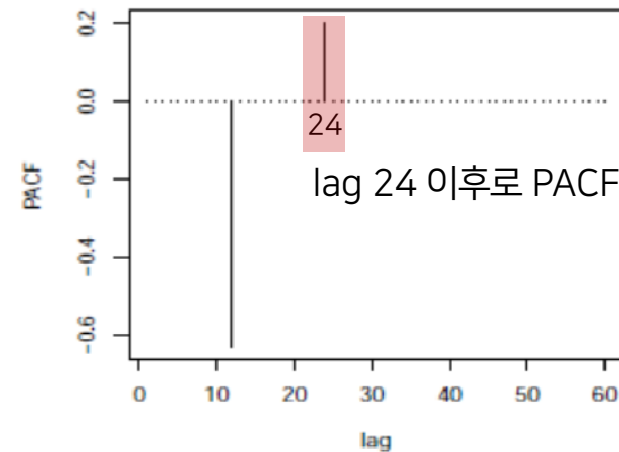
모형 적합 절차

(예시 1) 주기가 12인 시계열 데이터 A

SARIMA(0, 0, 0)(2, 0, 0) : ACF



SARIMA(0, 0, 0)(2, 0, 0) : PACF



lag 24 이후로 PACF plot **절단**

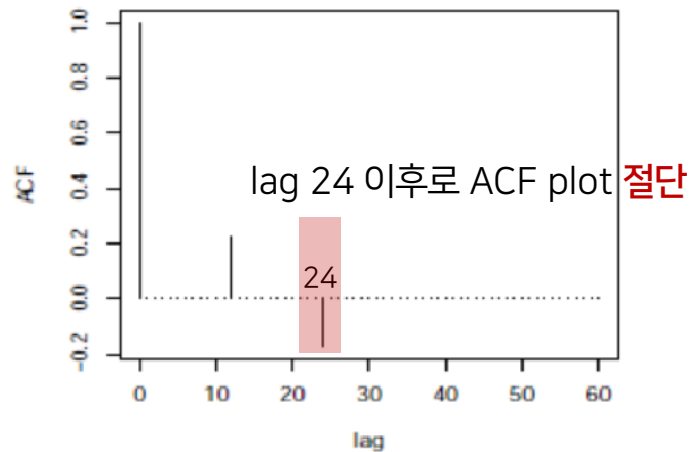
ACF는 지수적으로 감소, PACF는 시차 이후 절단되는 **AR 모형**을 따름

= 계절성분이 ARMA(2, 0)을 따름 (P=2, Q=0)

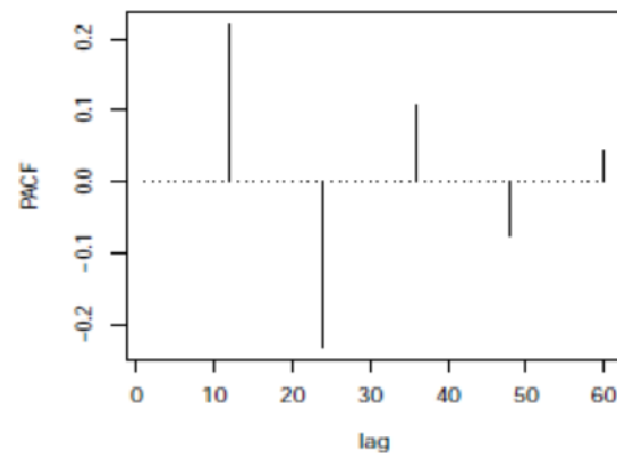
모형 적합 절차

(예시 2) 주기가 12인 시계열 데이터 B

SARIMA(0, 0, 0)(0, 0, 2) : ACF



SARIMA(0, 0, 0)(0, 0, 2) : PACF



ACF는 시차 이후로 절단, PACF는 지수적으로 감소하는 **MA 모형**을 따름

= 계절성분이 ARMA(0, 2)을 따른다 (P=0, Q=2)

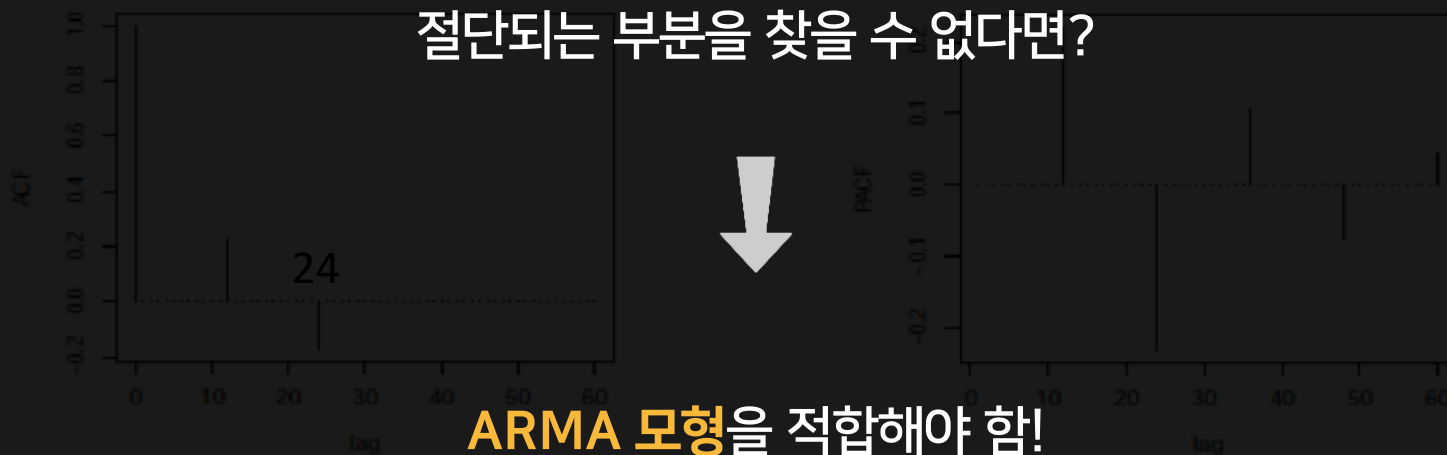
SARIMA 적합 절차



(예시 2) 주기가 12인 시계열 데이터 B

SARIMA(0, 0, 0)(0, 0, 2) : ACF SARIMA(0, 0, 0)(0, 0, 2) : PACF

ACF와 PACF 모두에서
절단되는 부분을 찾을 수 없다면?



ARMA 모형을 적합해야 함!

- ACF는 시차 이차곱을 10% 임계값을 넘지 않는 가장 낮은 차수의 차수를 선택
- Lag = 24 이후로 ACF plot이 절단됨

= 계절성분이 ARMA(0, 2)을 따른다 (P=0, Q=2)

모형 적합 절차

Step 4. 모수를 추정한 후 예측 진행

비정상 시계열 모형은 정상화과정을 진행하지 않기 때문에
예측 결과 역시 원본 데이터에 대한 최종 예측값 !



추세와 계절성을 따로 더할 필요 없음 !



4

이분산 시계열 모형

이분산 시계열 모형

지금까지 배운 시계열 모형은
분산에 변화가 없는 것을 가정 후 평균의 움직임에 관심을 갖는 모형



수익률, 주가, 환율 등 금융 관련 시계열 데이터에서는
분산이 과거 자료에 의존하는 특성을 가정



시간에 따른 이분산성

4

이분산 시계열 모형

이분산 시계열 모형

지금까지 배운 시계열 모형은
분산에 변화가 없는 것을 가정 후 평균의 움직임에 관심을 갖는 모형

경제학 분야의

수익 **변동성**

가, 환율 등 금융 관련 시계열

분산이 과거 자료에 의존하는 특성을 가정



통계학 분야의

조건부 분산



이분산 시계열 모형은

시간에 따른 **조건부 분산**을 **시간의 함수**로 표현하는 모형

수익률

수익률 (return)

$$\text{Simple return : } R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

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


- log return은 simple return 값에 **근사**
- log return은 덧셈으로 표현되며 log를 통해 **분산을 안정화**시키는 효과
- 클린업에서 다루는 모든 수익률은 **log return**을 의미 !

조건부 이분산성

조건부 등분산성

구조적인 변동성이 이전 기간 변동성의 영향을 받지 않는 특성



$$\text{Var}(r_t | \mathcal{F}_{t-1}) = \text{constant},$$

where \mathcal{F}_{t-1} is the σ - field generated by historical information



조건부 이분산성

조건부 이분산성

변동성이 시점에 의존하는 특성
즉, 미래의 변동이 현재까지의 상황에 의존 !

⋮

$$\text{Var}(r_t | \mathcal{F}_{t-1}) \neq \text{constant}$$

많은 시계열 자료가 조건부 이분산성을 갖기 때문에
수익률에 대해 모델링 하는 모형을 학습 !

ARCH 모형

ARCH (Auto-Regressive Conditional Heteroscedasticity)

t 시점의 변동성 σ_t^2 를
과거 시점의 오차항으로 설명하는 모델

 σ_t^2 은 조건부 분산

ARCH(1)

$$r_t = \sigma_t \varepsilon_t, \varepsilon_t \sim iid N(0,1)$$

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

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

$F_{t-1} : t-1$ 시점까지의 모든 정보의 집합
 $r_t | F_{t-1} : t-1$ 시점까지 모든 정보를 알고 있을 경우 t 시점의 수익률

ARCH 모형

ARCH (Auto-Regressive Conditional Heteroscedasticity)

t 시점의 변동성 σ_t^2 를
과거 시점의 오차항으로 설명하는 모델

 σ_t^2 은 조건부 분산

ARCH(1)

$$r_t = \sigma_t \varepsilon_t, \varepsilon_t \sim iid N(0,1)$$

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

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

σ_t^2 이 r_{t-1}^2 에 의존하기 때문에 조건부 이분산

ARCH 모형 | 비선형성

ARCH(m) 모형

$$r_t = \sigma_t \varepsilon_t$$

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

ARCH 모형은 σ_t^2 에 AR(m)을 가정

ARCH(m) 모형의 가장 큰 특징인 비선형성을

ARCH(1)을 통해 알아보자!

ARCH 모형 | 비선형성

$$\begin{aligned}
 r_t^2 &= \sigma_t^2 \varepsilon_t^2 = (\alpha_0 + \alpha_1 r_{t-1}^2) \varepsilon_t^2 \\
 &= \alpha_0 \varepsilon_t^2 + \alpha_1 \varepsilon_t^2 \{(\alpha_0 + \alpha_1 r_{t-2}^2) \varepsilon_{t-1}^2\} \\
 &= \alpha_0 \varepsilon_t^2 + \alpha_0 \alpha_1 \varepsilon_t^2 \varepsilon_{t-1}^2 + \alpha_1^2 r_{t-2}^2 \varepsilon_t^2 \varepsilon_{t-1}^2
 \end{aligned}$$

$$\vdots$$

$$= \alpha_0 \sum_{j=0}^n (\alpha_1^j \varepsilon_t^2 \varepsilon_{t-1}^2 \cdots \varepsilon_{t-j}^2) + \alpha_1^{n+1} r_{t-n-1}^2 \varepsilon_t^2 \varepsilon_{t-1}^2 \cdots \varepsilon_{t-j}^2$$

α_1 이 0과 1 사이에 있을 때 정상성 만족

$$\vdots$$

ARCH 모형은 곱셈으로 표현된 **비선형 모형!**

4

이분산 시계열 모형

ARCH 모형 | 비선형성



ARCH(m) 모형의 한계

$$r_t^2 = \sigma_t^2 \varepsilon_t^2 = (\alpha_0 + \alpha_1 r_{t-1}^2) \varepsilon_t^2$$

$$= \alpha_0 \varepsilon_t^2 + \alpha_1 \varepsilon_t^2 \{(\alpha_0 + \alpha_1 r_{t-2}^2) \varepsilon_{t-1}^2\}$$

ARCH(m) 모형의 m이 커지면

$$= \alpha_0 \varepsilon_t^2 + \alpha_1 \alpha_0 \varepsilon_{t-1}^2 \varepsilon_{t-1}^2 + \alpha_1^2 r_{t-2}^2 \varepsilon_t^2 \varepsilon_{t-1}^2$$

추정해야 할 **모수가 많아지고**, 추정량의 **정확도가 떨어짐**

$$= \alpha_0 \sum_{j=0}^n (\alpha_1^j \varepsilon_t^2 \varepsilon_{t-1}^2 \cdots \varepsilon_{t-j}^2) + \alpha_1^{n+1} r_{t-n-1}^2 \varepsilon_t^2 \varepsilon_{t-1}^2 \cdots \varepsilon_{t-j}^2$$



α_1 이 0과 1 사이에 있을 때 정상성 만족

나 불렀어?

ARCH 보다 확장된 **GARCH 모형** 사용 !

ARCH 모형은 곱셈으로 표현된 비선형 모형 !



GARCH 모형

GARCH (Generalized Auto-Regressive Conditional Heteroscedasticity)

σ_t^2 에 ARMA 모형을 가정해 t시점 수익률의 변동성을 표현한 모형

ARCH 모형에서 σ_t^2 에 AR 모형을 가정했던 것처럼

GARCH 모형에서는 σ_t^2 에 **ARMA 모형을 가정**하여

ARCH를 확장시킨 모델로 이해 !

GARCH 모형

GARCH(1,1) 모형

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

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

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

r_t^2 은 σ_t^2 에 근사,

$\eta_t = r_t^2 - \sigma_t^2$ (둘 사이의 오차)

일반화

GARCH 모형

일반화

GARCH(p,q) 모형

$$r_t = \sigma_t \varepsilon_t \sim N(0, \sigma_t^2)$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i r_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

너의 지식을
GARCH !



5

ARMAX

ARMAX

ARMAX (ARMA + eXogenous)

$$\phi(B)Y_t = \theta(B)Z_t + \beta_0 + \beta_1 X_t$$

Y_t 와 X_t 의 관측값 수는 동일해야 함

⋮

ARMA 모형에 독립변수로 **외부요인**(exogenous)을 추가한 모형
추가된 독립변수는 연속형일 수도, 범주형일 수도 있음

ARMAX

ARMAX (ARMA + eXogenous)

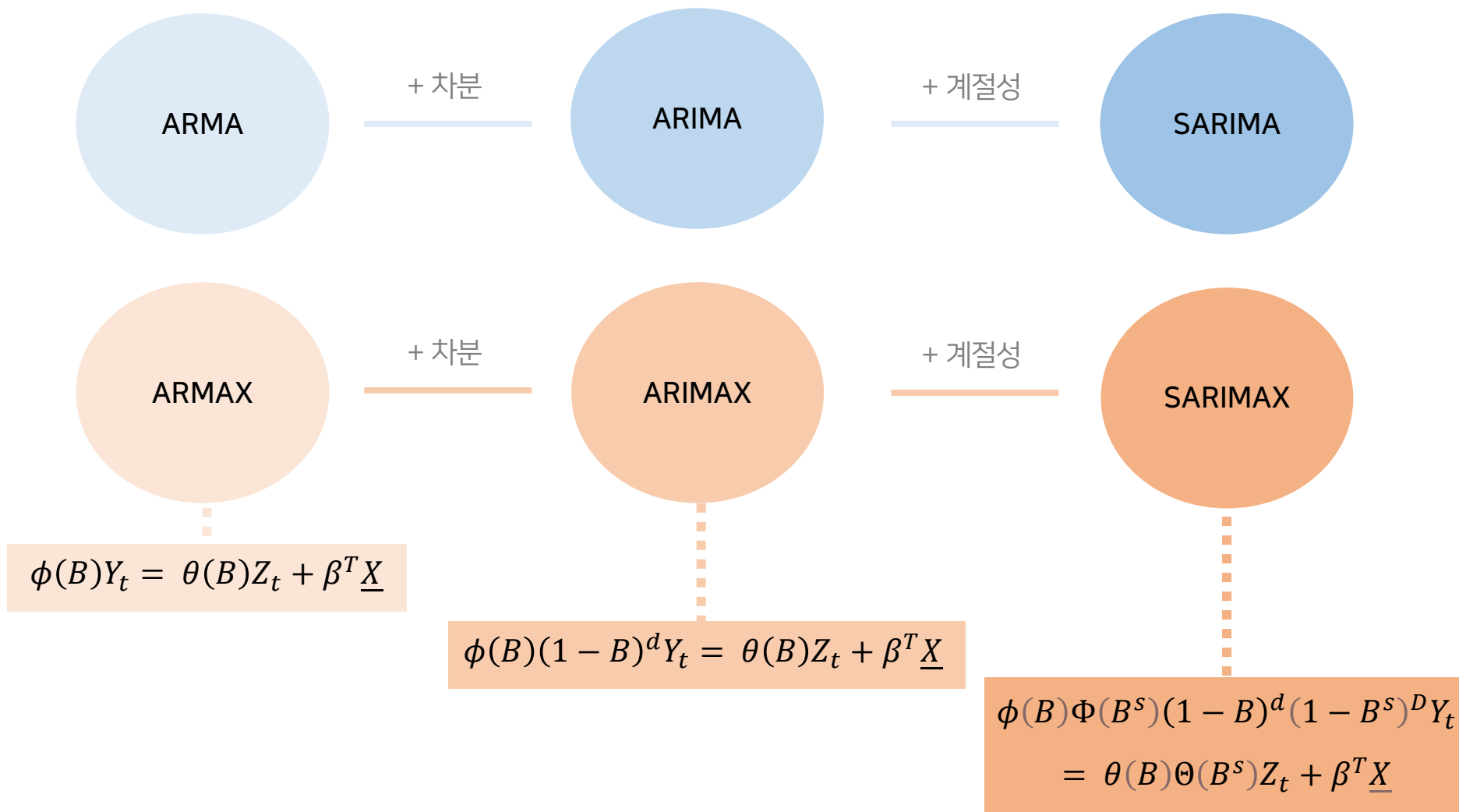
$$\phi(B)Y_t = \theta(B)Z_t + \beta_0 + \beta_1 X_t$$

Y_t 와 X_t 의 관측값 수는 동일해야 함

Y_t 를 예측하고자 하는 주가, X_t 를 날씨 데이터로 예를 들면,
날씨를 반영해 주가를 예측하고자 할 때 사용할 수 있는 모형식이 됨

ARMA 모형에 독립변수로 외부요인을 추가한 모형
추가된 독립변수는 연속형일 수도, 범주형일 수도 있음

ARMAX



6

VAR

VAR

VAR (Vector Auto Regressive)

AR 모형에 **Vector 구조를 결합**한 모형

AR

현재 관측값을
자기 자신의 과거 관측값들로 설명

VS

VAR

현재 관측값을
자기 자신의 과거 관측값들과
다른 변수의 과거 관측값으로 설명

VAR

VAR(1) 모형

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$



AR이 일변량 자기회귀모형이었다면,
VAR은 이를 확장한 **다변량** 자기회귀모형

VAR(1) 모형

$$\begin{aligned} X_t &= c_1 + \phi_{11}X_{t-1} + \phi_{12}Y_{t-1} + \varepsilon_1 \\ Y_t &= c_2 + \phi_{21}X_{t-1} + \phi_{22}Y_{t-1} + \varepsilon_2 \end{aligned}$$



다른 형태로 표현한 VAR(1) 모형

좌변에 들어가는 변수의 형태를 바꿔가며 식을 세워서 의존성, 상호작용을 설명할 수 있음!

VAR

VAR(1) 모형

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

AR이 일변량 자기회귀모형이었다면,
VAR은 이를 확장한 **다변량** 자기회귀모형

VAR(1) 모형

$$\begin{aligned} X_t &= c_1 + \phi_{11}X_{t-1} + \phi_{12}Y_{t-1} + \varepsilon_1 \\ Y_t &= c_2 + \phi_{21}X_{t-1} + \phi_{22}Y_{t-1} + \varepsilon_2 \end{aligned}$$

다른 형태로 표현한 VAR(1) 모형

좌변에 들어가는 변수의 형태를 바꿔가며 식을 세워서 의존성, 상호작용을 설명할 수 있음!

VAR

VAR(1) 모형

이를 통해 VAR 모형은 단순 과거 데이터만을 고려하는 것이 아니라
여러 변수들의 **의존성과 상호작용**을 고려하는 모형임을 알 수 있음

AR이 일변량 자기회귀모형이었다면,
VAR은 이를 확장한 **다변량** 자기회귀모형

VAR(1) 모형

상호연계성이 높은 경제구조를 분석할 때 용이하게 사용됨

$$X_t = c_1 + \phi_{11}X_{t-1} + \phi_{12}Y_{t-1} + \varepsilon_1$$

$$Y_t = c_2 + \phi_{21}X_{t-1} + \phi_{22}Y_{t-1} + \varepsilon_2$$

다른 형태로 표현한 VAR(1) 모형

좌변에 들어가는 변수의 형태를 바꿔가며 식을 세워서 의존성, 상호작용을 설명할 수 있음

VAR의 쓰임에 대해
자세히 알아보자!



모형의 활용

① 인과관계 분석

2개 이상의 변수가 존재할 때 변수 사이에 어떤 관계가 있는지,
특히 **시차 관계가 있는지 파악**하기 위한 분석



모형의 활용

② 충격반응 분석

여러 변수 중 특정 변수에 **예상치 못한 충격(Shock)**이 발생했을 때,
각 변수가 어떻게 변화하는지 파악하기 위한 분석



모형의 활용

③ 예측오차 분산분해

한 변수의 변화를 설명할 때

어떤 변수가 더 중요하게 작용하는지 **상대적 중요성**을 찾는 분석

다 와 간다
힘내자!



7

시계열과 ML

시계열과 ML

시계열
데이터

오차의 독립성 조건 만족 **X**

특별한
분석법 必



ARMA, SARIMA, ARIMAX,.. 등등

머신러닝을 사용할 수 있을까?

머신러닝
사용 가능 !



시계열과 ML

시계열
데이터

오차의 독립성 조건 만족

특별한
분석법 必

머신러닝을 사용할 수 있지만,
ARMA, SARIMA, ARIMAX,... 등등
시계열 데이터를 위한 전처리 방법들이 따로 있음!

지금부터 알아보자!

머신러닝을 사용할 수 없을까?

머신러닝
사용할 수
있음!

시계열 데이터의 전처리 | 결측치 보간

LOCF (Last observation carried forward)

직전 관측치 값으로 결측치 대체

NOCB (Next observation carried backward)

직후 관측치 값으로 결측치 대체

Moving Average / Moving Median

직전 N의 time window의 평균치 / 중앙값으로 대체

시계열 데이터의 전처리 | 결측치 보간

LOCF (Last observation carried forward)

직전 관측치 값으로 결측치 대체

일반적으로는 이 3가지 방법을 사용하지만,

NOCB (Next observation carried backward)

결측치를 기준으로 패턴이 급격하게 변화하는 경우에는

조금 더 복잡한 방법을 고려해야 함

Moving Average / Moving Median

직전 N의 time window의 평균치 / 중앙값으로 대체



시계열 데이터의 전처리 | 결측치 보간

선형 보간법

근사 함수가 선형(linear) 함수임을 가정

비선형 보간법

근사 함수가 비선형(non - linear) 함수임을 가정

스플라인(Spline) 보간법

전체 구간을 근사하는 것이 아닌, 소구간으로 분할하여 보간
각 구간마다 함수를 적합한 후 모든 구간에서 함수가 매끄럽게 이어지도록 함

시계열 데이터의 전처리 | 결측치 보간

선형 보간법



근사 함수가 선형(linear) 함수임을 가정

이 방법들로 보간하기 어려울 정도로 결측치가 많은 경우에는

비선형 보간법 **결측치 보간 모델링**을 통한 방법을 사용할 수 있음

근사 함수가 비선형(non-linear) 함수임을 가정

하지만 모델링을 진행하기 위해서는 **충분한 데이터**가 있어야 함!

스플라인(Spline) 보간법

전체 구간을 근사하는 것이 아닌, 소구간으로 분할하여 보간

각 구간마다 함수를 적합한 후 모든 구간에서 함수가 매끄럽게 이어지도록 함

시계열 데이터의 전처리 | 노이즈 처리

노이즈

다른 외부 요인의 간섭과 같이
데이터에 의도하지 않은 왜곡을 불러오는 모든 것

시계열 데이터의 시간의 흐름에 따라 변화하는 통계적 특성으로 인해
노이즈 역시 같이 기록될 가능성이 높음



시계열 데이터 특성에 맞는 노이즈 처리 방법을 사용해야 함!

시계열 데이터의 전처리 | 노이즈 처리

Moving
Average

평균값으로 관측치를 대체하여 평활화 하는 방법
노이즈가 적은 데이터에 적용할 수 있음

노이즈가 많은 경우 평균 역시 노이즈의 성격을 띠 수 있기 때문

Filtering

노이즈가 특정 분포를 따른다고 가정하고 해당 분포 값을 제거하는 방법

가우시안 필터링
(Gaussian Filtering)

노이즈가 정규분포를 따른다고 가정
중심에 가까울수록 큰 가중치를 부여

칼만 필터
(Kalman Filter)

잡음이 포함된 과거 측정값에서 현재 상태
의 결합분포를 추정하는 알고리즘데이터
의 특성에 맞는 분포를 모델링하는 것이 장
점

시계열 데이터의 전처리 | 노이즈 처리

Moving
Average

평균값으로 관측치를 대체하여 평활화 하는 방법
노이즈가 적은 데이터에 적용할 수 있음

노이즈가 많은 경우 평균 역시 노이즈의 성격을 띠 수 있기 때문

Filtering

노이즈가 특정 분포를 따른다고 가정하고 해당 분포 값을 제거하는 방법

가우시안 필터링
(Gaussian Filtering)

노이즈가 정규분포를 따른다고 가정,
중심에 가까울수록 큰 가중치를 부여

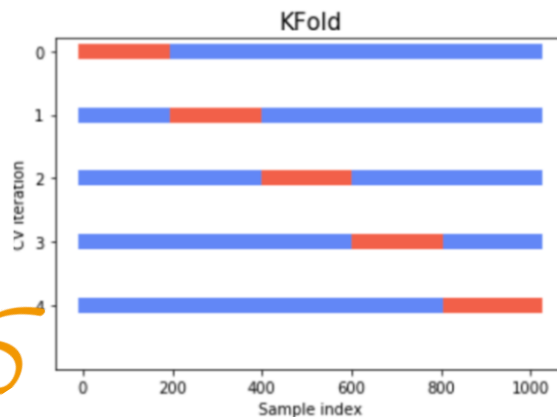
칼만 필터
(Kalman Filter)

잡음이 포함된 과거 측정값에서
현재 상태의 결합분포를 추정
→ 데이터의 특성에 맞는 분포를 모델링 가능

시계열 데이터의 CV

교차검증 (Cross Validation)

학습 데이터와 검증 데이터를 통해 모델의 성능을 측정하는 방법



데마팀 클린업 1주차 참고 !

일반적으로 사용하는 K-fold CV는 시간 순서를 고려하지 않고 진행

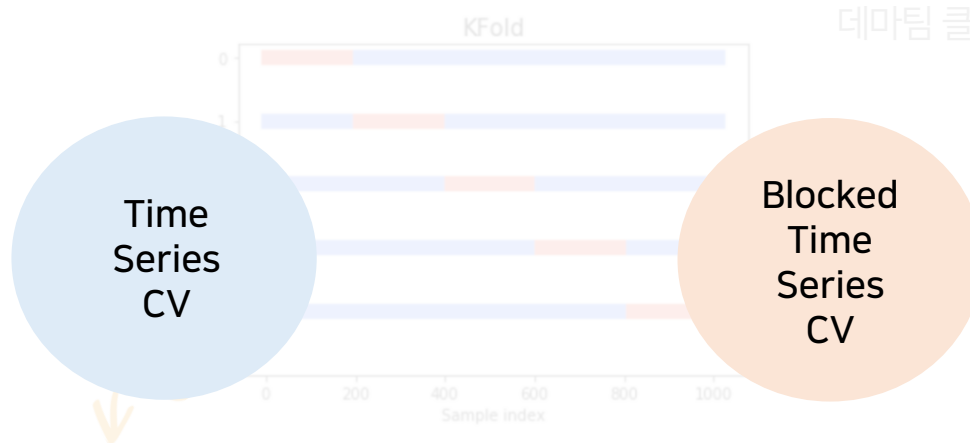
⋮

시계열 데이터는 시간에 따라 관측된 값이기에 **CV 역시 시간을 반영**해야 함 !

시계열 데이터의 CV

교차검증 (Cross Validation)

학습 데이터와 검증 데이터를 통해 모델의 성능을 측정하는 방법

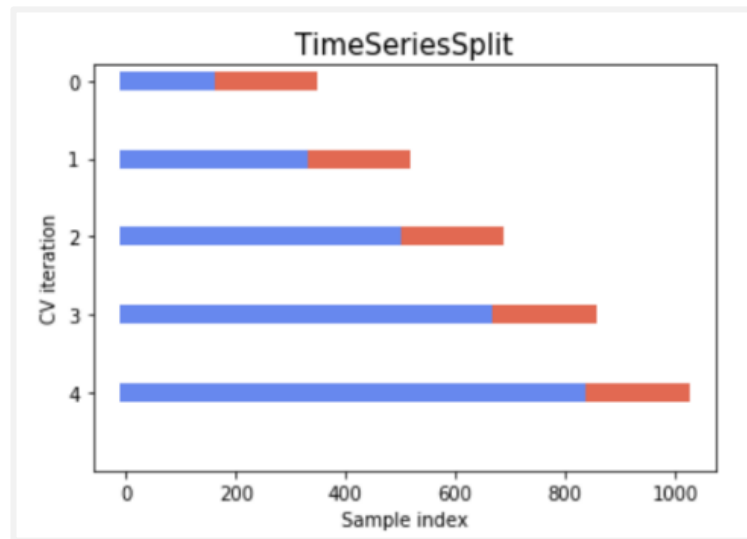


일반적으로 사용하는 **에 대해 한번 알아보시다!** 고려하지 않고 진행

시계열 데이터는 시간에 따라 관측된 값이기에 **CV 역시 시간을 반영해야 함!**

시계열 데이터의 CV | Time Series CV

Time Series CV

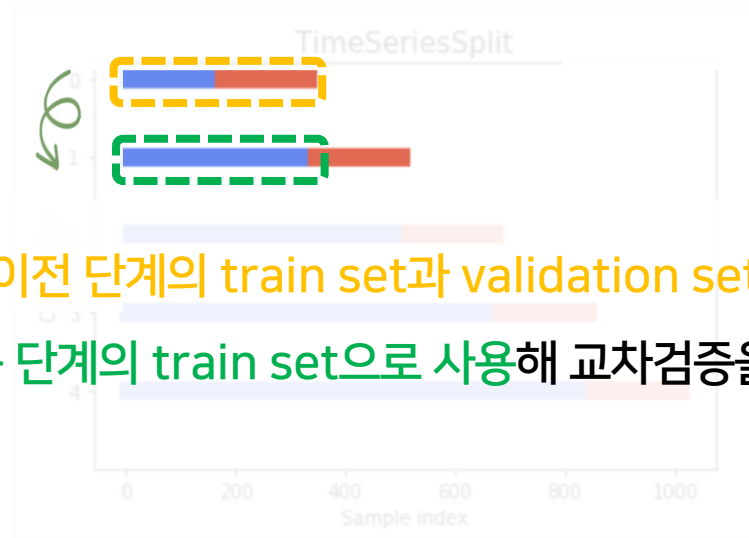


Expanding window CV라고도 부름

Expanding window란 window를 누적하여 이동한다는 의미

시계열 데이터의 CV | Time Series CV

Time Series CV



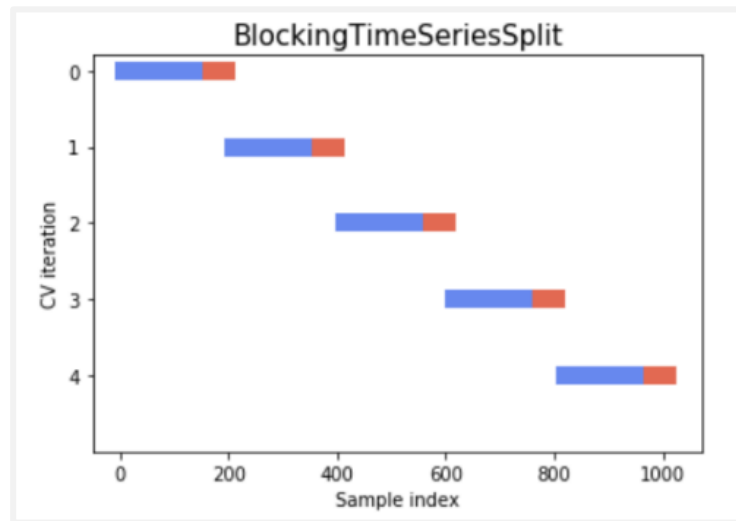
이전 단계의 train set과 validation set을
다음 단계의 train set으로 사용해 교차검증을 진행

Expanding window CV라고도 부름.

Expanding window란 window를 누적하여 이동한다는 의미입니다.

시계열 데이터의 CV | Blocked Time Series CV

Blocked Time Series CV



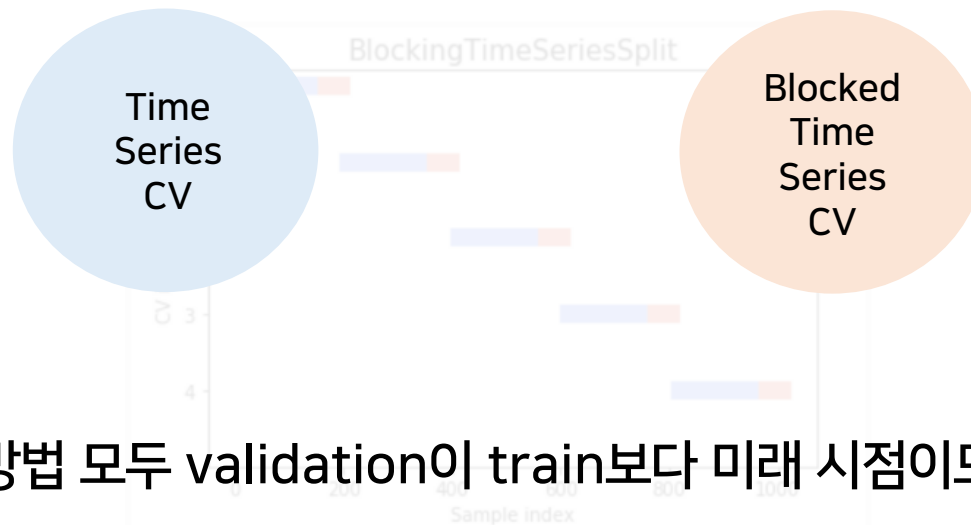
동일한 사이즈의 window를 옆으로 이동시킨다는 의미

Rolling window CV라고도 부름

같은 사이즈의 window 내에서 일정 비율로 train과 validation을 분할해 교차검증 진행

시계열 데이터의 CV

Blocked Time Series CV



두 방법 모두 validation이 train보다 미래 시점이므로,

시간의 흐름을 반영할 수 있는 교차검증 방법임!

Rolling window CV라고도 부름.

Rolling window란 동일한 사이즈의 window를 옆으로 이동시킨다는 의미
같은 사이즈의 window 내에서 일정 비율로 train과 validation을 분할해 교차검증 진행

클래스 불균형



클래스 불균형 문제



샘플링을 통해 해결



클래스 불균형 문제

.....

더 적은 수를 가지는 클래스에
가중치를 부여하는 방법 사용



샘플링



시간의 흐름 반영 X
문제 발생

NOPE



클래스 불균형



클래스 불균형 문제



샘플링을 통해 해결



클래스 불균형 문제



샘플링



시간의 흐름 반영X

문제 발생

더 적은 수를 가지는 클래스에
가중치를 부여하는 방법 사용



클래스 불균형 | 파라미터

Scale_pos_weight

이진분류 문제에서 사용하는 파라미터

Class_weight

샘플 수가 더 적은 쪽에 가중치 부여하는 방법

Ex) `model.fit(class_weight = {0:1, 1:2})` 와 같이 클래스 별 가중치 제시

Sample_weight

다중 분류에서 사용, 각 클래스 별 비율의 역수를 가중치로 계산하는 함수

Ex) `class_weight.compute_sample_weight(class_weight = "balanced")`

감사합니다

