

(보안데이터분석) 연습문제_11

1. 시계열 분석에서 정상성(Stationarity)에 대한 설명으로 가장 적절하지 않은 것은 무엇인가?
 1. 정상성을 띠는 시계열은 시간의 흐름에 따라 평균, 분산, 공분산이 일정하게 유지되는 특성을 보인다.
 2. 대부분의 시계열 모델(예: ARIMA 모델)은 정상성을 가정하고 개발되었기 때문에, 비정상 시계열 데이터는 분석 전에 정상성으로 변환하는 과정이 필요하다.
 3. 시계열 데이터가 정상성을 띠지 않으면, 과거 데이터의 패턴이 미래에도 동일하게 반복될 것이라는 가정을 하기 어렵다.
 4. 단위근(Unit Root)은 시계열 데이터의 정상성을 판단하는 중요한 요소 중 하나이며, 단위근이 존재하면 비정상 시계열일 가능성이 높다.
 5. 차분(Differencing)은 시계열 데이터의 분산을 안정화하여 정상성을 확보하는 가장 일반적인 방법이다. ☒ 정규차분은 평균을 안정화 시키는 방법이다. 이를 통해서 추세(Trend)를 제거하여 평균을 안정화 시키는 것이다. 계절성에 따른 분산의 변화는 계절 차분, log 변환이나 Box-cox 변환 등을 사용한다.
2. 시계열 데이터의 정상성(Stationarity)을 확보하기 위해 사용하는 '차분(Differencing)'에 대한 설명으로 가장 적절한 것은 무엇인가?
 1. 시계열 데이터의 분산만을 안정화시키고, 평균은 변화시키지 않는다.
 2. 데이터의 특정 주기성을 강조하여 계절성을 더욱 명확하게 드러내는 데 사용된다.
 3. 시계열 데이터에서 추세(trend)나 계절성(seasonality)과 같은 비정상적인 요소를 제거하여 평균과 분산을 안정화시켜 정상성을 확보하는 방법이다. ☒
 4. 이상치(outlier)를 자동으로 감지하고 제거하여 데이터의 품질을 높이는 데 기여한다.
 5. 데이터의 잡음(noise)을 제거하여 시계열을 평활(smoothing)하는 데 주로 사용된다.

일반적인 차분에 대한 설명이다. 계절성 차분으로 계절성을 제거 할 수 있다. 일반적인 개념으로는 쉬운 문제이지만 차분에 대한 이해가 높으면 어려운 문제
3. 시계열 데이터의 평활화(Smoothing) 기법에 대한 설명으로 가장 적절한 것은 무엇인가?
 1. 데이터의 추세(trend)나 계절성(seasonality)을 제거하여 시계열을 정상성(Stationarity)으로 변환하는 주된 목적을 가진다.

2. 시계열 데이터에 포함된 불규칙한 잡음(noise)을 줄여서, 데이터에 내재된 장기적인 추세나 주기적인 패턴과 같은 기본적인 움직임을 더 명확하게 파악할 수 있도록 돕는다. ✓
 3. 주로 미래 시점의 값을 정확하게 예측하기 위한 전처리 과정 없이 직접적인 예측 모델로 사용된다.
 4. 데이터의 비정상성 정도를 수치적으로 측정하여 통계적 검정의 기초 자료로 활용된다.
 5. 시계열 데이터의 모든 이상치(outlier)를 자동으로 감지하고 제거하는 유일한 방법이다.
4. 시계열 데이터에 대한 AR(자기회귀) 모델과 그 입력 시퀀스 길이(윈도우)에 대한 설명으로 가장 적절한 것은 무엇인가?
1. 입력 시퀀스 길이는 AR 모델이 예측 과정에서 외부 변수의 영향을 얼마나 많이 고려하는지를 나타낸다.
 2. AR(p) 모델에서 입력 시퀀스 길이(윈도우) p는 현재 시점의 값을 예측하기 위해 고하는 과거 관측값의 개수를 의미하며, 모델의 중요한 파라미터이다. ✓
AR(p) 모델에서 'p'는 모델의 차수(order)를 나타내며, 이는 현재 시점의 값(Y_t)을 예측하기 위해 과거 p개의 관측값($Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$)을 사용한다는 것을 의미합니다. 이 'p'가 바로 입력 시퀀스 길이 또는 윈도우 크기에 해당하며, AR 모델의 핵심 파라미터이다.
 3. AR 모델의 윈도우는 이동 평균(Moving Average) 모델에서만 사용되는 개념이다.
 4. AR 모델의 윈도우 길이는 미래 시점의 예측 오차를 줄이기 위해 동적으로 조절되는 값이다.
 5. AR 모델은 입력 시퀀스 길이(윈도우)와 관계없이 항상 모든 과거 데이터를 사용하여 미래 값을 예측한다.
5. ARIMA (AutoRegressive Integrated Moving Average) 모델에 대한 설명으로 가장 적절한 것은 무엇인가?
1. ARIMA는 '자기회귀(AutoRegressive)', '이동평균(Moving Average)', 그리고 '누적(Integrated, 차분)'의 세 가지 요소를 결합한 모델로, 비정상 시계열 데이터의 분석 및 예측에 널리 사용된다. ✓
 2. ARIMA 모델은 오직 정상성(Stationary)을 띠는 시계열 데이터에만 적용 가능하며, 비정상 시계열에는 사용할 수 없다.
 3. ARIMA 모델은 오직 선형 추세(linear trend)만을 예측할 수 있으며, 계절성이나 다른 복잡한 패턴은 모델링할 수 없다.
 4. ARIMA 모델은 주로 이미지 데이터의 패턴 인식에 사용되며, 시계열 데이터와는 관련이 적다.
 5. ARIMA 모델의 'I'는 이상치(Irregularity)를 의미하며, 모델이 자동으로 이상치를 감지하고 제거하는 기능을 포함한다.
6. 시계열 데이터 분석에서 '잔차 분석(Residual Analysis)'의 의미와 중요성에 대한 설명으로 가장 적절한 것은 무엇인가?

1. 잔차 분석은 모델이 시계열 데이터의 모든 체계적인 패턴(추세, 계절성 등)을 성공적으로 포착했는지 확인하고, 남은 잔차가 무작위적(white noise) 특성을 띠는지 검증하여 모델의 적합성을 평가하는 핵심 단계이다. 
- 잔차 분석의 가장 중요한 목적은 모델이 데이터의 중요한 정보를 얼마나 잘 설명했는지를 평가하는 것. 이상적인 잔차는 평균이 0이고, 일정한 분산을 가지며, 자기 상관관계가 없는 백색 잡음(white noise) 특성을 보여야 함. 잔차가 백색 잡음이 아니라면, 모델이 아직 설명하지 못한 패턴이 데이터에 남아있다는 것을 의미.
2. 잔차 분석은 시계열 모델의 입력 데이터를 평활화하여 노이즈를 제거하는 과정이다.
3. 잔차 분석을 통해 모델의 복잡도를 자동으로 최적화하여 과적합(overfitting)을 방지할 수 있다.
4. 잔차는 시계열 데이터에서 발견된 모든 이상치(outlier)의 집합을 의미한다.
5. 잔차 분석은 오직 모델의 예측 정확도(예: RMSE)를 계산하기 위한 수단이며, 모델의 통계적 특성 검증과는 무관하다.
7. 시계열 모델의 예측 성능을 측정하는 지표들에 대한 설명으로 가장 적절한 것은 무엇인가?
 1. 평균 절대 오차(MAE)는 예측 오차의 제곱 평균으로, 이상치에 민감하게 반응하여 모델의 안정성을 평가하는 데 주로 사용된다. 예측 오차의 절댓값 평균
 2. 평균 제곱근 오차(RMSE)는 예측 오차의 절댓값 평균으로, 오차가 커질수록 지표값이 선형적으로 증가하여 해석하기 용이하다. 예측 오차의 제곱 평균의 제곱근
 3. AIC(Akaike Information Criterion)는 시계열 모델의 예측 성능 지표 중 하나로, 값이 작을수록 예측 정확도가 높음을 의미한다. 모델의 적합도와 복잡성을 동시에 고려하여 최적 모델을 선택하는 데 사용
 4. R-squared (R^2)는 시계열 모델의 예측 성능을 측정하는 가장 보편적인 지표이며, 1에 가까울수록 예측 오차가 작음을 의미한다. R^2 는 시계열 모델에서 가장 보편적인 예측 성능 지표가 아님. R^2 가 1에 가깝다고 해서 예측 오차가 작다는 의미는 아님.
 5. 평균 절대 백분율 오차(MAPE)는 예측 오차를 실제값 대비 백분율로 나타내어, 서로 다른 스케일의 시계열 모델 성능을 비교할 때 유용하다. 