# 7장 - 시계열 분석

## 시계열 데이터

시간에 따라 변하는 데이터 (ex. 주가, 환율, 기온 등)

#### 시계열 데이터 형태 4가지

- 1. **불규칙 변동(irregular variation)** 시간에 따른 규칙적인 움직임이 없음. 우연적으로 발생하는 변동 (ex. 전쟁, 홍수, 파업 등)
- 2. 추세 변동(trend variation):

추세: 장기간에 걸쳐 지속적으로 증가/감소하거나 일정한 상태를 유지하려는 성향짧은 기간에는 추세 변동을 찾기 어려움. (ex. 국내총생산(GDP), 인구 증가율)

- 3. 순환 변동(cyclical variation) 일정한 기간을 주기로 순환적으로 나타나는 변동
- 4. 계절 변동(seasonal variation) 계절적 영향과 사회적 관습에 따라 1년 주기로 발생하는 것

#### 순환변동 vs 계절변동

계절변동은 고정된 주기(달, 분기, 주) 보통 1년 이내 반복

순환변동은 가변적 주기. 길이와 타이밍이 일정하지 않음. 1년 넘고 보통 2~3년 주기

#### 규칙적 시계열 vs 불규칙적 시계열

규칙적 - 트렌드와 분산이 변하지 않음

불규칙적 - 트렌드와 분산이 변함

→ 시계열 분석 - 불규칙적 데이터를 갖는 시계열 데이터에 기법을 적용해 패턴을 찾거나 예측

### 여러가지 모델들

1. AR 모델(AutoRegressive)[자기 회귀] - 이전 관측 값이 이후 관측 값에 영향을 준다는 아이디어에 대한 모형

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t$$

7장 - 시계열 분석

[좌측]  $X_t$  - 현재 시점/ [중앙] 과거가 현재에 미치는 영항을 모수( $\phi$ )에 과거 시점을 곱한 것 / [우측] $\epsilon t$  - 백색잡음(오차 항)

백색잡음(white noise) 패턴이 남아있지 않고 무작위로 야기되는 잡음

p시점에서 이전에 대이터에 의해 현재 시점의 데이터가 영향을 받는 모형

2. **MA모델 (Moving Average)[이동 평균]** - 트렌드가 변화하는 상황에 적합한 **회귀** 모 델

과거의 시계열 값을 사용하는 대신 과거의 오차를 활용해 회귀식을 세움

$$X_t = \mu + arepsilon_t + heta_1 arepsilon_{t-1} + heta_2 arepsilon_{t-2} + \ldots + heta_q arepsilon_{t-q}, arepsilon_t \ WN(0,\sigma^2)$$

3. ARMA(AutoRegressive Moving Average)[자기 회귀 이동 평균]

AR + MA (과거값의 선형 조합과 예측 오차의 선형 조합)

4. ARIMA (AutoRegressive Integreted Moving Average)[자귀 회귀 누적 이동 평균] AR + MA + I(Integrated- 차분)

차분 - 연속된 시점의 값 차이를 계산에, 추세나 느린 변동을 지워서 평균이 일정한 모습으로 만드는 변환

왜? - 랜덤워크, 추세 때문에 평균이 변하는 비정상 시계열을 평균이 거의 일정한 정상 시계열로 만들기 위해

랜덤워크가 뭔가? - 이전 값 + 우연한 충격으로만 움직이는 비정상 시계열 충격? - 예측 불가능한 새 정보(백색잡음 으로 가정)

# 순환신경망

RNN(Recurrent Neural Network) - 현재 입력뿐만 아니라 이전까지의 정보를 함께 사용해 순서를 가진 데이터를 처리하는 신경망

t시점에서 입력  $x_t$ 와 hidden state  $h_t$ -1을 받아 새 은닉생태  $h_t$ 를 만든다 → 이  $h_t$ 를 다음 은닉층에 입력으로 사용

입출력에 따른 유형

- 1. 일대다 : 이미지 켭션(image captioning)[이미지를 입력해 설명을 문장으로 출력]
  - a. 잠재코드 → 시퀀스 생성(문장 생성)/ 잠재코드 원본 데이터를 고정길이의 숫자로 압축(요약된 표현으로 더 쉽게 분류/생성/예측 하게 도와줌)
- 2. 다대일 문장입력 → 감성/문서 분류
- 3. 다대다 번역, 시퀀스 태깅(품사 태깅)

#### RNN계층과 셀

셀 - 시간 t에서 한번 상태를 업데이트 하는 계산 유닛

계층 - 하나의 셀을 모든 시점 1~t시간에 적용해 시퀀스 전체를 처리

#### 구조

- 1. 은닉층 이전 은닉층의 값과 현재 입력에 은닉층 가중치로 계산.
- 2. 출력층 DNN과 계산 동일
- 3. 오차 각 단계(t)마다 오차 특정(MSE)
- 4. 역전파 BPTT(BackPropagation Through Time) 각 단계마다 오차 측정하고 역전파
  - a. 기울기 소실 문제 발생 오래 전 시점으로 갈수록 여러 번의 곱셈을 거쳐 전 달되는데 이때 발생할 수 있음(tanh, sigmoid를 사용해서) → 긴 시퀀스에 서 오래된 정보는 기억하기 어려움

#### **LSTM**

망각 게이트, 입력 게이트, 출력 게이트를 은닉층에 추가해 기울기 소멸 문제 해결

**망각 게이트** - 과거 정보를 얼마나 기억할지 결정

**입력 게이트** - 현재 정보를 기억

과거 정보와 현재 데이터를 입력받아 현재 정보에 대한 보존량을 결정

메모리 셀 - 각 단계에 대한 은닉 노드

망각 게이트와 입력 게이터의 이전 단계셀 정보를 계산해 현재 단계 셀 상태를 업데이 트

출력 게이트 - 과거 정보와 현제 대이터를 사용해 뉴런의 출력 결정. 이전 은닉상태와 t 번째 입력을 고려해 다음 은닉상태 계산

게이트에서 얼마나 기억하고 이런걸 어떻게 정할까?

7장 - 시계열 분석

게이트들은 모두 시그모이드를 통과해 [0,1] 값을 가짐 학습된 가중치로부터 현재 입력 x\_t와 이전 은닉상태 h\_t-1을 보고 매 시점 계산됨 역전파 - 중단 없는 기울기(uninterrupted gradient flow)

#### **GRU**

reset gate, update gate 2개의 게이트만 사용

Reset Gate (망각 게이트)

이전 은닉상태의 값을 얼마나 활용할 것인지에 대한 정보

#### **Update Gate**

input, forget게이트와 비슷한 역할.

과거와 현재의 정보를 각각 얼마나 반영할지에 대한 비율을 구함

z → 현재 정보 반영할 수치 / 1-z → 과거 정보 얼마나 사용할지 반영

후보 은닉(후보군) - 과거 은닉층 정보를 그대로 이용하지 않고 리셋 게이트의 결과를 이용해 계산

최종 은닉층 계산 - 업데이트 게이트 결과와 후보군 결과를 결합해 현시점의 은닉층 계산 → 장기 의존 완화/ 계산 빠름

LSTM보다 간단한 구조

## 양방향 RNN

RNN은 이전 시점의 데이터를 참고해서 예측을 했지만 실제 문제에서는 미래 시점의 데이터에 힌트가 있는 경우도 많다. 과거 뿐만 아니라 이후 시점의 데이터도 활용해 예측하는 것을 양방향 RNN

구조 - 총 2개의 메모리 셀

이전 시점의 은닉상태를 받아 현재의 은닉상태를 계산하는 메모리 셀

다음 시점의 은닉상태를 받아 현재의 은닉상태를 계한하는 메모리 셀

LSTM, GRU에도 같은 개념

7장 - 시계열 분석 4