시계열자료분석팀 3주차

[목차]

1 2주차 복습

정상 시계열임을 가정하는 AR, MA, ARMA 모형에 대해 공부했습니다. 정상 시계열이란 추세 또는 계절성이 없고 약정상성 조건을 만족하지만, 오차항이 백색잡음이 아닌 시계열 자료입니다.

AR 모형은 과거 시점의 관측값들과 현재 시점의 오차항으로 설명되는 모델로, ACF는 지수적으로 감소하고 PACF는 p+1차부터 절단되는 특징이 있었습니다. MA 모형은 과거 시점의 오차항들로 설명되는 모델로, ACF는 q+1차부터 절단되고, PACF는 지수적으로 감소하는 특징이 있었습니다. ARMA 모형은 AR과 MA를 동시에 포함하는 모델로, ACF와 PACF 모두 지수적으로 감소하기 때문에 information criteria 등을 통해 모형을 선택했습니다. AR, MA, ARMA 모형은 정상성 또는 가역성 조건을 만족하기 위해 계수의 절댓값이 1보다 작아야 했습니다.

시계열 모형은 모형 식별, 모수 추정, 모형 진단, 예측 순으로 적합되었습니다. 모수 (파라미터) 추정에는 보통 MLE가 사용되고, 모형을 진단할 때는 모수와 잔차에 대한 검정이 필요했습니다. 예측을 할 때는 미래 시점의 관측값은 현재까지의 관측값들의 선형결합임을 가정했습니다. 참 값과 예측값 사이의 오차를 최소화하는 MSPE를 최소화하는 방향으로 예측값을 추정했습니다.

1주차에서 이미 확인했듯이, 대부분의 시계열 데이터는 비정상 시계열입니다. 즉, 시계열이 약정상성을 만족하지 않고, 평균 또는 분산이 일정하지 않아 추세나 계절성이 존재하거나, 공분산이 시점에 의존할 수도 있습니다. 시계열 데이터를 다룰 때마다 정상화 과정을 사전에 거쳐야 한다면 번거롭겠죠? 이번주에는 비정상 시계열 모형에 대해 공부해보겠습니다. 비정상 시계열 모형은 사전에 정상화 과정을 따로 진행할 필요 없이, 모델 자체에서 정상화가 이뤄집니다! (°0°)

2 ARIMA

2.1 ARIMA 모형의 정의

ARIMA 모형은 **추세(polynomial trend)가 있어 정상성을 만족하지 않는 시계열 자료**에 적용 가능한 모델입니다. ARIMA 모형은 **ARMA에 차분이 결합**된 형태입니다. d차 차분한 시계열이 정상 과정(stationary process) ARMA(p, q)를 따를 때, 해당 시계열에 ARIMA(p, d, q)를 적합할 수 있습니다. 다음과 같이 표현할 수 있습니다.

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

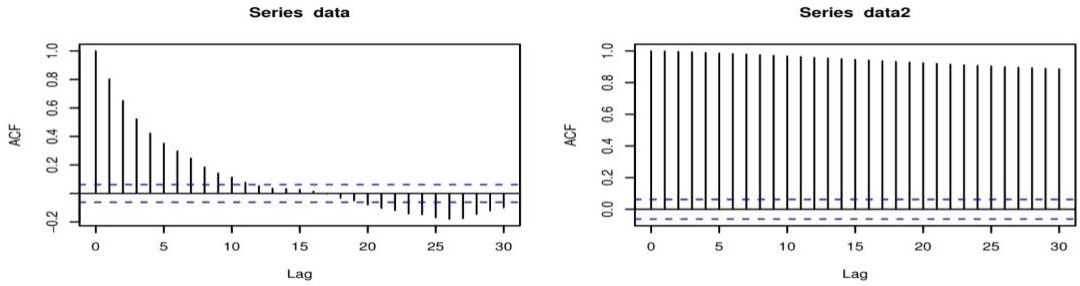
$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d X_t=\left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

d=0인 경우, ARMA(p, q) 모형과 동일합니다. d가 1 이상인 경우, d차의 추세를 제거하게 됩니다. 즉, **ARIMA(p, d, q) 모형은 주어진 시계열 데이터가 polynomial trend가 있고 오차가 ARMA(p, q)를 따를 때 적용**할 수 있습니다.

ARIMA 모형의 장점은 사전에 정상화할 필요 없이 모형 자체에 차분을 통한 추세 제거가 포함되어 있으며, 여전히 선형과정이라는 점입니다.

2.2 ARIMA 모형의 적합 절차

1) TS plot과 ACF 그래프를 통해 정상 / 비정상 시계열 여부를 판단합니다. 정상 시계열과 비정상 시계열의 ACF 그래프에서 차이는 ACF의 감소 속도에서 나타납니다. 정상 시계열의 경우 ACF가 빠른 속도로 지수적으로 감소하지만, 비정상 시계열의 경우 ACF가 느리게 선형으로 감소합니다.



2) TS plot에서 추세가 관측된다면 관측되는 추세에 맞춰 d차 차분을 적용합니다. 즉, ARIMA(p, d, q)의 모수 d의 차수를 결정합니다. 보통 1차 또는 2차를 많이 사용하는데요, 이는 과대차분의 위험이 있기 때문입니다.

   과대차분(Overdifferencing)이란 차분을 필요 이상으로 진행하는 것입니다. 이미 정상화가 되었음에도 불구하고 차분이 더 적용된다면, 정상성 자체에는 문제가 없지만 ACF를 복잡하게 만들거나 분산이 커지는 등의 문제가 발생합니다.

   비정상 시계열 데이터에 1차 또는 2차 차분을 적용해보고, ACF와 PACF 그래프를 확인했을 때 지수적으로 감소하는 그래프가 나타난다면 적절하다고 볼 수 있습니다.

3) 모형 적합 절차에 따라 p, q의 차수를 결정하고, 모수를 추정하고, 모형을 진단했다면, 최종 모형으로 예측을 진행할 수 있습니다.

추세 또는 계절성이 있는 비정상 시계열을 다룰 때, 1주차에서 배웠던 회귀, 평활 등으로 따로 정상화를 적용하고 정상 시계열 모형을 적용하는 방법도 가능합니다. 하지만 이 경우 입력값은 정상화를 거친 오차항이기 때문에 결과값은 오차항에 대한 예측값뿐이므로, 추세 또는 계절성을 추정하여 추가로 더해주어야 하는 번거로움이 있습니다.

그러나 비정상 시계열 모형은 연산 과정에 차분을 통한 정상화가 포함되어 있습니다! 따라서 입력값은 전처리를 거치지 않은 원본 시계열 데이터입니다. 예측 과정에서 ARMA 적합 후 차분에 반대되는 개념으로 '적분'을 거쳐 최종 예측값을 도출하게 됩니다. (왜냐하면 차분을 하면 시계열이 정상성을 만족(=ARMA 적합 가능)할 때 사용 가능하기 때문입니다.) 따라서 정상 시계열 모형을 사용했을 때와는 다르게 추세 또는 계절성을 더해주는 후처리 과정 없이, 도출된 예측값은 원본 시계열 데이터에 대한 것이므로 바로 사용할 수 있습니다. 이러한 장점은 곧 공부할 SARIMA에서도 동일합니다.

3 SARIMA

3.1 SARIMA 모형의 정의

SARIMA 모형은 seasonal ARIMA 의 줄임말입니다. ARIMA 모형은 추세가 있는 비정상 시계열에 적용이 가능하다고 했습니다. **SARIMA 모형은 추세뿐만 아니라 계절성도 존재하는 비정상 시계열에 적용 가능한 모형**으로, 가장 확장된 개념의 ARMA 모형입니다. ARIMA와 동일하게 연산 과정 내에 차분을 통한 정상화를 포함합니다.

1주차에 전통적인 분해법을 통해 시계열을 $Y_t = s_t + Z_t$ 과 같이 표현하고 계절성분 s_t를 추정하여 제거하는 방법에 대해 공부했습니다. 이 때의 계절성분은 결정적(deterministic) 계절 성분으로, 모든 주기에 있어서 계절성분이 동일함을 가정했습니다. 하지만, 현실에서는 주기마다의 계절성이 다를 수 있습니다. 계절성은 주기마다 동일하지 않지만, 각각의 계절성은 분명 서로 연관성을 가질 것입니다. 이러한 아이디어로부터 나온 SARIMA 모델은 **계절성 사이의 상관관계**에 대해 모델링을 하는 확률적 접근법입니다.

주기 s=12인 시계열 데이터를 예시로 SARIMA에 대해 이해해보겠습니다.

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

각 열을 고정합니다. 즉, 1월에 속하는 $Y_1, Y_{13}, Y_{25}, \ldots, Y_{1+12(r-1)}$를 하나의 열로 고정합니다. 1월부터 12월까지의 각 열(=월)은 동일한 ARMA(P, Q) 모형을 따른다고 가정합니다. 같은 월에 해당하는 시계열끼리의 상관관계를 나타낸 것입니다.

1주차에서 공부한 계절 평활법(seasonal smoothing)은 동일한 월에 속하는 시계열의 값을 평균내기 때문에, 1월의 계절성은 주기에 상관없이 모두 같은 값을 가졌습니다. 하지만 SARIMA의 경우 각 월은 ARMA(P, Q)를 따르는 계절성을 갖는다고 가정하기 때문에, 주기마다의 계절성이 동일하지는 않지만, Year 1의 계절성분과 Year2의 계절성분의 상관관계가 모델에 의해 설명될 수 있습니다.

여기까지는 동일한 월(month)에 속하는 시계열끼리의 상관관계를 설명한 것입니다. 그렇다면, 1월, 2월, 3월, ⋯ , 12월까지 각 월끼리의 상관관계도 설명해줄 필요가 있겠죠?

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

이번에는 각 행을 고정합니다. 즉, Year 1에 속하는 $Y_1, Y_2, Y_3, \ldots, Y_{12}$를 하나의 행으로 고정합니다. 동일한 연도에 속한 연속된 값들에 대해 ARMA(p, q) 모형을 따른다고 가정합니다. 연속된 시계열끼리의 상관관계를 나타낸 것입니다.

여기까지 각 열에 대해 ARMA(P, Q) 모형을 통해 주기마다의 계절성분을 표현하고, 각 행에 대해 ARMA(p, q) 모형을 통해 주기 내에서 연속된 시계열끼리의 상관관계를 나타

냈습니다. 지금까지의 내용을 식으로 표현하면 다음과 같습니다.

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

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

파란색 부분은 12시점 전과의 상관관계로, 주기마다의 계절성분을 나타냅니다. 빨간색 부분은 바로 전 시점과의 상관관계로, 연속된 시계열끼리의 상관관계를 나타냅니다. 첫번째 식을 정리하면 두번째 식과 같이 표현할 수 있습니다.

여기까지만 하면 SARMA(p, q)(P, Q) 모델입니다. 차분이 포함되지 않은 것이죠! 차분까지 포함하여 SARIMA(p, d, q)(P, D, Q)를 완성할 수 있습니다.

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

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

초록색 부분은 전체 시계열의 추세에 대해 d차 차분을 적용하는 것이고, 주황색 부분은 계절성분이 갖는 D차의 추세에 대해 lag-12 차분을 D번 적용하는 것입니다.

정리하자면, SARIMA(p, d, q)(P, D, Q) 모형에서 p, d, q는 연속된(serial) 시계열에 대해 모델링하는 부분이고, P, D, Q는 계절성분에 대해 모델링하는 부분입니다. d는 전체 시계열이 갖는 추세(polynomial trend)를 제거하는 부분이고, D는 계절성분이 갖는 추세를 제거하는 부분입니다. 모든 모수를 사용할 필요 없이, 필요에 따라 모형을 선택하면 됩니다. SARIMA(0, 0, 0)(1, 0, 0) 또는 SARIMA(1,1,0)(1,0,0) 등 다양한 조합으로 모델링할 수 있습니다.

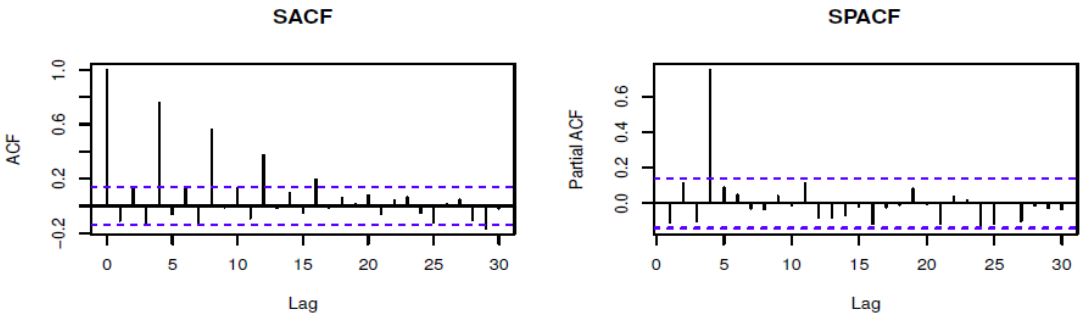
3.2 SARIMA 모형 적합 절차

1) TS plot 또는 잔차를 확인하여 이분산성이 나타나는 경우 변환을 통해 등분산성을 만족하게 합니다.

2) 차분 또는 계절 차분 여부를 결정합니다. 즉, d와 D의 차수를 결정합니다. ACF 그래프가 지수적으로 감소하지 않고, 느리게 감소한다면 차분이 필요하다고 볼 수 있습니다. 특히, ACF 그래프가 느리게 감소하면서도 규칙적으로 구불구불한 형태를 보인다면, 계절성분에 의한 비정상성을 의미합니다. 1차 또는 2차 차분을 적용해보고 ACF 그래프의 변화를 확인하며 차분의 차수를 결정합니다.

3) P, Q와 p, q의 차수를 결정합니다. ARMA가 아닌 AR 또는 MA만 사용되는 경우(SAR, SMA 모형이라고 합니다), ACF와 PACF 그래프가 절단되는 차수를 보고 모수의 차수

를 결정할 수 있습니다. 이 때, 계절성분의 모수인 P, Q를 결정할 때에는 주기마다의 ACF 또는 PACF를 참고하면 됩니다.



위 그래프는 주기가 4인 시계열 데이터로, 4시점 간격마다 ACF가 지수적으로 감소하고, PACF는 4 이후로 절단됩니다. 따라서 SAR(1) 모형을 적합할 수 있습니다. SAR(4)가 아닌 SAR(1)이라는 점을 주의해야 합니다!

하지만 ACF와 PACF 모두에서 절단되는 부분이 보이지 않는다면, ARMA를 적합해야 하기 때문에 information criteria를 참고하여 모수의 차수를 결정할 수 있습니다. 현실적으로 모든 모수의 IC를 비교하는 것은 불가능하기 때문에 범위를 정해놓고 해당 범위 내에서 가장 IC가 높은 모수의 차수를 찾는 식으로 모형을 선택합니다.

4) 모수를 추정합니다. 아래 R 커맨드를 통해 MLE를 구할 수 있습니다.

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

5) 예측 과정은 2주차에서 다뤘던 모형의 적합 절차에서의 예측의 과정과 동일합니다. 아래의 R 커맨드를 통해 예측할 수 있습니다. n.ahead 파라미터는 몇 시점 이후의 값에 대해 예측할 것인지를 설정하는 부분입니다.

   Forecast(model, n.ahead=h)

4 이분산 시계열 모형

지금까지 만나본 비정상 시계열 모형은 등분산성을 가정한 다음 평균 부분의 움직임에 관심을 갖는 모형이었습니다. 그러나, 환율, 주가 등 금융 관련 시계열 데이터에서는 불확실성을 나타내는 분산 부분의 움직임에 관심을 갖습니다!

경제학에서의 '변동성(volatility)'을 통계학에서 '조건부 분산(conditional variance)'을 통해 나타낼 수 있습니다. 일반적인 이분산성은 구조적인 변동성이 이전 기간의 변동성과

관련이 없습니다. 하지만 조건부 이분산성(conditional heteroscedasticity)은 미래의 변동이 현재까지의 상황에 의존하는 것입니다. 많은 금융 시계열 자료는 이러한 조건부 이분산성을 갖습니다.

4.1 수익의 통계적 특성

이번 장에서 공부할 이분산 시계열 모형 ARCH와 GARCH 모형은 수익율에 대해 모델링하는 모형입니다. 수익율에 대한 정의로는 simple return과 log return이 대표적입니다. t 시점의 가격을 P_t라고 할 때, 아래와 같이 표현할 수 있습니다.

$$[simple\ return]\quad R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

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

Log return은 simple return에 근사하고, 덧셈으로 표현되므로 통계적 계산에 더 편리합니다. 또한, log를 취하면 분산을 안정화해주는 효과도 있기 때문에, 앞으로 다루는 모든 수익율은 log return을 의미합니다.

수익율이 갖는 통계적 특성에 대해 알아보겠습니다.

- 평균에 대칭적이다.
- r_t 자체는 상관관계가 크지 않다. (ACF가 거의 백색잡음에 가까움)
- $|r_t|$ 또는 r_t^2은 강한 상관관계를 갖는다.
- Spikes 때문에 QQ plot의 양 끝이 정규분포에서 크게 벗어나는 heavy-tailed distribution을 갖는다.
- Local variance가 시간에 의존하여 발생하는 조건부 이분산성을 갖는다.

주목할만한 점은, 수익율 자체는 상관관계가 작아 거의 백색잡음에 가깝지만, 절댓값 또는 제곱을 취하면 강한 상관관계를 보인다는 것입니다. 절댓값 또는 제곱은 변동(=분산)을 나타냅니다. 즉, 수익율의 분산은 서로 강한 상관관계를 가지며, 변동성이 시점에 의존하는 조건부 이분산성을 갖습니다.

4.2 ARCH (Auto-Regressive Conditional Heteroscedasticity)

ARCH(1) 모형은 아래와 같이 가정합니다. 이 때, F_{t-1}는 t-1 시점까지의 모든 정보의 집

합으로, $r_t|F_{t-1}$는 t-1까지의 모든 정보를 다 알고 있을 때의 t 시점의 수익율입니다.

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

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

$r_t|F_{t-1}$의 분산인 ${\sigma_t}^2$에 대해 AR(1) 구조를 가정한 것이 ARCH(1) 모형이라고 할 수 있습니다. ${\sigma_t}^2$이 ${r_{t-1}}^2$에 의존하고, 따라서 ${r_t}^2$도 ${r_{t-1}}^2$에 의존함을 나타냈습니다.

ARCH(m) 모형은 ${\sigma_t}^2$에 대해 AR(m) 구조를 가정한 것입니다. 식으로 표현하면 아래와 같습니다.

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

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

ARCH 모형의 특징은 비선형적 모델이라는 점입니다. $r_t = \sigma_t \varepsilon_t$ 에서 볼 수 있듯이 곱셈을 통해 표현했기 때문에, ${r_t}^2$ 에 대해 전개한 식은 선형결합으로써 표현할 수 없습니다.

4.3 GARCH (Generalized Auto-Regressive Conditional Heteroscedasticity)

GARCH 모형은 ${\sigma_t}^2$에 대해 ARMA(p, q) 구조를 가정한 것으로, ARCH 모형보다 확장된 모형입니다. ${r_t}^2$ 은 ${\sigma_t}^2$ 에 근사하고, 둘 사이의 오차를 ${\eta_t}^2 = {r_t}^2 - {\sigma_t}^2$ 라고 할 때, GARCH(1, 1) 모형은 다음과 같이 표현할 수 있습니다.

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

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

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

세 번째 식은 두 번째 식에 ${\sigma_t}^2 = {r_t}^2 - {\eta_t}^2$를 대입해 정리한 것입니다. 빨간색 부분이 과거의 자기 자신의 함수로 표현된 AR(1) 부분이고, 파란색 부분이 오차의 함수로 표현된 MA(1) 부분입니다.

GARCH(p, q) 모형은 아래와 같이 표현할 수 있습니다.

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

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

5 시계열 데이터 with ML / DL

지금까지는 시계열 자료를 다루는 전통적인 통계적 모델들에 대해 다뤘습니다. 여전히 통계적 모델링에 의한 접근법이 좋은 성능을 보이고 있지만, 머신러닝과 딥러닝을 통한 접근법의 위력도 무시할 수 없겠죠? 시계열자료분석팀의 클린업 마지막 장에서는 머신러닝과 딥러닝을 통한 시계열 분석에 대해 약간의 맛보기(?)를 하고 마무리하도록 하겠습니다.

5.1 시계열 데이터의 전처리

1) 결측치 보간

결측치란 값이 존재하지 않는 것을 의미합니다. 실제 데이터를 다룰 때, 심심찮게 결측치를 만나볼 수 있습니다.

시계열 데이터의 경우, 결측치 제거를 통해 단순하게 문제를 해결하려 한다면, 타임스탬프가 일정해지지 않는 문제, 해당 시점의 평균과 분산에 왜곡이 생기는 문제 등이 발생합니다. 일반적으로는 평균 또는 최빈값으로 결측치를 대체하는데, 시계열 데이터의 경우 시계열의 특성을 반영한 방법으로 결측치 보간을 진행해야 합니다.

- Last Observation Carried Forward (LOCF)

  : 직전에 관측된 값으로 결측치 보간

- Next Observation Carried Backward (NOCB)

  : 직후에 관측된 값으로 결측치 보간

- Moving Average / Moving Median

  : window 내에서의 평균 또는 중앙값으로 결측치 보간

그러나 결측치 전후로 패턴이 급격히 변화하는 구간에서는 위와 같은 결측치 보간법으로는 불충분합니다. 예를 들어, 주가 예측 시에 주가가 급격하게 상승/하락하는 구간에 결측치가 존재할 때는 위의 방법을 통한 결측치 보간은 해당 구간에서 평균과 분산에 왜곡을 일으킬 가능성이 큽니다. 이 경우, 아래와 같은 방법을 고려해볼 수 있습니다.

- 선형 / 비선형 보간법

  : 결측치를 중심으로 window를 설정하고 해당 구간 내에 선형 또는 다항 함수를 적합하여 적합된 함수를 통해 결측치를 보간

- 스플라인 보간법 (Spline Interpolation)

  : 결측치를 중심으로 한 window의 전체 값을 한 번에 적합하지 않고, 이를 소구간으로 분할하여 스플라인(각 구간마다 함수를 적합하고 모든 구간에서 함수가 매끄럽게 이어지도록 하여 전체 구간에 함수를 적합하는 방법)을 적용하여 결측치를 보간 (데마팀 2주차 클린업 참고!)

- 모델링을 통한 결측값 예측

  : (결측치가 너무 많은 경우) 각 결측치마다 이전까지의 시계열 데이터를 활용해 모델링한 다음, 각 결측값을 예측하여 보간

2) 노이즈 처리 (Denoising)

노이즈란 의도하지 않은 데이터의 왜곡을 불러오는 모든 것을 의미합니다. 예를 들어, 평균적인 대중교통 이용량에 대해 분석하고자 할 때, 유명 가수의 콘서트로 인해 이용량이 급격하게 상승한 것은 연구자의 관점에 따라 노이즈로 볼 수 있습니다. 이런 경우에 노이즈를 제거할 수 있는 denoising 기법에 대해 간단히 알아보겠습니다.

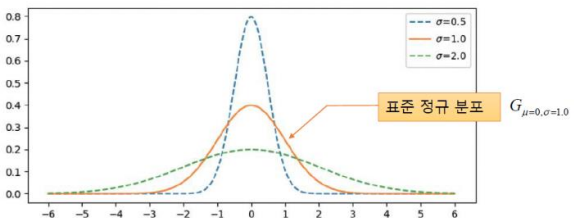
- 평활 (Smoothing)

: 이동평균평활법 또는 지수평활법을 통해 denoising 할 수 있습니다. 이 방법은 노이즈가 많이 발생하는 데이터에서는 노이즈 자체가 평균에 반영되기 때문에 적절하지 않고, 노이즈가 적은 데이터에 효과적입니다.

- 가우시안 필터 (Gaussian Filter)

$$G_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

• μ : 평균
• σ : 표준편차



: 가우시안 필터를 적용하여 denoising 할 수 있습니다. 가우시안 필터는 중심에 가까운 데이터에는 더 큰 가중치를, 멀어질수록 작은 가중치를 부여합니다. 가우시안 필터도 평활 방법의 일종으로 볼 수 있으며, 주로 이미지 처리에서 블러링 또는 노이즈 처리에 사용되지만 시계열 데이터에도 적용 가능합니다.

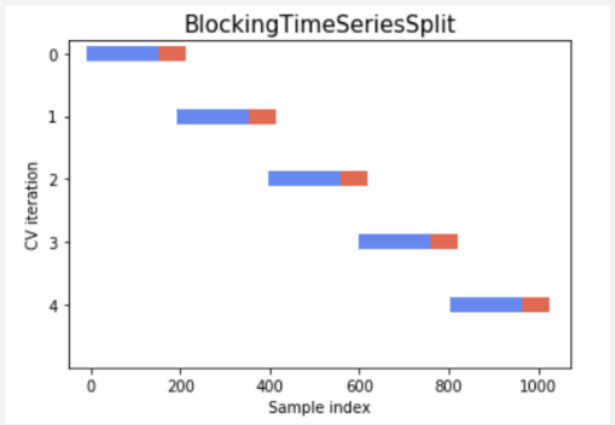
5.2 시계열 데이터의 교차검증

교차검증(Cross Validation; CV)은 과적합을 방지하기 위한 방법 중 하나로 신뢰성 있는 모델 평가를 진행하기 위해 필요합니다. 모델을 검증하기 위해 전체 데이터를 train set과 validation set으로 나누어 평가하는 것이 바로 CV입니다.

시계열자료분석에서도 이러한 CV 과정은 필요하나, 일반적으로 사용하는 K-fold CV 등을 사용할 수 없다는 문제점이 존재합니다. 그 이유는 시계열이라는 것은 말 그대로 시간의 순서가 굉장히 중요한 자료인데, 일반적인 CV 과정은 이러한 시간 순서를 고려하지 않기 때문입니다!

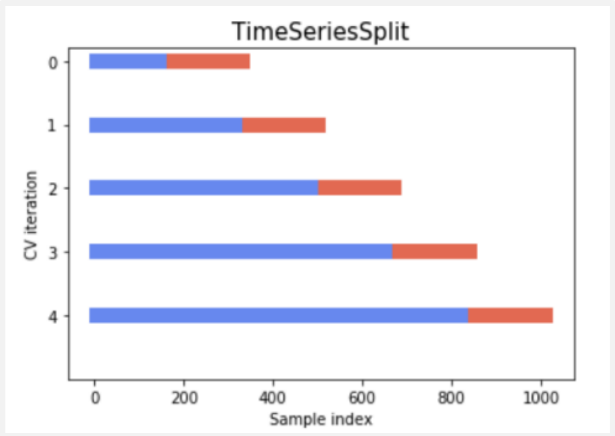
시계열 데이터를 위한 CV 기법이 따로 존재하는데요! 시계열 교차검증 기법 중 blocked time series cv와 일반적인 time series cv에 대해 알아보겠습니다.

1) Blocked Time Series CV



Rolling window CV라고도 부르는 CV 방법입니다. Rolling window란 동일한 사이즈의 window를 옆으로 이동시킨다는 의미입니다. 동일한 사이즈인 window 내에서 일정한 비율로 train과 validation을 분할합니다.

2) Time Series CV

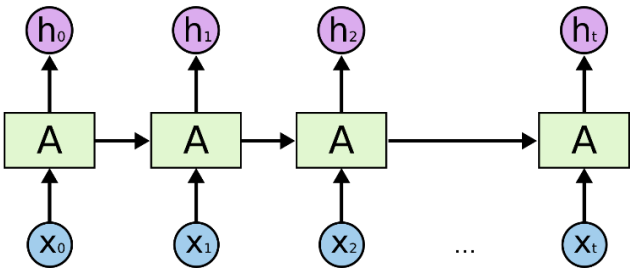


Expanding window CV라고도 부르는 CV 방법입니다. Expanding window란 window를 누적하며 이동한다는 의미입니다. 처음엔 가장 작은 사이즈의 train set을 이용하여 validation set을 통해 검증하고, 다음 단계에서는 이전 단계의 train set + validation set을 통째로 다시 train set으로 활용하는 과정을 반복하면서 점차 train set의 크기를 늘리는 방법입니다.

5.3 시계열 데이터와 딥러닝 기법

시계열 데이터 중 가장 핫한 도메인을 골라보라면 많이들 주식을 얘기하실 겁니다! 주식에 대한 관심이 늘면서 주식 데이터를 활용한 주가 예측과 관련한 데이터 분석이 활발하게 이뤄지고 있습니다. 특히 딥러닝 기법을 활용한 주가 예측 연구가 자주 보이는데요, 딥러닝 기법을 활용한 시계열 데이터 분석에 대해 아주 간단히만! 훑어보고 마무리하겠습니다.

1) LSTM (Long Short-Term Memory Network)

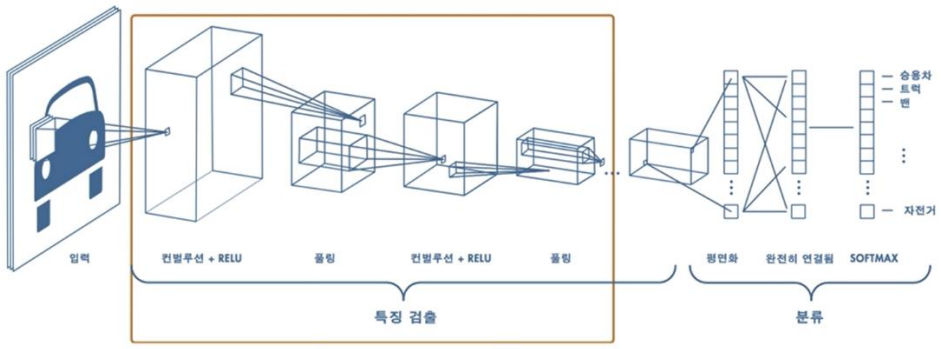


LSTM은 RNN에서 파생된 모델입니다. RNN과 LSTM 모두 모델의 개괄적인 흐름을 도식화하자면 위의 이미지와 같이 표현할 수 있습니다. (hidden state를 계산하는 부분(A)이 RNN에 비해 복잡해진 것이 LSTM입니다.) RNN은 시퀀스의 길이가 길어질수

록 초반의 정보가 손실되는 장기 의존성 문제가 발생합니다. LSTM은 이러한 장기 의존성 문제를 해결하기 위해 등장한 모델로, '정보를 잊는다'는 개념을 추가하여 최근 학습한 값과 함께 이전의 값 중 중요한 것들 위주로만 기억하도록 설계된 모델입니다. (구체적인 내용은 딥러닝팀 3주차 클린업 참고!)

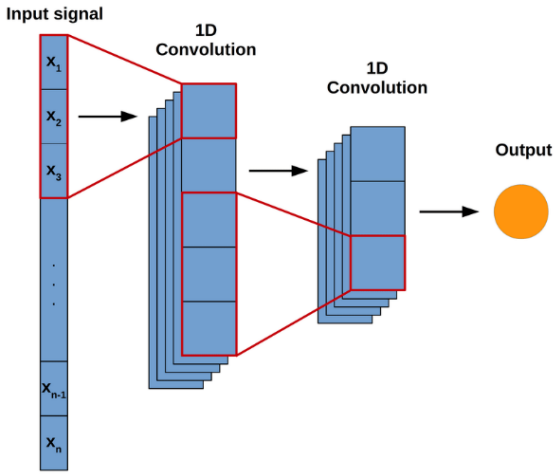
LSTM은 시계열 데이터의 time dependency를 잘 반영할 수 있는 모델입니다. RNN 기반 모델은 이전의 학습 결과가 다음 학습 단계에 영향을 주기 때문에, 종속적인 데이터를 처리하는데 효과적입니다. 다만 randomwalk와 같이 시계열 데이터 간의 종속성이 불분명한 데이터에는 이러한 기법이 효과적으로 작동하지 않을 수 있습니다.

2) CNN (Convolutional Neural Network)



CNN은 보통 이미지 처리에서 많이 사용이 되는 모델입니다. 이미지의 특징을 추출하여 학습하는 모델입니다. (구체적인 내용은 딥러닝 2주차 클린업 참고!) 이미지와 같은 다차원 데이터가 아닌 1차원 데이터에도 적용이 가능합니다! (Conv1D 함수 사용)

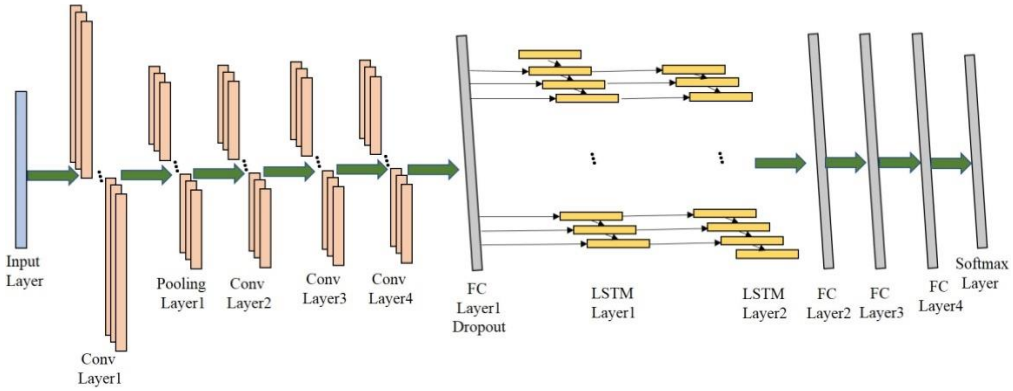
그렇다면 시계열 데이터에 대해 CNN은 어떻게 작동할까요? 예를 들자면, 주가를 볼 때 ~~(대부분은)~~ 순간순간의 가격보다는 주가 그래프의 전체적인 그림에 주목하게 됩니다.



이러한 관점에서 CNN은 시계열 데이터의 전체적인 추세와 패턴을 학습하고 이를 통해 다음 시점을 예측하는 방식으로 작동합니다. 특징을 추출하는 CNN답게, 개별 데이터 하나하나에 초점을 맞추기보단 전반적인 패턴을 파악하고 특징을 추출하여 학습하는 것에 집중합니다.

3) CNN + LSTM

CNN 모델과 LSTM 모델을 결합한 네트워크를 구성하여 시계열 데이터를 학습할 수도 있습니다! (우왕)



위 이미지에서 학습 과정이 잘 나타나 있습니다. 입력 시계열 데이터로부터 피처 맵을 추출하고, 하나의 피처 벡터로 flatten한 것을 LSTM 층에 입력하여 학습하고, 최종적으로 FC layer를 거쳐 결과값을 도출하는 방식으로 작동합니다.

고생하셨습니다!!! (쌍따봉) 시계열자료분석팀 3주차 클린업이 끝났습니다!!!

어질어질한 시계열자료분석 클린업 끝까지 함께해주셔서 다들 감사드리고… 이론적으로 깊게 들어가기보단 큰 흐름을 따라가보자는 느낌으로 최대한 쉽게 풀어보려고 노력했는데, 별로 성공적이었는지는 모르겠네요 ^^;; 쉽지 않은 내용임에도 잘 따라와주고 세미나 발표 때 저보다도 훨씬 더 잘 설명해준 우리 팀원들 넘 고맙고, 청강 와주신 분들께도 감사합니다. (유익했길 바라요..ㅠㅠ) 클린업 3주차까지 다들 수고 많으셨고, 주분 때도 파이팅해봅시다!