Time series analysis

Contents

I. Review

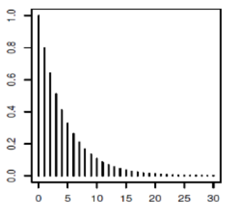
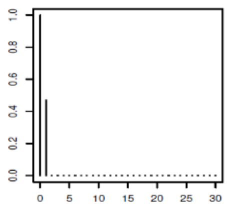
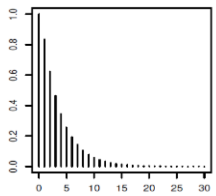
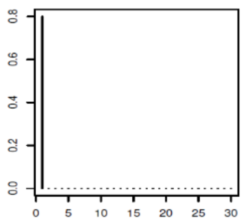
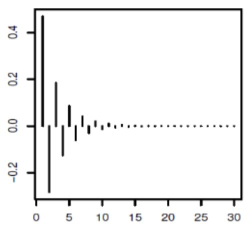
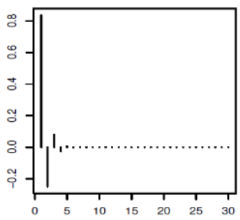
2 주차에는 정상시계열을 가정하는 시계열 모델 AR, MA, ARMA 에 대해 배웠습니다. 지난 주차에 배운 몇 가지 내용을 복습해보겠습니다.

1. **선형과정(Linear process)**

   선형과정이란 백색잡음들의 선형결합을 의미하였으며, 여기서 중요한 특징은 주어진 정상 확률 과정의 선형 결합은 또 다시 정상 확률 과정이라는 것입니다.

2. AR, MA, ARMA

   AR, MA, ARMA 는 ACF, PACF 를 통해 식별할 수 있음을 배웠고 각각의 ACF 와 PACF 는 아래와 같은 패턴을 보인다는 사실을 배웠습니다.

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

각 모형은 아래의 조건을 만족해야 합니다.

	AR(p)	MA(q)	ARMA(p,q)
정상성	조건필요	자체만족	조건필요
가역성	자체만족	조건필요	조건필요

II. ARIMA

1. 정의

이번주에 배울 모형은 비정상시계열을 다루는 모형들입니다. ARIMA 모델은 차분과 ARMA 가 결합된 모델이라고 생각하시면 됩니다. d차 차분된 $Y_t = (1-B)^d Z_t$ 가 정상과정 ARMA(p,q)를 따를 때 $\{X_t\}$ 는 **자기회귀누적이동평균과정 ARIMA(p,d,q)**을 따른다고 합니다. 다음과 같이 표현할 수 있습니다.

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

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

⇨ 따라서 ARIMA 모델은 주어진 데이터가 다항식의 추세를 갖고 있고, 그 오차가 ARMA(p,q)를 따를 때 사용하는 모델입니다. ARIMA 모델의 장점은 차분까지 포함하여 간결하게 표현 가능하다는 점과 여전히 선형과정이라는 점입니다.

ARIMA(1,0,1) =

ARIMA(3,0,2) =

ARIMA(0,0,1) =

ARIMA(0,0,2) =

여기서 잠깐!

ARIMA의 I는 무엇일까??

ARIMA는 자기회귀누적이동평균과정이라고 하였습니다. 여기서 I는 누적(Integration)을 의미한다는 점 눈치채셨나요? 그렇다면, 차분은 분명히 differencing인데, 왜 누적이라는 말을 사용하는 것일까요? 아래의 식을 보면 그 이유를 알 수 있습니다.

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

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

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

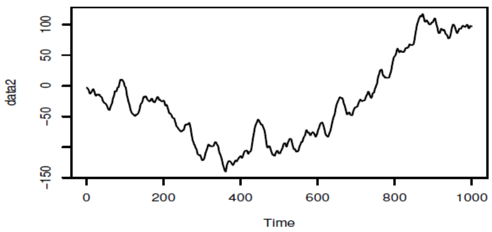
이와 같이 X_t의 관점에서 Y_t의 누적합이 되어 integration이라는 단어를 사용합니다.

2. 적합절차

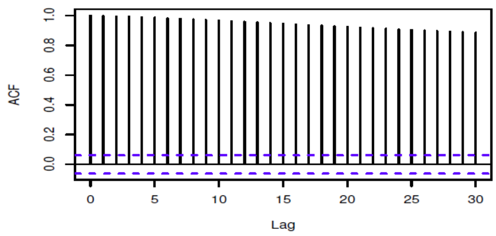
1) 시계열 플랏과 ACF 그래프를 통해 정상시계열인지 비정상시계열인지 판단합니다.

만약, 추세가 존재한다면 다음과 같이 천천히 ACF 가 감소하는 형태입니다.

ARIMA(1,1,0)의 시계열플랏



ARIMA(1,1,0)의 ACF



2) 위와 같이 추세가 존재한다면 차분을 통해 정상화를 합니다.

⇨ 이때 차분의 차수 d가 1,2 를 넘어가면 과대차분의 위험이 있기에 주의합니다.

과대차분(overdifferencing)을 조심하세요

과대차분이란 말 그대로 차분을 과하게 하는 것입니다. 이미 정상화가 되어있음에도 불구하고 차분을 또 시도하는 경우, 정상시계열의 선형 결합은 다시 정상 시계열이 되어 "정상성"자체에는 문제가 없습니다. 그러나, 지나친 차분은 ACF를 복잡하게 만들거나 분산이 커지는 문제가 생깁니다.

⇨ 이 성질은 모형 선택시에도 유용하게 사용됩니다. (분산의 크기 비교를 통해 적절한 모형 선택)

↑1 주차 교안 참고

Ex) MA(1) $[X_t = (1-\theta B)Z_t], Z_t \sim WN(0,\sigma^2)$

$Var(X_t) = (1+\theta^2)\sigma^2$

1 차 차분 후 ⇨ $Y_t = (1-B)X_t = (1-B)(1-\theta B)Z_t = [1-(1+\theta B)B+\theta B^2]Z_t$

$Var(Y_t) = 2(1+\theta+\theta^2)\sigma^2$

3) 모형의 모수를 추정합니다.

4) 적절한 모형인지 판단합니다.

5) 최종 모형을 선택합니다.

확률적(stochastic) vs 결정적(deterministic)

결정적이라는 단어의 의미는 랜덤에 반대된다고 생각하시면 됩니다. 즉, 예를 들어 결정적 추세의 경우 $\beta_0 + \beta_1 t$에 같은 t를 넣으면 항상 동일한 추세가 반환되며 이러한 추세가 지속됩니다. 반면 확률적 추세의 경우 인접한 자료 간 확률적 구조에 의해 생기는 경우로 인접한 시계열 간 강한 상관관계에 의하여 추세가 있는 것처럼 보이는 경우입니다.

계절성도 마찬가지입니다. 결정적 계절성은 우리가 1 주차에 배운 것과 같이 $a_0 + \sum_{j=1}^{k}(a_j\cos(\lambda_j t) + b_j \sin(\lambda_j t))$로 추정할 수 있으며 확률적 계절성은 확률적 구조로 인한 계절성으로 아래에서 배울 SARIMA 모형을 통해 추정 가능합니다.

III. SARIMA

1주차에 우리는 전통적 분해법에 의하여 계절성이 존재하는 시계열을 $Y_t = s_t + Z_t$과 같이 표현한 뒤 계절성분 s_t를 추정한 뒤 제거할 수 있음을 배웠습니다. 이때의 계절성분은 결정적(deterministic) 계절성분입니다. 그러나 항상 계절성분이 결정적인 것은 아니며 우리가 접하는 시계열들의 구성성분은 확률적이거나 다른 성분들과 상관이 있는 경우가 많습니다. 이러한 경우에 확률적 분석방법인 ARIMA를 활용할 수 있습니다.

1. IDEA

계절형 ARIMA모형 SARIMA는 확률적 계절성을 고려하는 모형입니다.

	month 1	month 2	⋯	month 12
Year 1	Y_1	Y_2	⋯	Y_{12}
Year 2	Y_{13}	Y_{14}	⋯	Y_{24}
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	⋯	$Y_{12+12(r-1)}$

다음과 같이 주기 $s = 12$인 시계열을 정리해보면, column을 고정했을 때 이 시계열들이 ARMA(P,Q)를 따른다고 생각해보겠습니다. 그때, 다음과 같이 표현할 수 있습니다. 여기서 $U_t \sim WN(0, \sigma_U^2)$ 입니다.

$$Y_{j+12t} - \Phi_1 Y_{j+12(t-1)} - \cdots - \Phi_P Y_{j+12(t-P)} = U_{j+12t} + \Theta_1 U_{j+12(t-1)} + \cdots + \Theta_Q U_{j+12(t-Q)}, t = 0,1, \ldots ,11$$

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

위 식이 어떤 의미인지 자세히 살펴보면, month가 달라져도 파라미터들은 같습니다. 위의 오차항 U_t를 백색잡음이 아닌 ARMA(p,q)를 따르는 시계열이라고 생각해봅시다. Month가 달라져도 이들 간 correlation이 있을 수 있잖아요! 다음과 같이 표현할 수 있겠죠?

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

앞선 두 과정을 합쳐서 표현하면 다음과 같습니다.

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

$$\phi(B)\Phi(B^{12})Y_t = \theta(B)\Theta(B^{12})Z_t, Z_t \sim WN(0, \sigma^2)$$

뿐만 아니라, 추세가 존재할 경우 d 차 차분까지 포함할 수 있습니다.

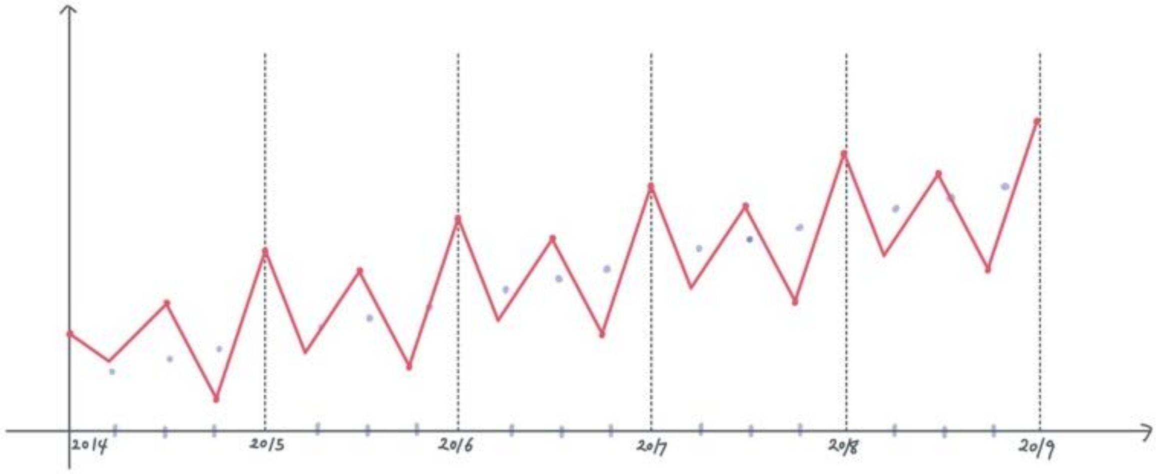
$$Y_t = (1-B)^d(1-B^{12})^D X_t$$

따라서 SARIMA(p,d,q)x(P,D,Q)모델은 다음과 같이 표현할 수 있습니다.

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

정리해보자면 SARIMA는 추세와 계절성을 동시에 포함할 수 있는 모형이며 SARIMA모형에는 순수SARIMA와 승법SARIMA 두 가지가 존재합니다. 각각의 경우를 더 자세하게 살펴보겠습니다. 방금 설명한 모형은 승법 SARIMA에 해당합니다.

예시) https://www.youtube.com/watch?v=WjeGUs6mzXg



ARIMA(1 , , 1)(1 , , 1)$_4$

2. 순수SARIMA

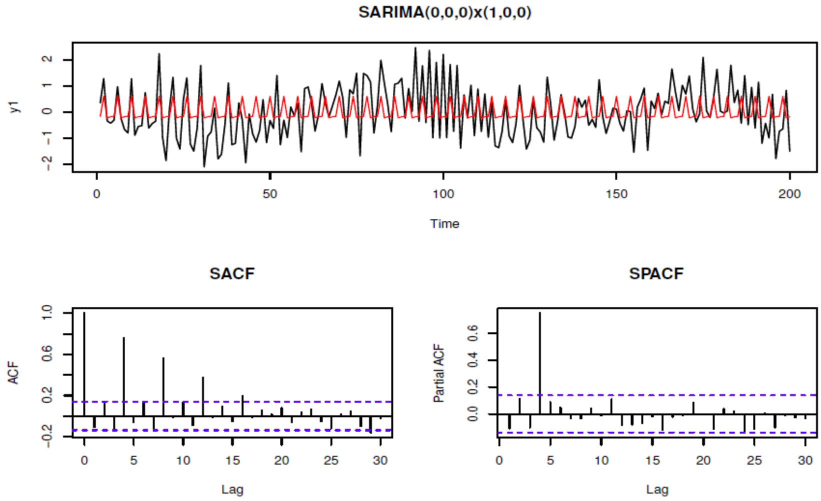
순수SARIMA 모델은 계절성만을 고려하는 모델입니다. (위의 표에서 row간 연관성을 고려하지 않는 모델이라고 생각하시면 됩니다.) D번의 계절차분과 과거 주기의 P개의 관측치 및 과거 주기의 Q개의 오차항으로 현재 관측치를 설명하는 모형입니다. 즉, $X_{t-s}, X_{t-2s}, \dots, X_{t-Ps}$와 $Z_{t-s}, Z_{t-2s}, \dots, Z_{t-Qs}$로 현재의 관측치 X_t를 설명하는 모델입니다. s는 계절의 주기라고 할 때 다음과 같이 표현할 수 있습니다.

ARIMA(0,0,0)(P,D,Q)$_s$모형

$$\Phi(B^s)(1-B^s)^D X_t = \Theta(B^s)Z_t$$
$$\Phi(B^s) = (1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$$
$$\Theta(B^s) = (1+\Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs})$$
$$Z_t \sim WN(0,\sigma^2)$$

ACF, PACF

$$(1 - .8B^4)X_t = Z_t.$$



위의 그림처럼 순수 SARIMA모형에서는 s에 해당되는 계절 주기에서만 0이 아니고 다른 시차에서는 0이 됩니다. 그 외의 패턴은 지난시간에 배운 ARMA와 똑같습니다. SACF는 감소하는 모양이고 SPACF는 첫번째 주기 이후 절단된 모형입니다. 따라서 위 SACF와 SPACF를 통해 SAR(1,0,0)모형임을 알 수 있습니다.

순수 SARIMA모형은 비계절 요소에 대한 부분은 고려하지 않기에 사용이 제한적입니다. 따라서 비계절 요소를 고려할 때 아래의 승법 SARIMA모형을 사용할 수 있습니다.

3. 승법SARIMA

승법 SARIMA모형은 계절 이외의 요소의 연관성도 고려하는 모형입니다. “승법”이라는 단어는 ARIMA × 순수 SARIMA의 형태이기 때문입니다. 위 순수SARIMA의 경우 오차는 백색잡음이었지만 승법 SARIMA의 경우 오차가 ARIMA를 따릅니다.

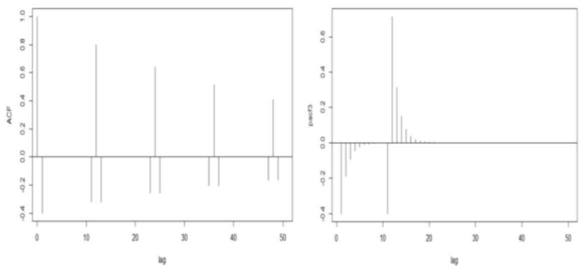
ARIMA(p,d,q)(P,D,Q)$_s$모형

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)Z_t$$
$$\Phi(B^s) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$$
$$\Theta(B^s) = \left(1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}\right)$$
$$\phi(B) = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)$$
$$\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)$$
$$Z_t \sim WN(0, \sigma^2)$$

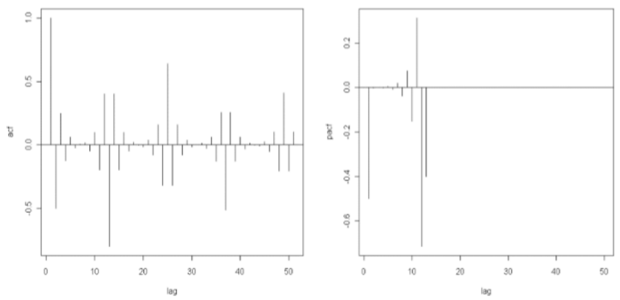
모수절약의 원칙에 따라 순수 SARIMA보다 더 자주 사용됩니다.

ACF, PACF

승법 SARIMA의 경우 비계절성 요소에 관하여 한 주기 내의 ACF와 PACF의 패턴을 통해 파악할 수 있으며 계절적 요소의 경우는 순수 ARIMA와 동일합니다.



ARIMA(0,0,1) × $(1,0,0)_{12}$



ARIMA(1,0,0) × $(1,0,0)_{12}$

IV. 이분산 시계열모형

우리가 지금까지 배운 시계열 모형들은 주로 평균 부분의 움직임에 관심을 갖는 모형이었습니다. 전통적인 시계열 자료의 분석에서는 시간에 따른 분산의 변화가 없는 모형을 생각합니다. 오차항이 $\{\varepsilon_t\} \sim IID(0, \sigma^2)$인 AR(1)모형의 조건부 평균과 분산은 다음과 같이 시간에 따라 변화가 없습니다.

$$y_t = \phi_1 y_{t-1} + \varepsilon_t$$

$$E(y_t | y_{t-1}) = \phi y_{t-1} \quad Var(y_t | y_{t-1}) = \sigma^2$$

이분산(heteroskedasticity) **시계열 모형**이란 조건부 분산을 시간의 함수로 표현하는 시계열 모형입니다. 그러나 주식 수익률, 이자율, 환율 등과 같은 경제 시계열 자료는 시간의 추이에 따라 변동성이 변하는 성질이 있습니다. 경제학 분야에서는 일반적으로 분산 대신 위험을 측정하는 수단으로서 **변동성**(volatility)이란 용어를 사용하며 이는 **조건부 분산**(conditional variance)입니다. 이는 강한 변동성이란 미래 값의 분산(변화가능의 폭)이 현재의 상황에 크게 의존하면서 결정될 때만 나타나는 현상이기 때문입니다. 변동성이 강한 자료에서 가장 특징적인 사실은 자료의 급락 또는 급상승이 빈번하게 반복된다는 사실입니다.

이분산성(heteroscedasticity)

이분산성이란, 분산이 시간이나 관측치별로 독립적으로 나타나지 않거나 일정하지 않음을 의미합니다.

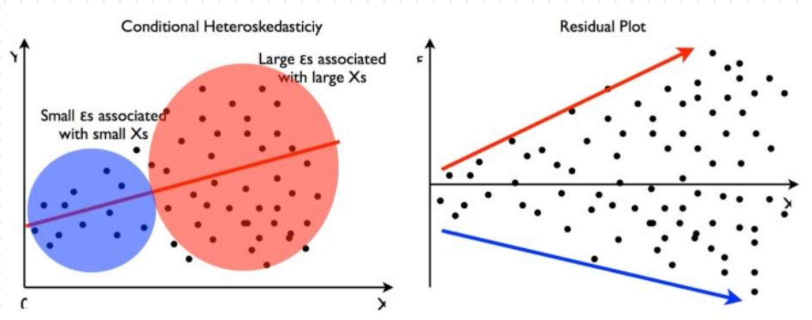
이분산성에는 조건부 이분산성과 비조건부 이분산성이 있습니다.

1) 비조건부 이분산성(unconditional heteroskedasticity)

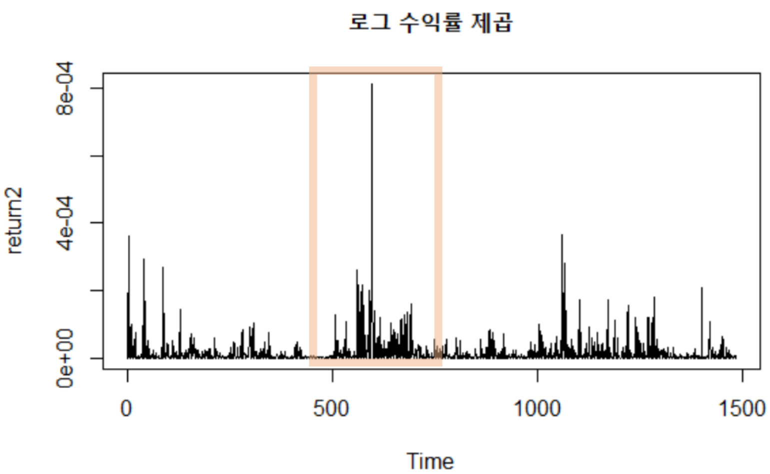
⇨ 일반적인 구조적 변동성의 변화가 이전 기간의 변동성과 관련이 없을 때

2) 조건부 이분산성(conditional heteroskedasticity)

⇨ 일정하지 않은 변동성이 이전 시점의 변동성과 연관이 있을 때



1. 변동성 집중(volatility clustering)

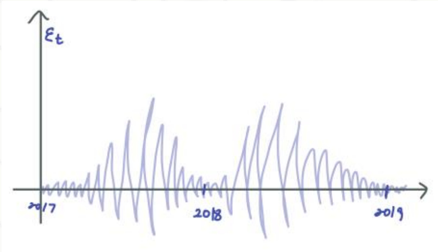


변동성 집중이란 위의 그림과 같이 한번 나타난 큰 변화가 당분간 계속 큰 변화를 유지하며, 작은 변화는 당분간 지속적으로 작은 변화를 유지하는 경향을 의미합니다. 위 그림에서 수익률(return)은 로그를 씌운 뒤 차분한 값으로 변동성 및 이분산성을 파악하고 분석하는데 유용한 표현입니다. 이러한 변동성 집중에 의한 변화 폭은 √나 log등의 자료 변환으로 상쇄되지 않으며, 오차의 제곱에 대해 적절한 시계열 모형을 적용하여 설명할 수 있습니다.

2. ARCH(Autoregressive conditional heteroscedasticity)

ARCH모델도 설명하기에 앞서 흐름에 대해 먼저 설명하겠습니다.

자, 우리는 지금 영화관 매표 기록에 대한 데이터를 모델에 적합시키려 한다고 생각해 봅시다. 먼저, 최적의 모델을 찾아 적합시킨 후 오차항만 남았다고 생각해보겠습니다. 오차항을 이용해 플랏을 그렸더니 다음과 같이 나왔습니다.



어라.. 뭔가 여기서 끝나면 안될 것 만 같은 느낌이지 않나요? 오차항이 변할 땐 크게 변했다가 또 조금 변할 땐 조금만 변하는, 우리가 앞서 봤던 변동성 집중과 비슷한 일이 일어난 것 같습니다! 따라서 오차항을 과거 시점의 오차항들로 설명해보려 합니다.

$Var(\varepsilon_t) = \sigma_t^2 = \alpha_0 + \alpha_1\sigma_{t-1}^2$ 변동성 (오차항의 분산)을 어제의 분산으로 다음과 같이 설명합니다. 이 식을 통해 오차는 다음과 같이 얻어질 수 있습니다. $\epsilon_t = w_t\sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}$,w_t: 백색잡음 ← 어제의 오차가 아주 컸다면 내일의 오차도 커지겠지요? 이제 양변에 제곱을 해보겠습니다 $\epsilon_t^2 = w_t^2(\alpha_0 + \alpha_1\varepsilon_{t-1}^2) = w_t^2\alpha_0 + w_t^2\alpha_1 + \varepsilon_{t-1}^2$ 자! 과거의 오차항을 이용하여 AR+type의 함수로 오늘의 오차항을 설명했습니다. 위와 같은 과정이 바로 ARCH(1)모델입니다.

자기회귀이분산모형(ARCH)은 로버트 엥글이라는 경제학자에 의해 처음 제시된 모형으로, 과거의 정보가 주어졌을 때 현재의 값에 대한 조건부분포를 표현하기 위해 도입된 모형으로 오차의 변동성이 자귀회귀적으로 변하는 것을 설명하는 비선형 모델입니다. 다음과 같이 표현할 수 있습니다.

🐻 표현

ARCH(p)

$$Z_t \sim iidN(0,1)$$
$$\varepsilon_t = \sigma_t Z_t$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i\varepsilon_{t-i}^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_p\varepsilon_{t-p}^2$$
$$\alpha_0 > 0, \alpha_j \geq 0, j = 1,2,\ldots,p$$

⇨ ARCH(p)모형은 t시점 오차항의 변동성을 p시점 전까지의 오차항의 제곱으로 설명합니다.

⇨ $E(\varepsilon_t^2|\varepsilon_{t-1}^2,\ldots) = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_p\varepsilon_{t-p}^2$: 시간의 함수로 변동성을 표현할 수 있습니다.

🐻 특징

⇨ 비선형적 모델

ARCH(1) : $\varepsilon_t = \sigma_t z_t$,$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$

$$\varepsilon_t^2 = \sigma_t^2 z_t^2 = (\alpha_o + \alpha_1 \varepsilon_{t-1}^2) z_t^2 = \left(\alpha_0 + \alpha_1 (\sigma_{t-1}^2 z_{t-1}^2)\right) z_t^2$$

$$= \left(\alpha_0 + \alpha_1 \left((\alpha_0 + \alpha_1 \varepsilon_{t-2}^2) z_{t-1}^2\right)\right) z_t^2$$

$$= \cdots = \alpha_0 \sum_{j=0}^{n} \alpha_1^j z_t^2 z_{t-1}^2 \dots z_{t-j}^2 + \alpha_1^{n+1} z_t^2 \dots z_{t-n}^2$$

🐻 추정과 검정

1) 추정 : 최대가능도추정방법(MLE)

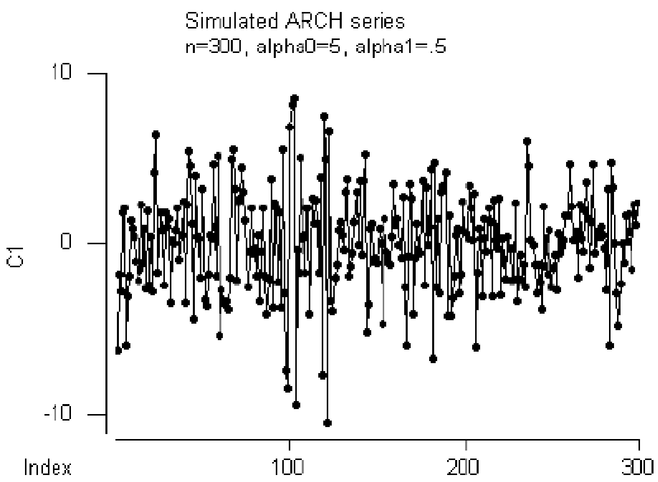
2) 검정 : H_0 : $\alpha_1 = \cdots = \alpha_p = 0$

   A. LM(Largrange multiplier): ε_t 가 $i.i.d\ N(0, \sigma^2)$을 따른다는 귀무가설 하에서 검정통계량 nR_ε^2 이 점근적으로 $\chi^2(p)$를 따른다는 사실을 활용하여 검정합니다.

      기각: ARCH효과 존재

   B. Ljung-Box Q 검정: ε_t^2이 AR(p)모형을 따름을 이용하여 ε_t^2의 자기상관계수가 유의한지 여부 판단하는 검정 방법입니다.

   C. 오차항의 정규성 검정: QQ-plot, Jarque-Bera test



3. GARCH(Genenralized Autoregressive conditional heteroskedasticity)

앞서 소개한 ARCH모형은 q값이 클 때 추정량의 정확도가 떨어지므로 추정값들이 ARCH모형의 정상성 조건을 만족시키지 못할 가능성이 더욱 커집니다. 이러한 이유로 모형의 도입 초기에는 작은 q값을 갖는 모형이나 ARCH 계수 α_i가 i가 증가함에 따라 기하급수적으로 감소하는 모형을 도입하였습니다. 그러나 이러한 모형이 실제 상황에는 잘 맞지 않는 경우가 대부분이었으므로 보다 일반적인 모형의 필요성이 대두되었습니다. 이러한 문제를 해결해주는 일반화된 모델이 **일반화자기이분산모형** GARCH입니다. AR모형을 ARMA모형으로 확장시키는 방법과 유사한 아이디어입니다.

🐻 표현

GARCH(p,q)

$$
\begin{gathered}
Z_t \sim iidN(0,1) \\
\varepsilon_t = \sigma_t Z_t \\
\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \\
= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \\
\alpha_0 > 0, \alpha_j \ge 0, \beta_j \ge 0
\end{gathered}
$$

⇨ GARCH(p,q)모형은 t 시점의 오차항의 변동성을 p시점 이전의 오차항들의 제곱과 q시점의 변동성으로 설명한 모형입니다.

🐻 특징

GARCH모형의 특징이자 장점은 앞서 소개한 것과 같이 모수의 절약입니다.

GARCH(1,1)모형은ARCH(∞)로 표현가능합니다.

$$
\begin{aligned}
\sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\
&= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta(\alpha_0 + \alpha_1 \varepsilon_{t-2}^2 + \beta \sigma_{t-2}^2) \\
&= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \alpha_0 + \beta \alpha_1 \varepsilon_{t-2}^2 + \beta^2 \sigma_{t-2}^2 \\
&= (1+\beta)\alpha_0 + \alpha_1(\varepsilon_{t-1}^2 + \beta \varepsilon_{t-2}^2) + \beta^2 \sigma_{t-2}^2 \\
&= (1+\beta)\alpha_0 + \alpha_1(\varepsilon_{t-1}^2 + \beta \varepsilon_{t-2}^2) + \beta^2(\alpha_0 + \alpha_1 \varepsilon_{t-3}^2 + \beta \sigma_{t-3}^2) \\
&= \alpha_0(1+\beta+\beta^2) + \alpha_1(\varepsilon_{t-1}^2 + \beta \varepsilon_{t-2}^2 + \beta^2 \varepsilon_{t-3}^2) + \beta^3 \sigma_{t-3}^2 \\
&\vdots \\
&= \alpha_0(1+\beta+\beta^2+\cdots) + \alpha_1(\varepsilon_{t-1}^2 + \beta \varepsilon_{t-2}^2 + \beta^2 \varepsilon_{t-3}^2 + \cdots) + \beta^\infty \sigma_0^2 \\
&= \frac{\alpha_0}{1-\beta} + \sum_{j=1}^{\infty} \alpha_1 \beta^{j-1} \varepsilon_{t-j}^2
\end{aligned}
$$

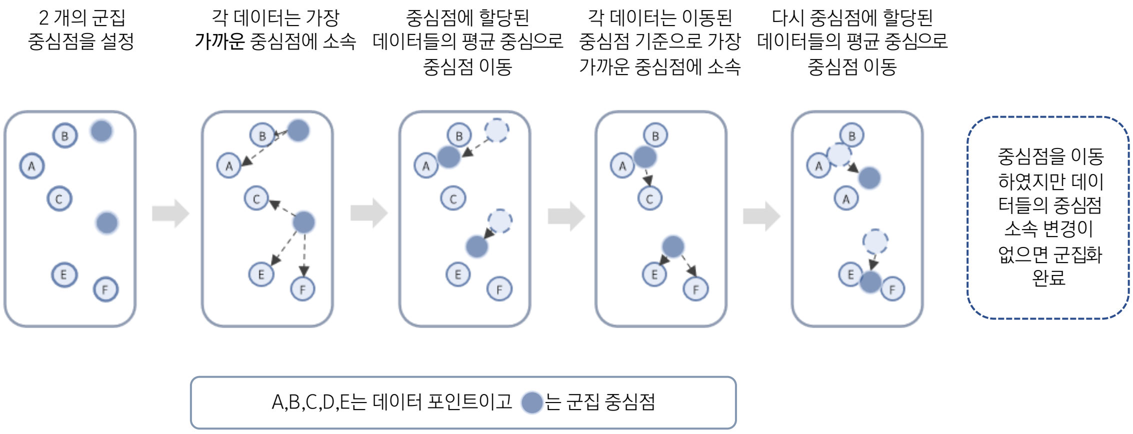
II. 시계열,,, 머신러닝과 만나다!

군집분석(Clustering)

군집분석이란 무엇일까요? 군집분석이 서로 다른 성질을 가진 구성요로들로 이루어진 집단에 대하여, 비슷한 패턴을 보이는 공통적인 특성을 지닌 구성요소들끼리 소집단(cluster)을 이루게 하여, 전체 집단을 이러한 소집단을 중심으로 분류하는 비지도 학습의 하나입니다. 즉 유사한 성격의 집단을 가까이 위치하고, 이질적인 집단은 멀리 위치하여 분석결과를 도출하는 방식입니다.

군집분석에도 여러 방법이 있는데요, 자료의 거리 또는 자료의 상관성을 토대로 한 유사도를 이용하여 dendrogram 형태의 군집형성을 수행하여 군집을 결정하는 계층적 군집분석 방법, 어떠한 개체가 어떠한 군집에 속하는가를 나타내는 비계층적 군집방법, 그리고 분석모형에 의하여 군집을 결정하는 모형 군집분석 방법 등으로 구분됩니다. https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/

클러스터링은 이번주 데이터마이닝에서 다뤄지는 주제이니 가장 간단한 k-means clustering 만 간략히 소개하고 넘어가겠습니다. K-means 클러스터링은 주변 k 개의 데이터들의 평균을 이용하는 군집분석입니다.



앞서 소개한 것과 같이 클러스터링 분석 시에는 거리 지표를 선택해야 합니다. 시계열 데이터는 시간을 인식하는 거리 측정법을 사용해야 하는데, 그중 가장 잘 알려진 **동적시간왜곡** DTW(dynamic time wraping)에 대해 알아보겠습니다.

동적시간왜곡 DTW

DTW 는 시간, 속도 또는 길이가 정확히 정렬되지 않는 두 시계열 간의 유사성을 측정하는 방법입니다. 이 알고리즘은 음성인식에서 처음 나온 방법으로 패턴 인식에서 이용되었고 시퀀스를 시간의 길이를 고려하지 않고 인식할 수 있는 방법입니다. 음성의 경우에는 두 음성의 빠르기를 생각하지 않고 패턴의 진행, 두 음성의 모양을 비교하는데 사용됩니다.

두 시계열 X 와 Y 가 있을 때, X 의 각 요소와 Y 의 가장 가까운 점 사이의 거리 제곱 합계의 제곱근으로 계산됩니다. 아래의 식을 통해 자세히 알아보겠습니다.

$$X = (x_0, \dots, x_n)\ , Y = (y_0, \dots, y_m)$$

$$DTW(x,y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_i)^2}$$

π 는 다음과 같은 조건을 만족합니다.

- $\pi_k = (i_k, j_k)\ with\ 0 \le i_k < n\ and\ 0 \le j_k < m$
- $\pi_0 = (0,0)\ and\ \pi_{K-1} = (n-1, m-1)$
- $for\ all\ k > 0, \pi_k = (i_k, j_k)\ is\ related\ to\ \pi_{k-1} = (i_{k-1}, j_{k-1}) as\ follows:$
  - $i_{k-1} \le i_k \le i_{k-1} + 1$
  - $j_{k-1} \le j_k \le j_{k-1} + 1$

위의 식이 어떤 과정으로 이루어지는지 더 구체적으로 알아볼께요!!

X = (1,2,3,3,2) Y = (1,1,2,3,3) 일때,

X \ Y	1	1	2	3	3
1	$\lvert 1-1\rvert = 0$	$0 + \lvert 1-1\rvert = 0$	$0 + \lvert 1-2 = 1\rvert$	$1 + \lvert 3-1\rvert = 3$	$3 + \lvert 1-3\rvert = 5$
2	$0 + \lvert 1-2\rvert = 1$	1	0	1	2
3	$1 + \lvert 1-3\rvert = 3$	3	1	0	0
3	$3 + \lvert 1-3\rvert = 5$	5	2	0	0
2	$5 + \lvert 1-2\rvert = 6$	6	2	1	1
1	$6 + \lvert 1-1\rvert = 6$	6	3	3	3

① 두 벡터의 길이를 행, 열로 가지는 행렬을 그립니다.

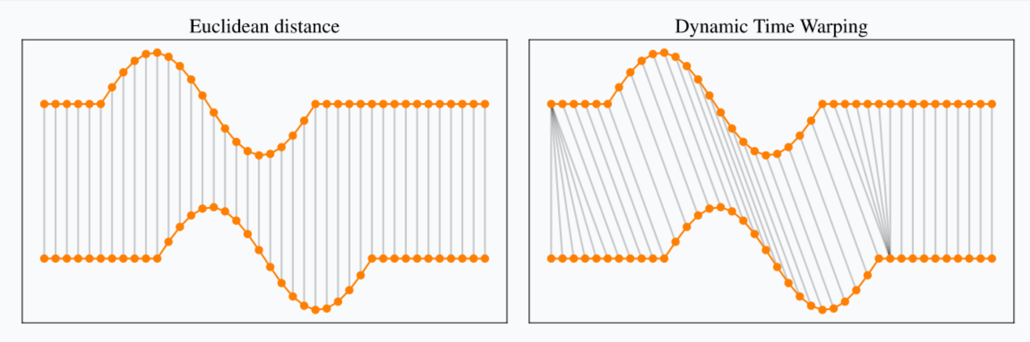
② 첫번째 열과 첫번째 행은 각각 첫번째 X와 Y, 첫번째 Y와 X의 유클리디안 거리를 누적합으로 구합니다.

③ i, j번째 값을 구할 때 아래의 예시와 같이 구합니다
예시) 현재 distance = d= |2-1| = 1
$\min\{d_{i-1,j} + d,\ d_{i,\,j-1} + d,\ d_{i-1,j-1} + 2d\} = \min\{0+1, 1+1, 0+2\} = 1$

DTW 에서 한 시계열의 모든 시간은 최소한 다른 시계열의 한 시간에 대응해야 합니다. 또한, 각 시계열의 처음과 끝은 서로 처음과 끝에 대응해야 합니다. 마지막으로 시간 간의 매핑은 과거가 아니라 미래로 이동하는 관계만 표현해야 합니다. 한 시계열에서 시간축상에서 이미 지난 시간을 다른 시계열의 시간에 대응할 수는 없습니다.

이러한 방식을 사용하면 나노 초 단위로 측정된 시계열과 수천 년 단위로 측정된 시계열을 비교할 수 있습니다. 이 알고리즘의 목적은 지나간 시간의 양보다 시각적인 모양을 비교하는 것에 가까운데요, 시간이란 고유한 시간보다는 x 축을 따라 균등한 간격으로 정렬된 점들이라는 일반화된 의미가 있습니다.

Euclidean distance vs DTW



두 거리를 이용한 유사도는 모두 매칭된 데이터들 간의 거리의 합이라는 점에서는 비슷합니다. 유클리드 거리의 경우 계산이 용이하고 연산속도가 뛰어나다는 장점이 있지만 패턴의 떨림과 움직임이 심해질수록 결과가 어긋나는 현상이 발생합니다. 또한 길이가 다른 시계열은 분석이 불가하다는 단점이 있습니다. 사실 앞서 소개한 K-means 클러스터링의 경우 유클리드와 비슷한 거리가 더 적절한 거리지표입니다. 따라서 거리 지표를 선택할 땐 해당 지표가 분석 방법의 가정, 강점 및 약점을 반영하는지 확인한 후 적절한 지표를 선택해야합니다.

그러나 항상 DTW 가 좋은 결과를 도출해주는 것은 아닙니다. DTW 의 왜곡의 정도에 아무런 제약을 주지 않는 경우에는 너무 과도하게 왜곡되어 유사하지 않은 시계열들을 유사한 것처럼 비교할 가능성이 존재하기 때문입니다. 따라서 왜곡할 수 있는 범위를 제약하는 contrained DTW 를 사용하기도 합니다.

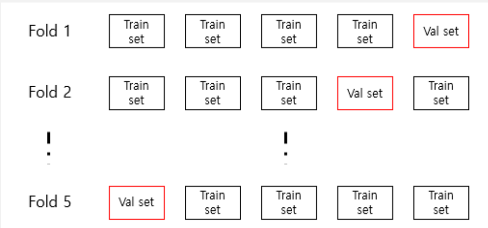
🐹 교차검증 (Cross Validation)

교차검증은 과적합을 방지하기 위한 방법 중 하나로 신뢰성 있는 모델 평가를 진행하기 위해 필요합니다. 우리에게 주어진 데이터는 오직 train set 일 때 성능을 어떻게 평가할까요? “주어진 train set 안에서 train set 과 검증할 validation set 을 나누어 평가하자”가 바로 교차검증의 핵심 아이디어입니다.

교차검증에 대해서 자세하게 다루지는 않겠습니다. 데이터마이닝팀에서 1 주차에 다뤘고, 2 주차에 패키지를 하면서 이제는 익숙해진 개념이라고 생각하고 다음으로 넘어가겠습니다.

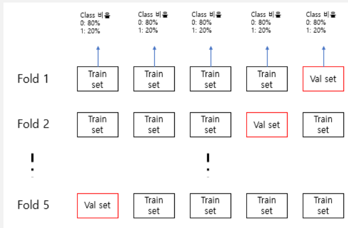
[참고]다양한 교차 검증 방법

K-fold CV



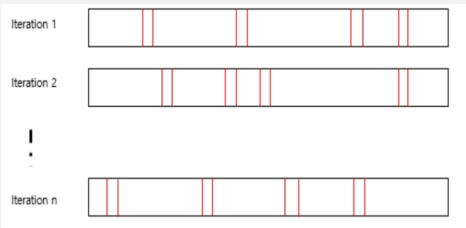
가장 일반적으로 사용되는 교차 검증 방법으로 k 개의 다른 데이터셋을 구성하고 모델을 학습시키는 방법
k=n(데이터의 개수)이면 LOOCV 와 같다

Stratified K-fold CV



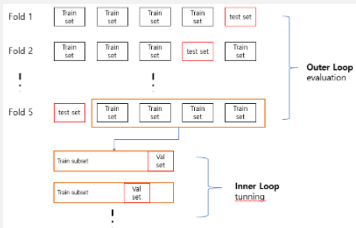
기존 k-fold 방법에서 계층까지 고려한 교차검증 방법
한국어로는 층화 k-fold 라고 하며 클래스가 있는 분류문제에서 활용회는 교차 검증 방법

Repeated Random Subsampling CV



임의로 validation set 을 추출하여 평가하고 하이퍼파라미터를 튜닝하는 교차 검증 방법
validation set 이 임의로 선정되어 중복 선택될 수도 있고 한번도 선택되지 않은 데이터가 있을 수도 있다

Nested CV

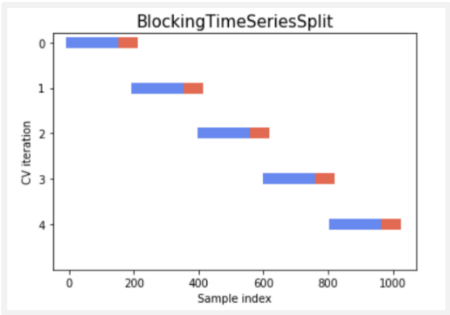


기존의 교차검증을 중첩한 방식
Outer loop 와 Inner loop 으로 구성되어 있으며
각 inner loop 에서 나온 최적의 파라미터를 이용하여
각 fold 에서 test set 으로 평가하는 방법

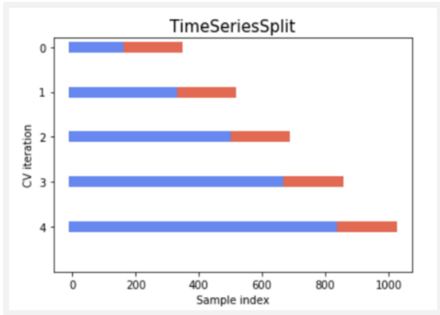
그렇다면 우리가 잘 아는 교차검증을 굳이 시계열자료분석팀에서 3 주차에 하는 이유는 무엇일까요??

➪ 시계열 데이터는 시간이라는 특별한 속성을 지니고 있는데 이러한 데이터를 무작위로 섞는다면 시계열 데이터라는 정보의 손실이 발생합니다. 따라서 시계열데이터를 다룰 때는 기존의 CV 와는 다른 방식으로 해주어야 합니다.

Blocked Time Series CV



Time series CV



위 그림만으로도 두 방법에 어떤 차이가 있는지, 각각이 어떤 방식으로 이루어지는지 이해가 되지 않나요? 두 방법에서 주목할 점은 모두 검증할 데이터 셋이 학습에 사용한 데이터셋보다 미래시점이라는 점입니다. 1 주차때의 기억을 떠올려보면 우리의 궁극적인 목표는 과거의 데이터를 활용하여 "미래"를 예측하는 것입니다. 각각에 대해 조금 더 자세히 알아보겠습니다.

1) Blocked Time series CV

Fixed rolling window forecast 라고도 합니다. Rolling window 란 동일한 사이즈만큼 옆으로 이동한다는 의미입니다. 동일한 사이즈의 윈도우가 움직이며 각 윈도우 내에서 일정한 비율로 train/test 를 나누는 방식입니다. 각 단계에서 사용되는 데이터의 양이 같다는 점이 아래의 blocked time series CV 와 가장 큰 차이점입니다.

Dataset: [1,2,3,4,5] ⇨ 아래의 규칙에 따라 train/test set을 만들려고 합니다.

- 모든 train:test의 비율은 2:1이다
- Train+test의 사이즈는 동일하게 유지된다

다음과 같이 데이터 셋을 형성하는 방식이 blocked time series CV입니다.

- Training: [1,2] Test: [3]
- Training: [2,3] Test: [4]
- Training: [3,4] Test: [5]

이 이후에는 다른 CV과정과 동일하게 각 데이터 셋에서 구한 정확도(오차)의 평균을 구하여 모델의 성능을 확인합니다.

2) Time series CV

Expanding window forecast 라고도 합니다. Expanding window 는 누적하며 이동한다는 의미입니다. 위 그림과 같이 데이터가 시간 순으로 정렬되어 있을 때 가장 먼저 제일 작은 사이즈의 train set 을 이용하여 test set 을 예측하고, 그 다음 이전단계에서 train, test 에 활용되었던 데이터를 전부 train set 으로 다시 활용하는 과정을 반복하면서 점차 train set 의 크기를 늘리는 방법입니다. 아래의 예시를 통해 알아보겠습니다.

Dataset: [1,2,3,4,5] ⇨ 아래의 규칙에 따라 4개의 train/test set을 만들려고 합니다.

- 모든 test set은 unique하다
- Train set의 관측치는 직전 단계에서 test로 활용된다

다음과 같이 데이터 셋을 형성하는 방식이 blocked time series CV입니다.

- Training: [1] Test: [2]
- Training: [1, 2] Test: [3]
- Training: [1, 2, 3] Test: [4]
- Training: [1, 2, 3, 4] Test: [5]

이 이후에는 다른 CV과정과 동일하게 각 데이터 셋에서 구한 정확도(오차)의 평균을 구하여 모델의 성능을 확인합니다.

IV. 부록

여기서 끝나면 너무 아쉽잖아요,,,,,,,,,,,,,,,,,,,,,,

1. ARFIMA

ARFIMA는 Autoregressive Fractionally Integrated Moving Average의 약자로, ARIMA모델에서 차분의 차수를 정수가 아닌 값까지 허용해주어 시계열이 가진 장기 기억(long term memory)를 보존해주는 모델입니다. 정수차원의 차분을 시행하면, 과거의 관측치를 빼 주어 장기 기억을 분석하는 것이 불가능합니다. 따라서 실수 차원의 차분을 통해 메모리를 최대한 보존하는 시계열을 만듭니다.

장기 기억을 가진 시계열은 ACF가 훨씬 천천히 감소한다는 특징이 있습니다. ACF가 아주 천천히 감소하는 경우 차분을 여러 번해도 추세가 온전히 제거되지 않을 수 있을 뿐 아니라 차분의 횟수가 늘어나면 추정해야 할 모수 또한 증가합니다. 이러한 이유로 ARFIMA모델이 사용됩니다.

실수 차원의 차분

이항급수

$$(1-B)^d = 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots$$

$$(1-B)^{0.2} = 1 - 0.2B + \frac{0.2(0.2-1)}{2!}B^2 - \frac{0.2(0.2-1)(0.2-2)}{3!}B^3 + \cdots$$

$$(1-B)^{0.2} = 1 - 0.2B - 0.08B^2 - 0.048B^3 - 0.0336B^4 \cdots$$

$$(1-B)^{0.2}x_t = x_t - 0.2x_{t-1} - 0.08x_{t-2} - 0.048x_{t-3} - 0.0336x_{t-4} \cdots$$

$$(1-B)^{0.4}x_t = x_t - 0.4x_{t-1} - 0.12x_{t-2} - 0.064x_{t-3} - 0.0416x_{t-4} \cdots$$

$$(1-B)^{0.6}x_t = x_t - 0.6x_{t-1} - 0.12x_{t-2} - 0.056x_{t-3} - 0.0336x_{t-4} \cdots$$

$$(1-B)^{0.8}x_t = x_t - 0.8x_{t-1} - 0.08x_{t-2} - 0.032x_{t-3} - 0.0176x_{t-4} \cdots$$

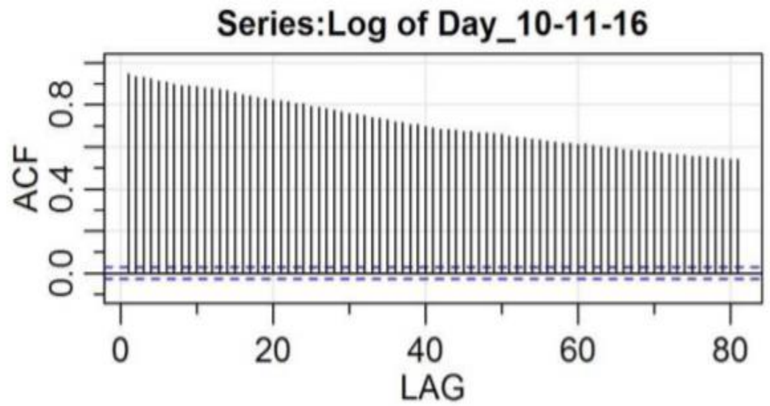
$$(1-B)^{1.0}x_t = x_t - 1.0x_{t-1} - 0x_{t-2} - 0x_{t-3} - 0x_{t-4} \cdots$$

위의 식을 통해 d가 1에 가까워질수록 보존되는 과거 정보의 양이 감소함을 알 수 있습니다. 또한 d 가 작을수록 먼 과거의 데이터들이 반영됨을 알 수 있습니다.

ARFIMA(p,d,q) 다음과 같이 모형을 표현할 수 있으며 이때 Z_t는 백색잡음입니다.

$$\phi(B)(1-B)^d Y_t = \theta(B)Z_t, \qquad 0 < d < \frac{1}{2}$$
$$\phi(B) = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)$$
$$\theta(B) = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)$$

- 정상성 조건을 만족하기 위하여 $|d| < 0.5$를 만족해야합니다. (절댓값이 0.5보다 커지만 분산 → ∞)
- $-0.5 < d < 0$인 경우 ACF의 합이 0이 됩니다.
- $d = 0$인 경우 ARMA과정이 됩니다.



장기기억을 가지는 시계열의 ACF

Long term memory process(장기기억 확률과정)

→ 정상성을 만족하는 확률과정 $\{Z_t\}$의 자기상관함수 $\rho(k)$가 $0 < d < \frac{1}{2}$인 어떤 실수에 대하여 $\rho(k) \sim Ck^{2d-1}, k \to \infty \ (C > 0)$을 만족할 때 장기기억 확률과정이라고 합니다.

특징 1) ACF의 합이 무한대로 발산한다.

특징 2) ACF가 빠르게 0으로 수렴하지 않는다.

2. ARMAX

여러분.. 이 세상의 많고 많은 일들이 어떻게 단순하게 설명될 수 있겠어요? ARMAX는 기존의 ARMA모형에 외부요인(eXogenous)을 추가시켜준 모형입니다. 예를 들어 주가를 예측하는데 날씨 데이터를 추가해주는 것입니다! 추가된 외부요인은 연속형 변수일수도 있고 범주형 변수일수도 있습니다. 이렇게 외부 요인을 추가하여 예측력을 높여주는 모델이 ARMAX이고, 여기에 차분을 더하면 ARIMAX가 됩니다. 일반적인 모형은 아래와 같습니다. Y_t와 X의 관측값 수는 일치해야 합니다.

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

추정은 다음과 같은 과정을 통해 이루어집니다.

$Y_t = \sum_{j=1}^{p}\phi_j Y_{t-j} - \sum_{i=1}^{q}\theta_i Z_{t-i} + Z_t + \beta_0 + \beta_1 X_t$ 에서 $\sum_{j=1}^{p}\phi_j Y_{t-j} - \sum_{i=1}^{q}\theta_i Z_{t-i} + Z_t$ 부분을 u_t로 치환하여 $Y_t = \beta_0 + \beta_1 X_t + u_t$ 로 표현하여 ARMA를 따르는 잔차를 가진 i회귀식으로 생각합니다. 회귀식을 추정한 뒤 나온 잔차에 대해 다시 ARMA 모형을 적합시켜 계수를 추정합니다.

- **ARMAX:**

$$\Phi(\mathbf{B})X_t = \boldsymbol{\beta}^T\mathbb{X} + \Theta(\mathbf{B})Z_t$$

- **ARIMAX:**

$$\Phi(\mathbf{B})\nabla^d X_t = \boldsymbol{\beta}^T\mathbb{X} + \Theta(\mathbf{B})Z_t$$

- **SARIMAX:**

$$\Phi(\mathbf{B})_p\Phi(\mathbf{B}^m)_P\nabla^d(\nabla_m)^D X_t = \boldsymbol{\beta}^T\mathbb{X} + \Theta(\mathbf{B})_q\Theta(\mathbf{B}^m)_Q Z_t$$

ar모형 이외에도 위와 같이 다양한 모델에 외생변수를 추가할 수 있습니다.

REFERENCE

· Introduction to time series and forecasting

· Forecasting: Principles and Practice

· 시계열 분석 이론 및 SAS 실습

· (SAS/ETS 와 R 을 이용한) 시계열분석

· (R 프로그램에 기반한) 시계열 자료 분석

· 실전 시계열 분석

· DTW
https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/
https://www.netinbag.com/ko/internet/what-is-dynamic-time-warping.html
https://ichi.pro/ko/sigyeyeol-deiteoe-k-pyeong-gyun-keulleoseuteoling-eul-jeog-yonghaneun-bangbeob-44875069226403
https://gist.github.com/bistaumanga/6023705
https://medium.datadriveninvestor.com/dynamic-time-warping-dtw-d51d1a1e4afc
https://rtavenar.github.io/blog/dtw.html

· CV
https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4
https://gmnam.tistory.com/230
https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9
https://www.kaggle.com/cworsnup/backtesting-cross-validation-for-timeseries/notebook

· stochastic vs deterministic
https://thesamuelsoncondition.com/2016/01/23/time-series-iii-deterministicstochastic-seasonality/
https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=muzzincys&logNo=220088405540

· ARMAX
https://sosoeasy.tistory.com/394

· ARFIMA
https://blog.naver.com/statstorm/222427718779
https://m.blog.naver.com/PostView.naver?blogId=chunjein&logNo=222071363797&proxyReferer=https:%2F%2Fm.search.naver.com%2Fsearch.naver%3Fquery%3D%25EC%258B%259C%25EA%25B3%2584%25EC%2597%25B4%2B%25EC%259E%25A5%25EA%25B8%25B0%25EA%25B8%25B0%25EC%2596%25B5%26where%3Dm%26sm%3Dmtp_hty.top

\+ 이전 시계열 교안(모두 감사합니다...)