데이터 분석 일반

예측분석

Predictive Analytics

*

지금 알고 있는 걸 그때도 알았더라면 by Kimberly Kirberger, 류시화 역

지금 알고 있는 걸 그때도 알았더라면
내 가슴이 말하는 것에 더 자주 귀 기울였으리라.
더 즐겁게 살고, 덜 고민했으리라.
금방 학교를 졸업하고 머지않아 직업을 가져야 한다는 걸 깨달았으리라.
아니, 그런 것들은 잊어 버렸으리라.
다른 사람들이 나에 대해 말하는 것에는
신경쓰지 않았으리라.
그 대신 내가 가진 생명력과 단단한 피부를 더 가치있게 여겼으리라.
더 많이 놀고, 덜 초조해 했으리라.
진정한 아름다움은 자신의 인생을 사랑하는 데 있음을 기억했으리라.
부모가 날 얼마나 사랑하는가를 알고
또한 그들이 내게 최선을 다하고 있음을 믿었으리라.
사랑에 더 열중하고
그 결말에 대해선 덜 걱정했으리라.
설령 그것이 실패로 끝난다 해도
더 좋은 어떤 것이 기다리고 있음을 믿었으리라.

아, 나는 어린아이처럼 행동하는 걸 두려워하지 않았으리라.
더 많은 용기를 가졌으리라.
모든 사람에게서 좋은 면을 발견하고
그것을 그들과 함께 나눴으리라.
지금 알고 있는 걸 그때도 알았더라면
나는 분명코 춤추는 법을 배웠으리라.
내 육체를 있는 그대로 좋아했으리라.
내가 만나는 사람을 신뢰하고
나 역시 누군가에게 신뢰할 만한 사람이 되었으리라.
입맞춤을 즐겼으리라.
정말로 자주 입을 맞췄으리라.
분명코 더 감사하고,
더 많이 행복해 했으리라.
지금 내가 알고 있는 걸 그때도 알았더라면.

* 목차

예측분석 이란?

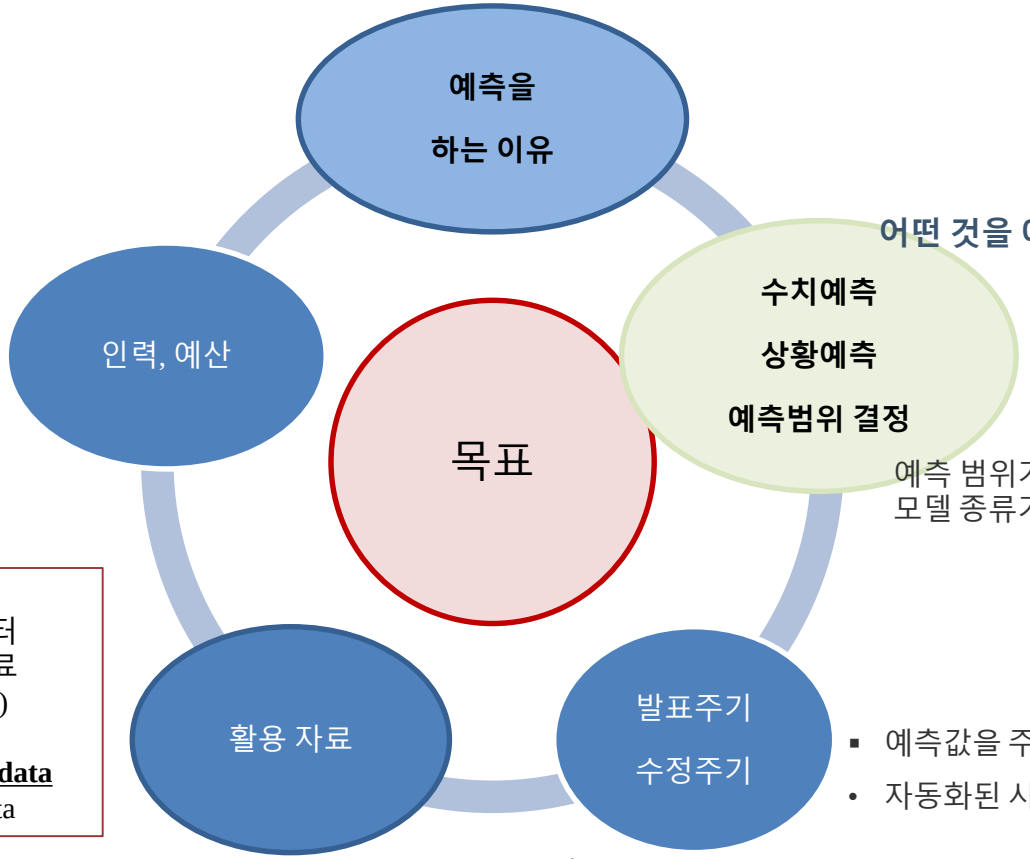
- 보다 나은 의사결정을 위해서 데이터로부터 학습하여 어떤 사안에 대한 미래행위를 예측하는 기술
- 특정한 Event 가 발생된 원인의 규명은 물론이고 향후에 발생될 가능성을 예측하고 이에 적합한 대응 행동을 제시해줌으로써 데이터 기반의 합리적이고 효율적인 의사결정을 가능하게 해주는 미래형 분석분야
- 최적화 예측 통계분석부터 최적화에 이르기까지 예측분석 (Predictive Analytics) 영역은 빅데이터를 도입하는 기업들이 중점을 두어서 반드시 갖추어야 하는 분석 역량
- 예측분석→ 새로운 경제적가치 창출
- ***(efficient market hypothesis)***

예측분석 주제영역

- 무엇을 예측할 것인가?
- 어떻게 활용할 것인가

모델 설계자는 모델을 통한 직관적 해석(good stories)과 예측력 높은 결과(good forecasts) 도출을 목표로 하며, 이에 따라 이론적 적합성 뿐 아니라 예측력을 모델의 평가 기준으로 활용

예측분석 : 목표



예측 범위가 가장 중요한 지에 따라 다른 모델 종류가 필요할 것

데이터 수집
- 정형데이터
- 시계열자료

(시계열 빈도)
- **월중data**
- **일별/일중data**
- 텍스트 data

- 예측값을 주기적으로 발표
- 자동화된 시스템 만들것 인가?

시계열 패턴
- 추세적요인
- 계절적 요인
- 순환적요인

- 데이터요약/시각화
- 확률과정
- 시계열모형

- Scatterplots
- lagged scatterplots
- 상관계수
- Auto correlation

예측분석 : 목표

- **예측기간 : 단기, 중기, 장기예측**
- **주식 가격 예측은?**

구분	적용
단기 예측	▪ 매출액 예측, 생산계획 • 카드회사 회원들의 가입정보를 통해 연 매출액을 알아 맞히는 것 • 전산 DOWN 예측 ▪ 환율예측, 주가 예측 ▪ BOK 경제 예측 , 통계청 인구예측, 기상청 기후예측, 선거예측
중기	원자재 구입, 신규 채용, 장비나 기계 구입 등 미래 자원 공급을 결정하는데 필요
장기	▪ 전략적 계획수립에 활용 ▪ 시장 기회, 환경 요인, 내부 자원을 고려하여 결정

예측의 목적 및 예측 기간의 길이를 고려하여 적절한 예측 기법 선정

- 단기 예측: 3개월 미만
- 중기 예측: 3개월~2년
- 장기 예측: 2년 이상

일반적으로 예측 기간이 길어질수록 예측의 정확도 낮아짐

점 예측과 구간 예측

- 점 예측(Point Forecast): 예측의 목적이 변수의 실제값 추정
- 구간 예측(Interval Forecast): 대략의 범위를 추정

예측분석 : 목표

예측의 목표 및 대상 설정

예측 목표 설정→ 대상 변수 설정→ 예측치 추정의 주기(월별, 분기별) 및 예측 기간의 길이(단기, 중기, 장기) 결정

자료 수집 및 자료의 특징 분석

변수에 대한 자료 수집 → 자기상관계수, 다른 변수와의 교차상관계수 등 기술통계량 계산 및 적합한 예측 모형 선정을 위해 전체적인 패턴을 나타내는 그래프 이용 시각화 (시간 그래프)

예측 모형의 선정 및 추정

- 자료의 특성에 맞는 **예측 모형 선정→ 추정 →모형의 적합성 검정**
- 미래 예측치의 정확도에 대한 사전 평가 불가능하므로 추정된 예측 모형을 이용해 실험적 예측치 도출
- 사용할 수 있는 관측치 중 마지막 일부를 제외하고 모형 추정→실제값을 알 수 있는 기간에 대한 사후적 예측치 도출 (Ex post Forecast)→실제값과 추정치 비교하여 적합성 기준으로 최적 모형 선택

예측치 분석 및 예측 모형 평가

최적 모형을 이용하여 사전적 예측치(Ex ante Forecast) 도출 → 이후 시간이 경과한 후의 실제값과 비교→예측 모형에 대한 재평가 → 필요한 경우 새로운 예측 모형 선택

구조변화에 대한 고려

- 예측치 추정은 기본적으로 과거의 자료가 미래 예측에 유용한 정보를 가지고 있다고 가정
- 가정이 성립하지 않는 경우, 즉 예측치를 추정하고자 하는 변수의 구조변화(Structural Break)가 나타나서 과거의 움직임과 다른 패턴을 나타낸다면 구조모형을 반영하지 않으면 예측오차가 크게 나타남
- Chow 검정, Quandt-Andrew 검정, Bai-Perron 검정 등으로 확인

예측분석 방법 구분

구분	내용	비고
주관적 방법	전문가들에 의한 경험, 식견	델파이방법 주관적
객관적 방법	모형에 의한 예측 이론적 모형 vs 비이론적 모형 이론적모형 : 경제이론에 근거하여 변수를 설정하는 모형	▪ **이론적 모형 (model with theory)** - 연립방정식 모형 ▪ **비이론적 모형** (atheoretical model) - 시계열모형 (모형 식별→ 모형 추정→ 모형 검증(진단) - **ARIMA** - **VAR** - **Machine Learning** - **Neural Network**
시나리오 모형	미래 일어날 시나리오 설정	

예측분석 방법 구분

구분	분석방법		
정량적 방법	**시계열분석 : 추세분석 (Trend Analysis)**	이동평균법	과거 일정기간(n)일 이동평균 값으로 예측 → 추세변동, 계절변동, 순환변동 등의 요인이 없을 때 적용 가능
-**과거 데이터가 충분히 존재하는 경우** - 과거 패턴의 몇 가지 양상이 미래에도 계속 될 것이라고 가정하는 것이 합리적일 때		지수평활법	▪ **최근의 실적치에 가장 큰 가중치를 부여하고** 오래된 데이터의 가중치는 지수함수적으로 적게 적용하는 것 ▪ 단기예측에 주로 활용
		최소제곱법 (Least-squares Method)	- 실제치와 예측치와의 편차 제곱의 총합이 최소가 되는 **추세선(회귀식)을 도출 후에** 이를 통해 미래 수요를 예측
(사례) 주가 예측, 매출액, 수익 예측	인과형 분석 (theoretical) 모델링	회귀분석	단순, 다중회귀분석, 선형회귀분석, 비선형회귀분석, 직관적 해석이 가능
		계량경제모형	경제이론에 입각하여 방정식을 통한 예측 인과관계를 추정할 만한 충분한 과거 자료가 있는 경우 사용
	시뮬레이션 방법		
(머신러닝)	상관관계에 기반 축약형 (reduced-form)		빅데이터에 담긴 정보를 추출•가공하여 경제변수를 예측 틀 제공 검증 데이터(testing set)에서의 결과 우수 (1) 결과의 직관적 해석이 어렵다 (2) 실제 경제예측에서 어떤 알고리즘도 뚜렷한 우월성을 지니고 있다는 것이 입증되지 않았다는 점
정성적방법	델파이 조사법 (Delphi Technique)	- 여러 전문가의 판단을 조직적으로 수렴시켜 일치된 의견이나 예측을 도출하는 기법 시간과 비용이 과다 소요 , 통계적 지식 불필요 . → **주관적 예측방법**	
(신제품 예측, 과거 데이터가 없는 경우)	시장조사법	- 정성적 기법 중 가장 계량적이고 객관적인 방법 - 소비자로부터 직접 수요에 관한 정보를 얻으려는 방법 - 조사대상자 : 일반소비자, 특정고객, 도소매업자 - 조사 방법 : 설문조사, 면접, 소비자패널	
	전문가 의견 이용		
	패널조사		
	판매원 의견		

예측모델링 기법 : 통계기법과 ML

금융시장 예측을 위해 사용되는 통계기법

통계기법	요약 설명
시계열 분석 (Time Series Analysis)	▪ 시계열 분석은 **시간에 따라 변화하는 데이터를 분석**하여 미래의 값을 예측하는 데 사용 ▪ 금융 시장에서 주가, 금리, 환율 등의 예측에 주로 사용
	▪ **이동 평균(Moving Average)**: 최근 데이터의 평균을 사용하여 미래 값을 예측하는 기법으로, 데이터의 변동성을 줄이고 추세를 파악한다. ▪ **ARIMA (AutoRegressive Integrated Moving Average)**: 과거 데이터를 기반으로 시간적 패턴을 찾아 미래를 예측하는 시계열 모델 ▪ **GARCH (Generalized Autoregressive Conditional Heteroskedasticity)**: 금융 데이터의 변동성을 예측하는 데 사용되며, 주가 및 옵션 가격 예측에 유용
회귀 분석 (Regression Analysis)	▪ **변수 간의 관계를 통해 미래 값을 예측**하는 기법 ▪ 독립 변수를 바탕으로 종속 변수(목표 변수)를 예측하는 데 유용하며, 특히 주가와 같은 연속형 데이터 예측에 자주 사용
	▪ **선형 회귀(Linear Regression)**: 두 변수 간 선형 관계를 기반으로 미래 값을 예측 ▪ **다중 회귀(Multiple Regression)**: 다수의 독립 변수를 사용하여 목표 변수의 변동성을 설명하고 예측한다.
판별 분석 (Discriminant Analysis)	▪ 판별분석은 **여러 그룹 간 차이를 분석하고 분류**하는 기법 ▪ 금융에서는 특히 신용 리스크 분석, 고객 세분화 등에 활용
	▪ **선형 판별분석(LDA)**: 금융 상품의 수익성과 관련된 여러 변수를 바탕으로 안정적이고 위험이 큰 상품을 분류한다. ▪ **이차 판별분석(QDA)**: 각 클래스 간의 비선형 관계를 반영하여 더욱 정교하게 분류한다.
주성분 분석 (PCA, Principal Component Analysis)	▪ 주성분 분석은 **고차원 데이터를 저차원으로 축소**하여 데이터의 주요 패턴을 파악하는 기법 금융 시장에서는 다수의 경제 지표나 자산에 대한 데이터를 분석할 때 유용하게 사용
	▪ **활용 예시**: 수십 개의 주식 종목을 포함한 포트폴리오에서 주요 요인(주성분)을 추출해 전체 시장의 움직임을 설명할 수 있다.

예측모델링 기법 : 통계기법과 ML

금융시장 예측을 위해 사용되는 통계기법

머신러닝 기법	요약 설명
LSTM (Long Short-Term Memory)	▪ LSTM은 **시계열 데이터를 처리**하는 데 유리한 순환 신경망(RNN)의 한 종류로, 장기 의존성을 학습하는 데 특화되어 있다. 주가나 금리와 같은 금융 데이터의 변동성을 예측하는 데 효과적이다.
	▪ **활용 예시**: 주가 예측에서 과거 데이터뿐만 아니라, 특정 패턴이 여러 시점에 걸쳐 나타나는 경우 장기적인 데이터를 고려하여 예측을 수행한다.
랜덤 포레스트 (Random Forest)	▪ 랜덤 포레스트는 **다수의 결정 트리**를 조합해 예측하는 앙상블 기법으로, 과적합을 줄이면서 높은 정확도로 예측할 수 있다. 주가 예측, 신용 리스크 평가, 고객 세분화 등에 사용된다.
	▪ **활용 예시**: 고객의 신용 이력을 여러 특성으로 분석하여, 특정 고객의 신용 리스크를 평가하고 분류
부스팅 기법 (Boosting Methods)	▪ 부스팅은 **약한 학습기**를 결합해 강한 학습기를 만드는 앙상블 기법 ▪ 금융 시장에서는 XGBoost, LightGBM, AdaBoost 등이 주가 예측, 리스크 평가, 투자 전략 개발 등에 활용
	▪ **활용 예시**: 주가의 변동성을 예측할 때 여러 부스팅 알고리즘을 적용하여 예측 성능을 높일 수 있다.
강화 학습 (Reinforcement Learning)	▪ 강화 학습은 **에이전트가 보상을 최대화하도록 학습**하는 기법으로, 트레이딩 봇이나 포트폴리오 관리와 같은 자율적인 금융 의사결정 시스템에 유용하게 사용
	▪ **활용 예시**: 주식 매매에서 강화 학습을 통해 트레이딩 전략을 학습하고, 주가 변동에 따른 최적의 매수/매도 시점을 찾아내는 트레이딩 봇을 개발할 수 있다.
서포트 벡터 머신 (SVM, Support Vector Machine)	▪ SVM은 **최적의 분류 경계**를 설정하여 데이터를 분류하거나 예측하는 모델 ▪ 금융 시장에서는 가격 변동성 예측이나 자산 분류, 투자 추천 시스템 등에 사용
	▪ **활용 예시**: 주가 변동에 영향을 미치는 지표를 기반으로 특정 자산이 상승할지 하락할지를 분류한다.
클러스터링 (Clustering)	▪ 클러스터링은 **유사한 특성을 가진 데이터를 그룹화**하는 비지도 학습 기법 ▪ 금융에서는 고객을 세분화하거나, 주식 종목 간의 유사성을 파악하여 포트폴리오 구성에 활용
	▪ **K-평균(K-Means)**: 데이터를 K개의 군집으로 나누어 유사한 특성을 가진 주식 종목을 그룹화한다. ▪ **계층적 군집화(Hierarchical Clustering)**: 자산이나 고객을 계층적으로 그룹화하여 분석할 수 있다.

머신러닝 분류 시계열 예측 모델

일반 예측 모델	Support Vector Machine (SVM)
	Random Forest (RF)
	Deep Feed-forward Neural Network (DNN)
	Softmax
시계열 예측 모델	Long Short Term Memory(LSTM)
	stateful Long Short Term Memory (stateful LSTM)
	Timedistributed Mode
	Convolution LSTM

시계열 예측을 위한 LSTM 기반 딥러닝: 기업 신용평점 예측 사례 : 이현상* . 오세환* 2020

구분 : 회귀와 분류

Regression

Multiple Linear Regression
k-NN
Decision Tree Regression
Neural Networks

Classification

Logistic Linear Regression, Discriminant Analysis
k-NN, Naïve Bayese
Decision Tree Classifier
Neural Networks
SVM
Random Forest AdaBoost

Shallow learning

Scikit-Learn

- Neural Network
- CNN (Convolutional Neural Network)
- RNN(Recurrent Neural Network)
- LSTM(Long Short-Term Memory)

Deep learning

TensorFlow/ Keras

예측분석 절차

예측분석 절차

모형의 식별 → 모형의 추정 → 모형의 진단

기획단계 문제정의 정보수집 → 데이터 준비단계 EDA → 예비분석 예측모델 생성 → 모델 검증 → 모델 적용-보완 과정

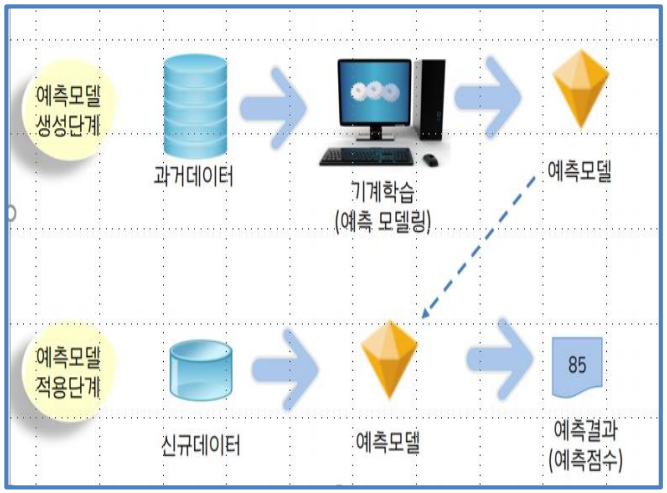
데이터 탐색 준비

- 일관된 패턴이 존재 여부?
- 의미 있는 추세(trend)가 존재 여부? 계절성(seasonality)이 중요 여부?
- 경기 순환(business cycle)이 존재한다는 증거 여부?
- 데이터에 전문적인 지식으로 설명할 수 있는 어떤 특이점 여부?
- 분석에 사용할 수 있는 변수 사이의 관계가 얼마나 강한지 여부?

모델 선택

- 회귀(regression) 모델, 지수 평활(exponential smoothing) 기법,
- Box-Jenkins ARIMA 모델, 계층적(hierarchical) 예측,
- neural network, Vector Auto Regression

예측 모델 생성과 적용



예측분석 방법 선택

예측 방법 선택 프로세스

data간에 순서 관계가 존재여부?

존재 O

Time series analysis
→ Statsmodels

존재 X

비-시계열 데이터 예측
Simple, K-fold, Holdout

예측하려는 데이터와 예측에 사용되는 데이터가 같은 그룹인가?

같은 그룹의 데이터를 예측 경우
정상적

1) 안정적 시계열 모형

- 자기회귀 (Auto Regressive, AR) 모형
- 이동평균 (Moving Average, MA) 모형
- ARMA (Auto Regressive Moving Average) 모형
- 벡터자기회귀 (VAR : Vector Auto Regression) 모형

2) 불안정적 시계열모형

- 확률보행(Random Walk) 모형
- ARIMA (Auto Regressive Intergrated Moving Average) 모형

3) 비선형 시계열모형

- ARCH (Autoregressive Conditional. Heteroskedasticity)
- GARCH (Generalized Autoregressive Conditional. Heteroskedasticity)
- EGARCH (Exponential Generalized Autoregressive Conditional. Heteroskedasticity)
- TAR (Treshold Auto Regressive)

특정 그룹의 데이터로 다른 그룹의 데이터를 예측하는 경우

예측하려는 데이터의 값의 크기를 비교 가능 여부?

비교 가능한 값의 경우
→ 회귀분석

비교 불가능한 경우
→ 분류

분류→

*예측모델링 구분

예측 모델링

모델링은 모델 설계자가 모델을 통해 답하고자 하는 질문과 연구 목적을 바탕으로 이루어짐

이론적 (theoretical) 모델링

(1) **인과적 질문에 대한 수학적 표현**

(2) 모델에 깔려 있는 가정을 상호 소통할 수 있도록 하는 정확한 수학적 언어를 제공

(3) 답하고자 하는 질문을 풀 수 있는 체계적인 틀을 제공해야 한다는 조건을 충족해야 함

사례: "미국 연준의 기준금리 인상이 신흥국 자본유출을 야기하는가?

축약형 모델링

변수들 간의 **상관관계에 기반하여** 경제변수를 설명·예측

: 상관관계란 두 변수의 선형 동행성(linear co-movement)에 기반한 유추

: 상관관계에만 기반해서는 인과관계를 도출해낼 수 없으며 인과관계는 인과적 가정에 의해 정의 됨.

*예측모델링 구분

머신러닝 - 빅데이터 시대의 축약형 접근법

- 머신러닝은 설계된 알고리즘을 따라 빅데이터에 담긴 정보를 학습하여 예측력 제고에 유용한 관심변수 정보를 추출하는 귀납적 러닝 메커니즘
- 머신러닝은 변수 간 관계 등과 같은 모델러(분석자)의 사전적 지식(prior)을 요구하지 않으며, 주어진 알고리즘 틀을 이용해 데이터라는 Blackbox에 담긴 정보를 학습

머신러닝 알고리즘은 빅데이터에 담긴 정보를 학습하는 방법과 변수들에 가중치를 부과하는 방법에 따라 구별됨

- 머신러닝의 장점은 다양한 구조를 가진 데이터를 다룰 수 있게 해주며, 기존 축약형 모델링의 문제점으로 지적된 변수의 자의적 선택 문제와 과적합(overfitting) 문제를 최소화
- 머신러닝은 빅데이터를 학습•요약하는 데 초점이 있는지(clustering, classification, dimension reduction) 또는 특정 종속변수의 예측성 제고에 도움이 되는 설명변수들을 추출하는 것인지에 따라 크게
- ① 지도학습 (Unsupervised Learning)과 ② 비지도학습(Supervised Learning)으로 분류

빅데이터 시대의 예측 도구, Machine Learning의 올바른 활용법 (포스코 경영연구원)

*예측모델링 구분

머신러닝 - 빅데이터 시대의 축약형 접근법(한계점)

빅데이터 방법론은 기존 데이터에 담긴 정보를 학습하는 데 초점을 맞추므로 데이터 자체가 편향된(biased) 경우 편향된 예측결과로 이어짐

(사례)

○ Amazon은 기존 직원들의 이력서에 담긴 정보를 머신러닝 기법으로 학습하여 인사채용에 활용하였지만 성차별적인 결과를 도출한다는 점을 발견하고, AI 채용 절차를 폐지 (Reuters, 2018.10.10일자)

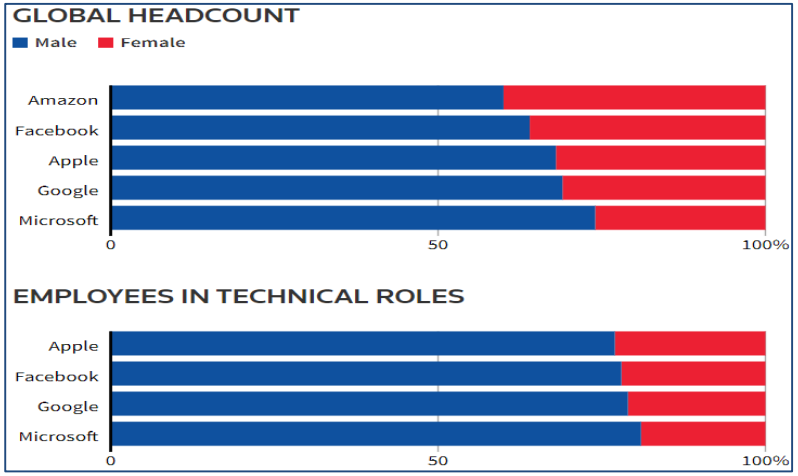
- 테크 산업은 남성 엔지니어의 비중이 큰 분야인데, 기존 인력 풀의 정보 학습에 기반한 머신러닝 알고리즘이 이력서에 쓰인 여성 관련 단어에 페널티를 부과하였고, 이는 기존의 남성지배적 구조를 고착화하는 결과를 초래 - 맥락에 대한 이해가 뒷받침되지 않은 기계적 학습의 폐해를 단적으로 드러내는 사례로서 모델러의 조정적 역할이 더욱 강조됨

빅데이터 시대에 모델러(분석자)에게 요구되는 자질

맥락에 맞는 데이터 활용 능력

○ 알고리즘은 정형화되고 슈퍼컴퓨터를 활용한 실시간 머신러닝도 가능해졌지만, 핵심은 문제 상황에 어떻게 적용하고 해석할 것인가임

빅데이터 시대의 예측 도구, Machine Learning의 올바른 활용법(포스코경영연구원)



https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

*예측모델링 구분

모델의 예측력을 현 시점에서 평가하는 것은 현재 확보 가능한 데이터를 기반으로 할 수밖에 없다는 근본적인 한계가 있음

○ 미래 추정값을 산출할 수는 있으나 실제값과의 비교는 불가능하므로, 경제예측력 평가에 대한 학술적 논의는 현재까지 주어진 데이터를 활용해 이루어짐

- 현재 활용 가능한 샘플을 모델 fitting을 위한 in-sample
- 예측력 평가를 위한 out-of-sample로 나누어 모델의 예측력을 평가
- 실제 예측력 평가는 사후적(ex-post)으로 이루어질 수밖에 없고, 현재까지 확보 가능한 데이터로 순위를 매긴 예측력 평가가 미래에도 유효한지는 또 다른 문제

out-of-sample 기간 동안 가장 작은 예측오차 값을 산출해내는 모델을 예측력이 가장 우수 모델로 선택

적시성이 강조되는 산업현장의 경우 시시각각 변하는 현장의 정보와 모델 설계자의 통찰력
• 직관을 반영한 정성적 평가도 큰 부분 담당
○ "economic forecasting is an art, not a science"라는 비판은, 근본적인 한계를 지적하며 과학적 접근법에 결합되는 모델 설계자의 직관적 조정의 중요성을 강조

예측력을 강화하기 위해서는 시시각각 변하는 정치경제적 상황과
각 산업의 특수성을 고려한 **미세 조정(fine-tuning)**이 필요
- 빅데이터를 활용할 수 있는 프레임이 주는 장점을 활용하되,
- 해당 알고리즘의 장단점을 잘 인지하여 실시간 데이터와 모델을 업데이트할 필요

빅데이터 시대의 예측 도구, Machine Learning의 올바른 활용법(포스코경영연구원)

시계열 data

시계열 데이터 고유한 특성

1) 시간 순차성(Time Step)

: 시간축에서 직접 추출 가능하며 시작부터 끝까지 일정시간 간격으로 측정된 년, 월, 일, 시간 특성이 대표적

2) 지연 값(Lag)

: 관측값에 시간 차이로 발생되며 현재 관측값들은 이전 관측값들로 표현

→ 자기상관(Auto Correlation) 이 높아짐.

3) 시간의 종속성 : 통계분석의 가정에 위배 독립성 위배

시계열 패턴 (변동 요인)

주파수 영역(정보) (frequency domain)	**주기적으로 반복되는 정보**	**추세요인 (trend)**	**형태가 오르거나 내리는 장기적 추세, 선형, 이차식, 지수 형태**
		계절요인 (seasonality)	일정 주기 마다 자료가 변화 - 제거 : 연데이터 활용, 전년동기대비, 이동평균 등
		순환요인	특정한 경제적, 자연적 이유없이 알려지지 않은 주기로 자료가 변화 – 경기순환에 따라 반복하는 변동
		불규칙요인	위 세 가지의 요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인
시간 영역(정보) (time domain)	시간에 따라 전개되는 정보		

시계열 data

시계열 요인 특성

구분	내용
추세 요인 (Trend factor : T_t)	인구의 변화, 자원의 변화, 자본재의 변화, 기술의 변화 등과 같은 요인들에 의해 영향을 받는 **장기 변동**
계절 요인 (Seasonal factor : S_t)	**12개월(1년)의 주기**를 가지고 반복되는 변화
불규칙 요인 (Irregular facto : I_t)	우연적 요인에 의해 발생되는 변동
순환 요인 (Cycle factor : C_t)	경제활동의 팽창과 위축과 같이 불규칙적이며 반복적인 **중기 변동요인**

시계열 구성요인 간의 결합 방식에 따른 구분

구분	가정	표현	사용
가법모형 (additive model)	구성요인 간 **독립적**이라고 가정	가법모형 = 추세 요인 + 순환 요인 + 계절 요인 + 불규칙 요인	자료가 음(-)의 값을 포함하고 있는 경우
승법모형 (multiplicative model)	구성요인 간 독립적이지 않고 상호작용을 한다고 가정	승법 모형 = 추세 요인× 순환 요인 × 계절 요인× 불규칙 요인	시간에 따라 요인이 비례적으로 증가하는 경우 사용
준가법형 (pseudo- additive model)		$y_t = (T_t + C_t) \times (S_t + I_t - 1)$	시계열자료의 일정기간 월 또는 분기가 0 에 가까운 값을 갖는 경우 적정

시계열 data : 종류

시계열 자료 종류 데이터 생성확률과정(DGP:Data Generating stochastic Process)의 한 표본경로(Sample Path)에 대한 관측치

선형(linear)시계열 모형 vs 비선형(nonlinear) 시계열 모형

1) 선형 시계열모형(Linear Time Series Models)
- 선형 시계열 모형은 변수들이 선형적으로 결합되어 데이터를 설명하는 모델
- 이해가 쉽고, 분석 결과가 안정적이지만 설명에 한계 → AR, MA, ARMA, ARIMA

2) 비선형 시계열모형 (Nonlinear Time Series Models)
- 비선형 시계열 모형은 데이터가 비선형적인 동적 패턴을 가지는 경우에 사용
- 다양한 설계가 가능, 선택된 모형을 설명하기 어려움.

특징	선형 시계열 모형	비선형 시계열 모형
관계	선형 관계	비선형 관계
복잡성	상대적으로 단순	복잡
해석 가능성	높음	낮음
데이터 패턴 설명 능력	제한적	복잡한 패턴 설명 가능
예시 모형	AR, MA, ARIMA	ARCH, GARCH, TAR, Neural Networks

Nonlinear Time Series Models
- ARCH 모형 (Autoregressive Conditional Heteroskedasticity)
- GARCH 모형 (Generalized ARCH)
- Threshold 모형 (TAR, Threshold AutoRegressive Models)
- 비선형 신경망 모형 (Neural Network Models)

선형시계열 모형 : 비정상성 시계열 data vs 정상성 시계열 data

- **Weakly Stationary :1차, 2차 적률 만족**
- **Strongly Stationary : 1차, 2,차, 3차,4차 적률 만족**

- **비정상성** 시계열 data : **시간의 변화에 영향을 받는 자료**→ 시계열 분석을 실시할 때 다루기 어려운 자료로 대부분의 시계열자료가 이에 해당
- **정상성** 시계열 data : 비정상 시계열을 변환해 다루기 쉬운 시계열 자료로 변환한 자료

定常性

→ **시간에 변해도 통계적 속성(평균, 분산, 공분산)이 일정한 시계열 데이터 : Weakly Stationary**

시계열 분석 : 종류

시계열 자료 종류

정상성(stationary)시계열 중요성

1) 예측 가능성: 정상 시계열은 시간에 따라 일정한 통계적 속성을 가지므로, 미래 값을 예측하기 위한 모델링이 용이

2) 모델의 단순화: 정상 시계열 데이터를 사용하면 시계열 모델을 더 단순하게 만들 수 있으며, 모델의 안정성과 예측력을 향상시킬 수 있다.

3) 통계적 추론의 용이성: 데이터가 정상성을 만족하면, 통계적 추론과 가설 검정이 더 간단하고 정확해진다.

➔ ***정상성시계열은 항상 그 평균값으로 회귀하려는 경향이 있으며, 그 평균값 주변에서의 변동은 대체로 일정한 폭을 가짐***

定常性 시계열 자료가 왜 필요한가?

- **통계분석(가정)** : 정규성, 등분산성, 독립성
- **시계열 분석** : 정상성을 만족해야 분석 결과에 대해 신뢰

시계열의 평균과 분산이 일정해야 시계열 값을 예측할 수 있기 때문

시계열 분석에서 시계열 자료는 시간에 따른 확률 과정 (stochastic process)에서 실현된 값이라는 전제.

(만약)

시간의 흐름에 따라서 이 확률분포가 크게 변동하는 경우, 그에 따른 실현값들의 평균, 분산 등 변화가 의미가 없기 때문에 ➔ 적어도 평균과 분산이 다루고자 하는 확률 과정을 설명하기에 문제가 없도록 하기 위해 필요한 조건이 바로 定常性 !!

시계열 분석 : 종류

시계열 자료 종류

시계열 자료를 변환을 취해 정상성(stationary) 자료로~

1. 차분 (Differencing)

- **1차 차분** ; $\Delta y_t = y_t - y_{t-1}$
- **2차 차분** : $\Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$

차분 후의 plot을 보고 1차 차분이 적절할지? 2차 차분이 적정할지? 여부를 확인

2. 로그 변환 또는 제곱근 함수 → power transformation

1) 로그 변환 : y_t → $\ln(y_t)$

2) 제곱근 함수 : y_t → $\sqrt{y_t}$

3. 로그 차분

: $\Delta \ln(y_t) = \ln(y_t) - \ln(y_{t-1})$

정상성 검증 방법

1. 그래프로 파악하는 방법

- ACF(Auto Correlation Function)을 이용하는 것
- 시계열 그래프

2. 통계적인 검정 방법단위근 검정 (Unit Root Test)

- Dicky-Fuller 검정
- ADF 검정 (Augmented Dicky-Fuller Test)

시계열 분석 : 종류

시계열 자료 종류 : 선형추세분석

확정적 시계열 모형 (Deterministic Time Series Model)

$y_t = \beta_0 + \beta_1 t + \eta_t$
η_t : ARMA 과정

○ 방법: 시계열 자료의 **확률적 변동요인을 제거하고 확정적 추세 요인을 도출하여 예측**
확정적 추세(deterministic trend)에는 **추세의 기울기가 시간에 따라서 변하지 않을 것이라는 가정**
○ 자료: 과거자료가 충분하지 않지만 뚜렷한 추세를 나타내는 경우 단기예측 가능
○ 특징: 이해가 쉽고 비용이 적게 드나 확정적 변동요인의 구조변화가 나타나지 않는 한 예측의 정확도는 높음

확률적 시계열 모형 (Stochastic Time Series Model)

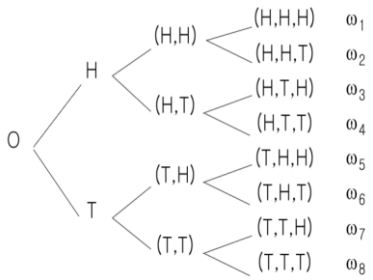
$y_t = \beta_0 + \beta_1 t + \eta_t$
η_t : ARIMA 과정

○ 방법: 시계열 자료의 과거 패턴을 기초로 ARIMA 종류의 시계열 모형을 추정하여 예측에 사용

○ 자료: 주어진 시계열 자료 자체에 대한 정보는 많은 반면, 다른 변수와의 상관관계가 미약하거나 불분명한 경우 사용 ○ 특징: 경제 이론을 도입되지 않는 반면 통계적 기법이 많이 사용되는데 저비용에 전환점 추정 등 변수 움직임을 보는데 유용
확률적인 추세(stochastic trend)는 추세의 기울기가 변할 수 있고 추정된 증가량에는 과거 기간 동안 평균 증가만 가정하고, 미래에 나타날 성장률일 필요는 없다.
예측구간이 미래 성장에 있어서 더 큰 불확실성을 허용하기 때문에 확률적 추세(stochastic trend)로 예측하는 것 더 안전 → **오차에 정상성이 나타나지 않기 때문에 예측구간이 훨씬 넓다.**

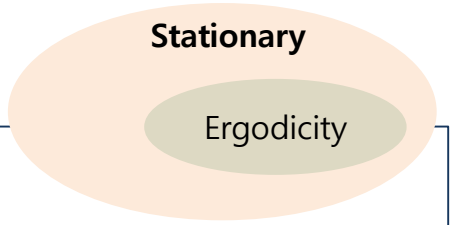
시계열 분석 : 입문

확률 과정 , 확률변수, 표본 경로



- 확률과정: 시간과 표본의 함수 X (t , ω)
- 확률변수 : 표본의 함수 X (ω)
- 표본경로 : 특정 표본에 대한 시간의 함수X (t , ω_2)

확률변수 : 사건 ω 함수
$X(\omega_1) = 3$, $X(\omega_2) = 2$,...



Ergodicity정의

- 표본경로를 통한 data 반복 생성이 불가능한 **금융시계열의 경우"Ergodicity"가정 필요**
Ensemble average를 사용하기 위해서는 무수히 많은 표본경로 생성이 필요하다.
그러나 금융시계열에서는 어려움
→ Ensemble average 대신에 **Time average를 이용하여 stationary 자료 계산 가능**
(전통적 통계분석) : 대수의 법칙과 중심극한정리를 바탕으로 추론방법 정립
(Ergodic 시계열) 시계열 중에도 시간에 따라 의존성이 낮은 경우 표본의 수가 충분히 크다면 대수의 법칙과 중심극한정리가 성립 -주어진 하나의 표본경로를 사용하여 모든 통계적 추정, 검증, 예측작업이 이루어져야 한다는 현실적 제약 → 시계열이"Ergodic"하다는 가정이 필요

Ensemble average : stationary stochastic process X (t, ω)의 기대값 계산 E[X (t, ω)] = p $\lim_{I\to\infty}\left(\frac{1}{I}\right)$ $\sum_{i=1}^{I} X(t,\omega_i)$

Time average = $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X(t,\omega)$

stationary stochastic process 가 평균에 대해 Ergodic 하는 것 → , E[X (t, ω)] = p $\lim_{T\to\infty}\frac{1}{T}$ $\sum_{t=1}^{T} X(t,\ \omega)$

*시계열 : 주파수 분석

주파수(frequency) 정보

- 시계열에는 주기적 변동이 포함 : 이런 주기적 정보가 주파수 정보
- 프랑스 수학자 Fourier(1768년~ 1830년) : 주기적 시계열을 sine , cosine 삼각함수로 구성된 급수로 표현 함

푸리에 변환(Fourier transform, **FT**)
: 시간이나 공간에 대한 함수를 시간 또는 공간 주파수 성분으로 분해하는 변환을 말한다.
푸리에 변환은 이 변환으로 나타난 주파수 영역에서 함수를 표현한 결과물을 가리키는 용어로도 종종 사용

주파수(frequency) 시계열의 표현

- 시계열 y_t 가 시간이 지남에 따라 특정한 주기로 순환 → 시계열을 cosine 또는 sine 커브로 표한 함

$$y_t = A\cos(2\pi\omega t) \qquad t = 1,2,3,\ldots\ldots,n$$

A : 진폭 , ω : 주파수 (단위시간당 순환의 수)

- 주기가 p인 변동 → 시계열의 n기간 중 $\frac{n}{p}$ 번의 순환 존재
- 주파수(frequency) ω : 주기(p)의 역수 $\frac{1}{p}$: 월별 시계열에서 12개월 주기의 순환변동은 p = 12 , 주파수(ω)= $\frac{1}{12}$
- 저주파 (low frequency) 변동 : 주기가 긴 변동
- 고주파 (high frequency) 변동 : 주기가 짧은 변동

시계열 y_t

주기적 함수 $g(t)$, 평균= 0, 분산 σ^2 ⟶ 백색잡음 ε_t 로 구성

$$y_t = g(t) + \varepsilon_t$$
$$= \alpha_0 + \sum_{i=1}^{k}[\alpha_i \sin(2\pi\omega_i t) + \beta_i \cos(2\pi\omega_i t)] + \varepsilon_t$$

회귀계수는 최소제곱법으로 추정
→ Fast Fourier Transform 이용

주기도 (periodogram)

주파수 ω_i 를 x축에, 회귀 제곱합을 y축으로 하여 작성된 도표

시계열 분석 : data 전처리

1. 변수 변환이 필요하다. !!

- 시계열은 시간의 따라 **변동성이 커지면서 지수적으로 증가하는 경향이 있음**(비정상성)

→ **로그함수를 이용하여 시계열 선형화**

→ **변수 변환 : 역변환 ($\frac{1}{x}$) , 제곱근 변환 ($\sqrt{x}$) Box-Cox 변환**

Box-Cox Transformation

모형의 **정규성 가정이 성립한다고** 보기 어려울 경우 종속변수를 정규분포에 가깝게 변환시키는 기법

Box, G. E. P. & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society, 2 211–252.

2. 차분 (d : difference)

추세를 갖는 시계열 (비정상)을 추세가 없는 시계열 (정상성)로 전환

- 일반차분 : 직전 시점의 자료를 빼는 방법
- 계절차분 : 여러 시점 전의 자료를 빼는 방법
- **(목적) 차분은 시계열의 추세가 없는 시계열로 전환하는 목적 !!!**
- 단기변동을 증폭시킴

- **일반차분(1차 차분)** $\Delta y_t = y_t - y_{t-1}$
- **계절차분(1차 차분)**
  - **분기별 :** $\Delta_4 y_t = y_t - y_{t-4}$
  - **월별 :** $\Delta_{12} y_t = y_t - y_{t-12}$

시계열의 증감률 data : 시계열을 로그변환 후 차분한 것과 근사적으로 같으므로 증감률도 일종의 차분임

3. 평활화 (smoothing)

- 시계열에 **불규칙변동요인, 계절 요인** 등 주기가 짧은 변동요인으로 흐름(추세) 파악에 애로

(목적) 평활화를 통하여 **주기가 짧은 변동요인을 제거하여 시계열의 흐름(추세)을 파악하고자 하는 경우**

- 평활법은 모형식과의 관련성이 분명하지 않으나 시계열의 여러 구성성분들을 고려하기 위한 첫 단계로 활용
- 평활화와 유사한 용어 : 필터링(filtering)

(종류) 중심화 이동평균, 가중이동평균, 이중이동평균 등

시계열 QNSTJR : data 전처리

*** 평활화 smoothing 방법**

구분	내용	사례
중심화 이동평균	시계열의 단기 변동을 시차 구분 변경 없이 제거 → 불규칙 변동을 제거	3분기 중심화 이동평균 사례 $z_t = \frac{1}{3}(y_{t-1} + y_t + y_{t+1})$
이동평균법	**일정기간별 이동평균을 계산하고**, 이들의 추세를 파악하여 다음 기간을 예측하는 방법 →금융시장에서 활용 →최근 시점의 불확실성을 배제하여 동일가중평균한 것임.	5일 이평, 20일 이평 5일 이평 $z_t = \frac{1}{5}(y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t)$
가중이동평균	항마다 가중값을 달리하는 가중이동평균	* '핀테크' 증권분석 참조
지수 평활법 (exponential smoothing)	▪ 최근의 자료에 더 큰 가중치를 주고 현 시점에서 멀수록 작은 가중치를 주어 지수적으로 과거의 비중을 줄여 미래값을 예측하는 방법 ▪ 시점 t까지의 자료($z_1\ z_2\ \ldots z_t$) 에 과거자료일수록 지수적으로 감소하는 가중치 (w^t , w^{t-1} , …… w^1) (단, $0 \le w \le 1$) 를 주어 시점 t를 진행하면서 t시점에서 예측치(즉 평활값)를 구하는 과정	▪ 단순지수평활(Simple Exponential Smoothing) ▪ 이중지수평활(Double Exponential Smoothing) ▪ Winters 계절지수평활 (Winters Seasonal Exponential Smoothing)
이중이동평균	이동평균한 값을 다시 이동평균한 것	$z_3 = \frac{(\frac{y_1 + y_2 + y_3}{3} + \frac{y_2 + y_3 + y_4}{3} + \frac{y_3 + y_4 + y_5}{3})}{3}$

* 변동성 군집성

변동 군집성 (Volatility Clustering)

- 시계열 자료들에서는 **변동폭의 변화가 어떤 경향을 가지는 경우가 많다**.
- 한번 나타난 큰 변화는 당분간 계속 큰 변화를 유지하며, 작은 변화는 당분간 지속적으로 작은 변화를 나타낸다는 것

변동성 군집성이 갖고 오는 문제점

데이터 변환 (예를 들어 $\sqrt{\ }$, log 변환으로 상쇄되지 않으므로 자료의 변화 폭 (오차의 제곱 등)을 통해서 설명 가능하게 됨.

금융시장의 변동성은 시간에 걸쳐 변화(time varying) 하는 것이 일반적임.

→ 변동성이 커진다는 것은 일반적으로 시장으로 유입되는 정보의 양이 많아 짐

* 변동성 군집성

금융시계열 일반적 특성

- 금융시계열은 정규분포에 비해 **두터운 분포꼬리(fat tail)형태**
- 이상값 (outlier)이 발생할 확률이 정규분포에 비해 높음
- 첨도(kurtosis)가 정규분포에서의 3보다 큰 첨예분포(leptokurtic)임. 즉, 관찰값들이 독립적이지 못하고 일정한 의존성을 가지고 있음.
- **대부분의 금융시계열은 변동성 군집성 (volatility clustering) 현상이 나타남.**

→ 충격에 의해 분산이 한번 커지면 큰 상태로 어느 정도 지속되고 또한 상대적으로 분산이 작은 기간이 이를 따르는 현상

- ***전통적 통계분석 : 오차항의 등분산성을 가정***

→ 금융시계열에서는 등분산성을 만족하지 못하는 경우가 많다.

→ 이 경우 최소제곱법에 의한 계수 추정이 적절하지 못하다.

→ 변동성이 일정 한 것이 아니라 변동성이 증대하는 상황의 지속성 현상 , 즉 변동성 군집성이 나타난다.

Engle(1982) : 변동성 충격의 지속성을 고려한 ARCH 모형 제안

Engle, R.F. (1982) Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50, 987-1007.

확률과정

확률 과정(Stochastic Process)란 ?

→ **시간의 진행에 대**해 확률적인 변화를 가지는 구조를 의미

→ 통계분석에서 **표본들의 평균이 일정한 평균과 분산을 가지는 정규분포를 따른다는 가정 + 시간 진행을 추가 한 것.**

확률 과정(Stochastic Process) 종류

① **백색잡음과정 (white noise process)**

② **임의보행과정 (random walk process)**

③ **자기회귀과정 (Auto Regressive process)**

④ **이동평균과정 (Moving Average process)**

⑤ **자기회귀이동평균과정 (ARMA process)**

⑥ **자기회귀적분이동평균과정 (ARIMA process)**

Markov Process, Wieners process , Brownian motion process,
Ito process, poisson process

시계열의 종속성

- 종속성이 시계열의 과거, 현재, 미래를 연결하는 구조
- 시계열의 종속성은 시계열의 패턴 생성 → 이를 기반으로 미래 예측
- 전통적 통계분석에서 가정하는 확률변수의 독립성에 위배되는 특징

자기상관 (Auto Correlation)

시간에 따른 시계열의 의존구조 여부는 자기상관이나 자기공분산을 통해 파악 가능

- *상관계수*(correlation coefficient) : 두 변수 사이의 관계의 강도를 측정할 때 계산하는 양
- 자기상관 (Auto Correlation): 시계열의 *시차 값(lagged values)* 사이의 선형관계를 측정

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

자기상관(autocorrelation)이 없는 시계열 → 백색잡음(white noise)

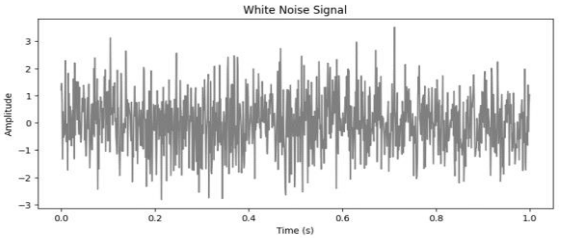
시계열 분석: 확률과정

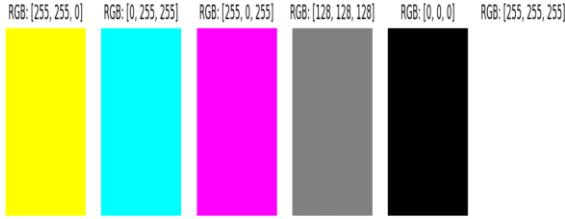
확률 과정(Stochastic Process) 종류

백색 과정 (White Noise Process)

- **백색잡음은 서로 다른** 주파수 에서 동일한 강도를 갖는 무작위 신호로 , 일정한 전력 스펙트럼 밀도를 제공
- 가시광선에서는 여러 색상이 모두 겹치게 되면 **백색**이 됨.

모든 주파수에서 동일한 전력 스펙트럼 밀도를 갖는 확률적 과정을 말함





파이썬 실습 파일 참조
: deeplearning : CNN입문 .. 참고)

∘ 통계적 특징

1) 평균= 0 → 평균적으로 소멸함 (zero mean)

2) 분산이 일정→ 변동성이 일정함 (finite variance) 백색과정에 속한 개별 랜덤변수들이 모두 동일한 확률분포를 갖음.. 대부분 평균이 0인 가우시안 확률분포를 갖는다고 가정함

3) 시간에 따른 독립적 : 자기상관(autocorrelation)이 없는 시계열 → 서로 독립이고 동일한 분포를 따르는 확률과정 –→ 임의 시간 구간 간에 상관성이 없음

4) 정상성 (Stationarity)
백색잡음은 약한 의미의 정상성을 만족한다. → 평균, 분산이 시간에 따라 변하지 않으며, 공분산도 시간 차이에만 의존한다.

- 평균 : $\mu_t = E(\varepsilon_t) = 0$
- 분산 : $\sigma_t^2 = Var(\varepsilon_t) = \sigma_\omega^2$
- 자기상관계수 : $\rho(i,j) = 0 \quad (i \neq j)$

백색잡음의 예

- 금융 데이터 : 주식의 미세한 가격 변동이 백색잡음으로 근사될 수 있다.
- 시계열 분석 : ARIMA 모델의 잔차(residual)는 백색잡음을 따라야 한다.
  → 만약 잔차가 백색잡음이 아니라면 모델이 데이터 패턴을 충분히 설명하지 못한 것이다.
- 신호 처리 : 라디오나 전자기기에 나타나는 전자적 간섭 신호.

시계열 분석: 확률과정

확률 과정(Stochastic Process) 종류

백색 과정 (White Noise Process)

White Noise ***과정의 성질 ➔- 정상성 과정 (stationary)***

이동평균(MA) 계열

$Y_t = \theta_1 \varepsilon_{t-1} + \epsilon_t$

- 평균 : $\mu_t = E(\varepsilon_t) = 0$
- 분산 : $\sigma_t^2 = Var(\varepsilon_t) = (\theta_1^2 + 1)\,\sigma_\omega^2$
- 자기상관계수 :
- $\rho(|i - j|) = \rho(i, j) = \frac{\theta_1}{1 + \theta_1^2}$ $i = j \pm 1$
- $\rho(i, j) = 0$, $i =$ 나머지

확률보행 계열

$Y_t = \Phi_0 + Y_{t-1} + \epsilon_t$

- 평균 : $\mu_t = \Phi_0 t$
- 분산 : $\sigma_t^2 = t\ \sigma_\omega^2$
- 자기상관계수 : $\rho(h) = \rho(t, t+h) = \frac{\sqrt{t}}{\sqrt{t+h}}$, $h > 0$

Stochastic Process

확률 과정(Stochastic Process) 종류

확률보행과정 (Random Walk Process)

Random / Stochastic
- 시간적으로 미리(사전에) 결과에 대해 정확히 예측,정의할 수 없다는 의미
- 단, 어떤 확률적 분포를 가질 수 있다는 통계적 규칙성은 있음 ☞ 랜덤성
-→ 여기서, `Random(무작위)` 및 `Stochastic(추계적)`를 같은 의미로 씀

Process (과정)
- **시간을 고려한 상태를** 말할 때는 주로 `과정`
- **시간을 고려하지 않는 상태는** 주로 `사건`이라고 함

Random walk

- 임의 방향으로 향하는 연속적인 걸음을 나타내는 수학적 개념
- 1905년 칼 피어슨이 소개
- 생태학, 수학, 컴퓨터 과학, 물리학, 화학 등의 분야에서 광범위하게 사용
- 대표적인 예 : 브라운 운동

- **절편 = 0 , 시계열의 평균은 일정 ,** 분산은 시간에 비례하여 증가 → 앞뒤로 움직일 확률이 동일하다고 해도 시간이 흐름에 따라 평균에서 점차 벗어나는 경향
- **절편 ≠ 0 , 시계열의 평균과** 분산이 모두 시간에 비례하여 증가

Brownian motion

- 1827년 스코틀랜드 식물학자 Robert Brown이 발견
- 액체나 기체 속에서 미소입자들이 불규칙하게 운동하는 현상

Stochastic Process

랜덤 과정 / 확률 과정 (Random Process, Stochastic Process, Probabilistic Process)

연속적 확률과정(Continuous-Time Stochastic Process)

연속적 확률과정은 시간이 **연속적인 값**을 가질 때 확률 변수들이 어떻게 변하는지를 설명하는 과정

즉, 확률 변수가 모든 시간 t∈R+ 정의된다.

- 정상성(stationary) : 연속형 확률과정 X(t)가 같은 시간의 주기를 가지면 같은 확률분포를 가진다는 의미
- **독립성 :** 연속형 확률과정 X(t)가 서로 겹치지 않는 구간들에 대해서 전부 독립적인 확률 변수라는 의미

특징
시간 변수의 연속성 : 시간 t는 실수(real numbers) 값. 예: t=0,0.1,0.2,...
상태 공간 : 상태 공간은 이산적(discrete)일 수도 있고 연속적(continuous)일 수도 있다.
연속 상태 공간: 브라운 운동. 이산 상태 공간: 특정 마르코프 과정.
예시 : 브라운 운동 (Brownian Motion): 주가와 같은 랜덤한 연속적 변화를 모델링.
포아송 점프 과정 (Poisson Jump Process): 일정한 시간 간격 동안 이벤트가 발생하는 빈도를 나타냄.
응용 분야
금융: 주식 가격 모델링. 물리학: 입자 운동 모델링. 공학: 신호 처리 및 통신 시스템.

이산적 확률과정(Discrete-Time Stochastic Process)

이산적 확률과정은 시간이 **이산적인 값**(간격이 일정하거나 불규칙적)에서만 정의될 때 확률 변수들이 어떻게 변하는지를 설명한다.
즉, 확률 변수는 특정 시점 t∈Z+서만 관찰된다.

이산적 확률과정 사례

- Markov Process
- Wieners process (Standard Brownian Motion)

Stochastic Process

마코프 확률과정(Markov Stochastic Process)

- 1906년 러시아의 수학자 안드레이 마코프가 도입
- T+1시점은 **과거의 값에는 전혀 영향을 받지 않고** 오직 오늘의 값(T시점)에만 영향을 받는 확률과정
- 현재에 대한 조건부로 **과거와 미래가 서로 독립인 확률 과정**

마코프 성질(Markov Property)

$$E(S_{t+1}|S_t, S_{t-1}, \cdots, S_0) = E(S_{t+1}|S_t)$$

마코프 과정의 응용

경제학 및 금융
- 주식 가격 모델링: 상태 전환 확률을 통해 시장 조건 예측.
- 고객 이탈 분석: 고객 행동의 변화 예측.

공학
- 통신 네트워크에서 데이터 패킷의 흐름 모델링.
- 로봇의 경로 계획(현재 위치에서 다음 위치로의 이동).

자연어 처리 (NLP)
- 마르코프 모델을 이용한 문장 생성 및 언어 모델링.

마코프 과정의 한계

- 단기 의존성: 과거 정보가 현재 상태를 통해서만 요약되므로 장기적인 의존성을 모델링하기 어려움.
- 현실 모델링의 제약: 실제 데이터는 종종 마코프 성질을 완전히 만족하지 않음

Stochastic Process

브라운 운동(Brownian motion)

- 1827년 스코틀랜드 식물학자 로버트 브라운(Robert Brown)이 발견
- 액체나 기체 속에서 미소입자들이 불규칙**(무작위)**하게 운동하는 현상

https://ko.wikipedia.org/wiki/%EB%B8%8C%EB%9D%BC%EC%9A%B4_%EC%9A%B4%EB%8F%99

Wieners process

- **Norbert Wiener(1894년 11월 26일 ~ 1964년 3월 18일)에 의해** 브라운운동을 수학적으로 정립
- 이러한 수학적 모델을 **Wieners process** 라고 부른다.

→ **Wieners process 를 통해 무작위성에 대한 수학적 추상화(abstraction) 를 달성**

Wieners process, → 표준 브라운 운동(Standard Brownian Motion, SBM)

- 마르코프 과정 중 변화량의 평균 0 →향후 예상되어지는 변화가 평균적으로 0 + 연간 분산은 1을 따르는 확률과정
- **Wieners process를 따르는 확률변수 z의 특징** $\Delta Z = \varepsilon \sqrt{\Delta t}$
- z의 변화량은 시간의 변화에 영향을 받고,z는 마르코프 과정을 따르기 때문에, 각각의 시간에 따른 z값은 다른 시간 간격의 영향을 받지 않는다.
- z의 변화량 : 평균은 0, 분산은 t의 변화량을 값으로 갖는 **정규분포를 따른다**. T시점의 분산은 T만큼의 값을 갖는다.
  → 시간의 변화량이 커질수록, z의 변동성도 함께 커지게 된다.
- (위너 과정 특징)
  → 위너과정의 **평균과 분산이 시간이 지남에 따라 커지는 선형함수로 표현된다는** 것

$$\Delta Z = \varepsilon \sqrt{\Delta t} \Rightarrow dX = \varepsilon \sqrt{dt}$$

- ε : 정규분포를 통해 뽑아낸 난수
- $\sqrt{dt}$: 시간 프레임에 따른 스케일링 팩터

➔ *위험자산의 통계적 성질을 모델에 담고 있지 못하는 한계*

1) *시간의 흐름에 따른* ***장기적 추세 (drift) – 평균***
2) *단기적으로 추세성보다 큰* ***변동성*** *– 표준편차*

➔ *수익률의 평균과 표준편차는 주가(확률변수)의 통계적 구조를 형성*

Stochastic Process

Wieners process, → 표준 브라운 운동(Standard Brownian Motion, SBM)

(특징)

1) 시작점 : 표준 브라운 운동은 보통 시간 t=0에서 0의 값에서 시작
2) 독립적 증분 : 브라운 운동의 증분은 서로 독립적 : 과거의 경로가 미래의 경로에 영향을 주지 않는다.
3) 정규분포 증분: 브라운 운동의 증분은 정규 분포를 따릅니다. 구체적으로, 어떤 시간 간격 Δt 에 대해, 그 간격 동안의 운동은 평균이 0이고 분산이 Δt 인 정규 분포를 따른다..

→ $B(t+\Delta t)-B(t)$는 평균이 0이고 분산이 Δt인 정규 분포를 따른다. 즉, 이 증분은 $N(0,\Delta t)$으로 모델링된다.

3) 연속성: 브라운 운동의 경로는 연속적 : 급격한 점프나 불연속적인 변화가 없다.

→ ***금융 연구에서 주가의 움직임이 위너과정을 따른다고 전제***

Stochastic Process

Arithmetic Brownian motion ABM

평균변화율이 0이 아닌 확률 과정 : Generalized Wiener process , Arithmetic Brownian motion

$$dS(t) = \mu\, dt + \sigma dW$$

- **S(t) : Stochastic Process**
- μ : S의 변화율의 **평균**, drift parameter –**장기적 추세**
- σ : 표준편차를 나타내는 상수항 - **단기적변동성**
- W: 위너 과정 – **예측 불가능성**

주가 Drift 현상
주가가 상승하든지 하락하든지 관계없이 매 기마다 무위험자산 수익률, 인플레이션 영향으로 Drift 만큼은 오르게 된다.

정의: 산술 브라운 운동은 표준 브라운 운동에 **선형 추세(drift)를 추가한 것**

이는 시간에 따라 일정한 속도로 증가하거나 감소하는 경향이 있는 확률 과정을 모델링한다.

특징: drift 항 때문에, 산술 브라운 운동 (ABM) 은 시간이 지남에 따라 평균값이 변화한다.

사용 : 경제 지표나 이자율과 같이 시간에 따라 선형적으로 증가하거나 감소하는 현상을 모델링하는 데 적합

Standard Brownian Motion(SBM) 과 Arithmetic Brownian motion(ABM)

구분	SBM	ABM
추세(DRIFT)의 유무(분포)	drift 항 없음 (정규분포)	drift 항 포함 (정규분포)
모델링 대상	순수한 무작위 움직임을 모델링	시간에 따라 일정한 추세를 가지는 현상을 모델링
평균값의 변화	평균값은 시간에 따라 변하지 않는다.	평균값은 시간에 따라 선형적으로 변한다.
사용	보다 일반적인 무작위 움직임을 이해하는 데 사용	경제학과 금융에서 이자율, 환율, 경제 지표 등의 시간에 따른 변화를 모델링하는 데 유용

Stochastic Process

Geometric Brownian motion (GBM)

- **정의**: 표준 브라운 운동을 기반으로 하지만, 로그 정규분포를 따르는 비율 변화를 모델링한다.

→ 이는 주로 금융에서 자산 가격의 시간에 따른 변화를 모델링하는 데 사용

- **특징**: 로그 정규 분포를 따르며, 이는 자산 가격이 음수가 되지 않고, 비율 변화가 중요한 경우에 적합

- $\frac{dS(t)}{S(t)} = \mu\, dt + \sigma dW$

- ➔ **Stochastic Process의 변화량을 변화율로 바꿔**, Stochastic Process의 값이 **음수(-)가 나오게** 하지 않는 것
- → 별도 참조 (뒷면)

구분	SBM	GBM
분포 전제	정규 분포	로그 정규 분포
값의 범위	양수 또는 음수	항상 양수 값 → 이는 자산 가격이 음수가 될 수 없다는 현실을 반영
수학적 복잡성		상대적으로 복잡성→ GBM이 자산 가격의 비율 변화를 모델링하기 때문
적용 분야	일반적인 확률 과정을 모델링하는 데 사용	주로 금융에서 자산 가격의 변동을 모델링하는 데 사용

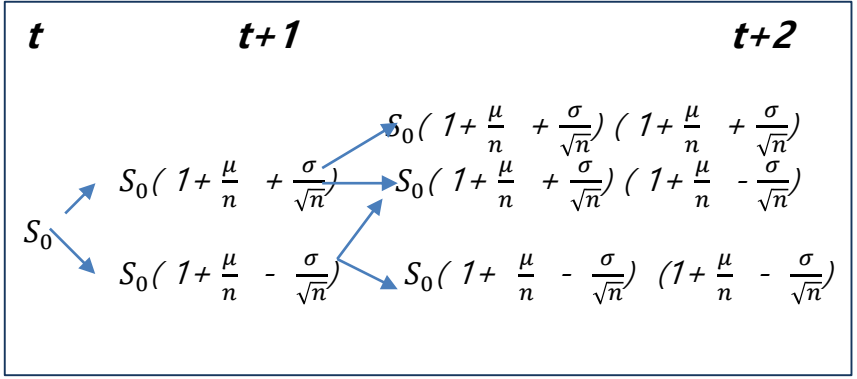
사례) 블랙-숄즈 모형 , Geometric Brownian motion

- 주식의 가격을 브라운 운동으로 모델링
- 이러한 모델은 가격이 불규칙적(무작위적)으로 움직인다는 가정을 기반

1. 주가의 변화는 연속적이다.2. 주가는 로그정규분포를 따른다.
3. 주가수익률은 정규분포를 따른다.4. 주식의 기대수익률과 수익률의 불확실성은 보유기간에 비례한다.

* Geometric Brownian Motion (GBM)

현재 주가(S_0) 변화 (Drift 효과 고려 ($\frac{\mu}{n}$) 매 기에 적용되는 주가 수익률 표준편차 $\frac{\sigma}{\sqrt{n}}$ 는 동일한 것으로 가정)



테일러 급수 (a=0)를 2차 항 까지만 적용

$$\ln(1+\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}}) \approx (\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}}) - \frac{1}{2}(\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}})^2$$
$$= \frac{\mu}{n} + \frac{\sigma}{\sqrt{n}} - \frac{\mu^2}{2n^2} - \frac{\mu\sigma}{n\sqrt{n}} - \frac{\sigma^2}{2n}$$

$$\ln(1+\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}}) \approx (\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}}) - \frac{1}{2}(\frac{\mu}{n}-\frac{\sigma}{\sqrt{n}})^2$$
$$= \frac{\mu}{n} - \frac{\sigma}{\sqrt{n}} - \frac{\mu^2}{2n^2} + \frac{\mu\sigma}{n\sqrt{n}} - \frac{\sigma^2}{2n}$$

주가의 모형 일반화

$$S_t = S_0(1+\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}})^u(1+\frac{\mu}{n}-\frac{\sigma}{\sqrt{n}})^d$$

-총시행횟수(nt)= u(상승 횟수) + d(하락 횟수)

-랜덤워크 시행결과 (M_{nt}) = u – d

-u = $\frac{1}{2}$(nt + M_{nt}) , d= $\frac{1}{2}$(nt - M_{nt})

$$\rightarrow S_t = S_0 (1+\frac{\mu}{n}+\frac{\sigma}{\sqrt{n}})^{\frac{1}{2}(nt+M_{nt})}(1+\frac{\mu}{n}-\frac{\sigma}{\sqrt{n}})^{\frac{1}{2}(nt-M_{nt})}$$

$$lnS_t = \ln S_0 + ut - \frac{\sigma^2}{2}t + \frac{M_{nt}}{\sqrt{n}}\sigma - \frac{\mu^2 t}{2n} - \frac{\mu t}{n}\frac{M_{nt}}{\sqrt{n}}$$

여기서 n→∞ *(무한 반복 수행)*

: $\frac{M_{nt}}{\sqrt{n}} = W_t$, $\frac{\mu^2 t}{2n} = 0$, $\frac{\mu t}{n} = 0$

$$lnS_t = \ln S_0 + ut - \frac{\sigma^2}{2}t + \sigma W_t$$

결국 $S_t = S_0\, e^{\sigma W_t + \left(u - \frac{\sigma^2}{2}\right)t}$

자연로그

$$lnS_t = \ln S_0 + \frac{1}{2}(nt + M_{nt}) \ln (1+\frac{\mu}{n}-\frac{\sigma}{\sqrt{n}}) + \frac{1}{2}(nt - M_{nt}) \ln (1+\frac{\mu}{n}-\frac{\sigma}{\sqrt{n}})$$

GBM : $S_t = S_0\, e^{\sigma W_t + \left(u - \frac{\sigma^2}{2}\right)t}$

Stochastic Process

Ito 과정 (Ito Process)

Ito 과정은 확률 미분 방정식(Stochastic Differential Equation, SDE)을 기반으로 연속 시간에서 확률적인 변화를 설명하는 **연속 확률 과정**의 일종이다.

- **Stochastic Process의 일반화**
- Brownian Motion을 기초로 하는 어떤 임의의 Stochastic Process의 Stochastic Differential Equation은 일반화 한 것

Ito 과정의 정의

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t$$

결정적 파트 확률론적 파트

- .X t : 시간 t에서의 Ito 과정.
- $\mu(Xt,t)$: 드리프트(drift) 항. Xt 가 시간에 따라 변화하는 평균적인 속도를 나타냄.
- $\sigma(Xt,t)$: 확산(diffusion) 항. Xt 의 랜덤 변동성을 나타냄.
- Wt : 브라운 운동(Brownian Motion) 또는 위너 과정(Wiener Process).
- dt: 무한소 시간 변화량.
- dWt : 브라운 운동의 무한소 변화량 ($dWt \sim N(0,dt)$)

Ito 과정의 구성 요소

1.드리프트 항 (μ(Xt,t)dt)➔ 확률 과정의 평균적인 방향성 예: 주식 가격의 장기적인 성장률.
2.확산 항 (σ(Xt,t)dWt) ➔ 과정의 무작위적 변동성 예: 주식 시장의 변동성.
3.브라운 운동 (Wt) ➔ 연속 시간에서 랜덤한 움직임을 나타내는 확률 과정. 주어진 시간 간격에서 정규분포를 따른다.

Ito 과정의 주요 성질

- 연속성 : Ito 과정은 시간에 대해 연속적이지만, 미분 가능하지는 않는다.
- 확률적 성질 : 과정의 미래 상태는 현재 상태와 브라운 운동의 랜덤 성분에 의존한다.
- Ito 보조정리 (Ito's Lemma) : Ito 과정의 함수 $f(Xt,t)$를 모델링할 때 사용된다.

$$df(X_t, t) = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial X_t}dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}(\sigma(X_t, t))^2 dt$$

이는 미분 계산에서 체인 룰을 확률 과정에 적용한 형태이다.

시계열 모형 구분

선형 시계열 모형

안정적 시계열 모형

AR, MA, ARMA모델의 경우 시계열이 정상성이라는 가정이 있는 상황에서 진행

- **안정적시계열(covariance-stationary)** : 3가지 조건 만족 (시계열의 평균이 일정, 시계열의 분산 일정, 시계열의 공분산 일정)
- 백색잡음계열 ε_t 의 시차변수의 선형결합과 시간 t 의 함수인 확정적 확률과정 d_t의 합으로 표현

Wold Decomposition : 모든 안정적인 확률과정은 확정적요소(deterministic component)와 확률적요소(stochastic component)의 합으로 나타낼 수 있음

$$Y_t = \mu + e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + \cdots \text{ 단, } e_t \sim \text{i.i.d}(0,\sigma^2)$$

현재의 Y_t는 평균(μ)과 현재 및 과거의 충격(shock)으로 나타낼 수 있는데 이를 Wold Representation이라고 함

$$y_t = \emptyset_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \ldots + \emptyset_p y_{t-p} + \varepsilon_t$$

- 자기회귀모형(AR (Auto-Regressive 모형)
  : 일정시점 전의 자료($y_{t-1}, y_{t-2} \ldots y_{t-p}$)가 현재 자료($y_t$)에 영향을 주는 모형 (p차) 현재의 데이터를 과거의 데이터의 선형 결합으로 설명한다. : 예시) AR(1) , AR(2)
  오늘의 값은 어제의 값 자체에 영향을 받는다.
  만약 계수가 양수이면 어제의 값이 큰 값이면 오늘의 값도 큰 값을 가지며, 음수이면 어제의 값이 큰 값이면 오늘의 값은 작은 값이 된다.

$$y_t = \theta_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

- 이동평균모형(MA (Moving Average) 모형)
  : 유한한 개수의 백색잡음(오차항)의 결합 (q차)
  → 현재의 자료 (y_t) 를 과거의 자료의 오차항(ε_{t-1}..) 의 선형 결합으로 설명한다. 예시) MA(1) , MA(2)
  → 오늘의 값은 어제의 값의 오차 항에 영향을 받는다.
  → 계수가 양수이면, 어제의 값이 상승하는 추세이면, 오늘의 값도 상승하는 추세
- ARMA 모형

$$y_t = \emptyset_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots \emptyset_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

시계열 모형 구분

불안정시계열 모형

확률보행 (Random walk) 모형

시계열 data (X_t) $X_t = X_{t-1} + Z_t$

- **임의 방향으로** 향하는 연속적인 걸음을 나타내는 수학적 개념
- **금일 주가 = 전일 주가 + Noise** → X_t 가 X_{t-1}보다 높거나, 낮아질 확률이 반반이 되는 의미!!
- **Karl Pearson (1905)** 소개
- **절편 = 0 , 시계열 평균은 일정 , 분산은 시간에 비례하여 증가** → 앞뒤로 움직일 확률이 동일하다고 해도 시간이 흐름에 따라 평균에서 점차 벗어나는 경향
- **절편 ≠ 0 , 시계열 평균과** 분산이 모두 시간에 비례하여 증가
- 대표적인 예 : 브라운 운동

Random Walk with Drift 모형

$X_t = X_{t-1} + \mu + Z_t$

- $\mu = \mathrm{E}(X_t - X_{t-1})$

ARIMA (Auto Regressive Integrated Moving Average)

: **자기회귀모형(AR)**과 **이동평균모형(MA)**과 데이터의 정상성을 확보하기 위한 **차분 (differencing)을 통해** 데이터를 정상성으로 변형한 모델

→ **과거의 관측값과 오차를 사용해서** 현재의 시계열 값을 설명하는 **BOX-JENKINS 모델(ARMA)를 일반화한 것**

→ **1차 차분, 2차 차분 수행 결과를 시각화하고, 시계열 곡선이 특정한 트렌드(constant average trend)를 보이면 1차 차분을, 시간 변화에 따라 분산이 변하는 경우 2차 차분을 수행**

→ ARMA모델에 차분이라는 차수 d가 포함되어 ARIMA(p,d,q)로 표현

예시) ARIMA (1, 1, 1) : AR(1), MA(1) , 1차 차분

ARIMA모델에서는 차분 (difference)을 통해 non-stationary한 자료에서도 좀 더 나은 예측을 하는 것이 목표

시계열 모형 : 변동성 모형

비선형 시계열 모형 analysis of volatility

ARCH (Autoregressive conditional heteroskedasticity) 모형

ARCH 모형은 시계열 데이터의 분산이 시간에 따라 변화하는 특성(이분산성, Heteroskedasticity)을 모델링하기 위해 개발된 모형(

(ARCH 모형의 기본 아이디어)

- 과거 오차 항의 크기(제곱된 값)에 따라 현재 시점의 분산이 달라진다는 가정을 기반으로 함.
- 분산이 일정한 시점이 아니라, 조건부 분산(conditional variance)을 고려함.

이분산 (heteroskedasticity) : 독립변수 값이 변화할 때 이에 대응하여 변화하는 종속변수값의 **분산이 상이하거나 어떤 형태(pattern)를 가지는 것** → 추정량은 불편성은 유지할 수 있지만 최소분산을 갖는 효율성은 지니지 못한다는 것임.

Volatility 모형

Historical volatility models	**과거변동성의 단순평균을 통해 변동성의 예측값을 추정하는 모형**
Exponentially weighted moving average models)	단순평균이 아닌 지수적 가중치를 한 이동평균을 통해 예측값을 추정하는 모형
Autoregressive volatility models	변동성에 대한 ARMA모형을 추정하여 예측값을 만들어 내는 모형
implied volatility models	거래되는 옵션가격에 내재(implied) 된 해당옵션의 존재기간 동안의 변동성에 대한 예측값을 만들어내는 모형

시계열 모형 : 변동성 모형

비선형 시계열 모형 **변동성 분석(analysis of volatility)**

ARCH (AutoRegressive Conditional Heteroskedasticity) 모형 , 자기회귀 조건부 이분산 모형

- ARCH 모형은 수익률의 t기 분산이 회귀모형 **오차항 제곱의 전기 변수들로 설명될 것이란 아이디어를 기반**
- 종속변수의 분산은 그 자신의 과거값과 독립변수들의 함수로 모형화 됨.

➔ **t기의 조건부 분산이 t-1기의 잔차제곱에 의존하는 모형임.**

- Robert F. Engle(1982) 제시

ARCH 모형 변동성의 특징

1. 시간에 따라 변한다. → 실제로 많은 시계열 data는 시간에 따라 분산이 변한다.
2. 군집성을 반영한다. → 현재의 변동성은 과거의 오차항 제곱이 클수록 클 것이기 때문이다.
3. 계수가 특정 조건을 만족하는 경우 초과첨도 현상이 설명된다.
4. 시차 p를 길게 설정하면 지속성이 설명된다.

GARCH (Generalized ARCH) 모형

- Tim P. Bollerslev (1986) : GARCH(Generalized Autoregressive Conditional Heteroskedasticity) 모형으로 일반화됨.
- ARCH 모형에선 **변동성의 지속성을 설명하기 위해 긴 시차의 p가 요구** → ARCH의 특징을 살리기 위해 **일반화된 자기회귀조건부 이분산 (GARCH)** 모형이 고안 → **과거 q개의 오차제곱항 대신 무한개의 오차제곱항을 고려**
- 현재 분산이 과거 오차 항뿐만 아니라 과거의 분산에도 의존한다고 가정한다.
- ARCH 모형은 차수 q가 커질수록 많은 파라미터를 추정해야 하기 때문에 비효율적일 수 있으며, 이를 해결하기 위해 GARCH가 제안되었다.
- ARCH 모형과 전반적인 구조는 동일하지만 분산 방정식에서 전기 오차항의 제곱들로 이뤄진 ARCH 항에 전기 분산들로 이뤄진 GARCH 항이 추가된 형태의 모형

시계열 모형 : 변동성 모형

비선형 시계열 모형 **변동성 분석(analysis of volatility)**

ARCH (Autoregressive conditional heteroskedasticity) 모형

- ARCH(1) 모형 : t기의 조건부분산이 **t-1기의 잔차제곱에** 의존하는 모형

- 평균방정식: $y_t = \mu_t + u_t$, $u_t = \sqrt{h_t}\ \varepsilon_t$
- 분산방정식: $h_t = \alpha_0 + \alpha_1 u_{t-1}^2$

μ_t : 조건부평균(conditional mean)
u_t : 잔차항
ε_t : 평균=0, 분산=1 인 백색잡음(white noise)
h_t : u_tt의 조건부 분산(conditional variance)

평균방정식(mean equation)에서

조건부평균 $\mu_t = y_t$ 의 기대값(expected value)

•일반적인 **회귀모형이나 시계열모형** : 잔차항 u_t 는 모든 t시점에서 평균=0, 일정한 분산을 갖는 백색잡음(white noise)으로 가정함.

•**조건부변동성모**형 : μ_t 가 시간 가변적인(time varying) , 분산 h_t의 제곱근과 평균=0, , 분산=1인 백색잡음 εt의 곱이기 때문에 μ_t는 0의 평균과 시간 가변적인 분산 h_t 를 가짐.

분산방정식(variance equation)은 시간 가변적인 분산 h_t 의 정의식임.

• ARCH(1) 모형에서 h_t는 잔차제곱의 시차값에 의해 결정됨.

•분산 h_t 은 음(-)의 값이 될 수 없기 때문에 모수 α_0 와 α1 역시 음(-)이 될 수 없다., α1은1보다 작아야 함.

•분산방정식은 t-1기에 커다란 충격➔ **즉 u_{t-1}^2 이 크면 t기에도 커다란 변동성이 예상된다는 것을 시사함.**

시계열 모형 : 변동성 모형

비선형 시계열 모형

GARCH (Generalized ARCH) 모형

과거 q개의 오차제곱항 대신 무한개의 오차제곱항을 고려

ARCH 모형의 한가지 단점은 시계열의 조건부변동성을 기술하기 위해 너무 많은 수의 모수를 필요로 한다는 것임.
•이런 문제를 해결하기 위하여 Tim P. Bollerslev (1986)는 GARCH(Generalized Autoregressive Conditional Heteroskedasticity)라는 일반화된 모형을 제시함.
•가장 단순하고 널리 이용되는 GARCH 모형 → GARCH (1,1) 모형

Kospi 일별수익률은 백색잡음계열 처럼 보이지만 ,
수익률의 제곱인 변동성에는 군집현상과 지속성이 나타남. → 시계열의 분산(변동성)은 시간이 변함에 따라 특정시점에는 변동성이 커지거나 작아 짐.

평균방정식: $y_t = \mu_t + u_t \quad , u_t = \sqrt{h_t}\ \varepsilon_t$
분산방정식: $h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \boldsymbol{\beta_1 h_{t-1}}$

GARCH (1,1) 모형에도 제약조건

- 분산 h_t : 음(-)의 값이 될 수 없기 때문에 → 모수 α_0 , α_1 , β_1모두 음(-)의 값이 될 수 없음
- 잔차항 u_t가 일정한 무조건부 분산을 갖기 위해서는 아래를 만족해야 한다.
- $\beta_1 \leq 1$,
- $\alpha_1 + \beta_1 < 1$ 의 안정성(정상성) 조건을 만족해야 함.

ARCH와 GARCH의 차이점

특징	ARCH	GARCH
조건부 분산 정의	과거 오차 항에 의해서만 결정	과거 오차 항 + 과거 분산에 의해 결정
모수 개수	차수 q 증가 시 모수 개수 급증	상대적으로 적은 모수로 동일 효과 구현
적용성	짧은 메모리 의존성이 있는 데이터에 적합	장기 메모리 의존성이 있는 데이터에 적합

시계열 모형 : 변동성 모형

비선형 시계열 모형

ARCH 및 GARCH 모형의 주요 특징

1. 이분산(Heteroskedasticity) 모델링 : 분산의 시간 가변성을 설명할 수 있다.
2. 변동성 클러스터링 : 변동성이 큰 시점이 연속적으로 발생하는 금융 데이터 특성을 잘 반영한다.
3. 응용 분야 : 금융 시장 데이터(주가, 환율, 채권 수익률 등)., 경제 데이터(물가 변동, 실업률).

모형 선택 및 확장

모형 선택

- 데이터에 분산의 시간 변화가 없다면 ARCH/GARCH 모형이 필요하지 않음.
- ACF(자기상관 함수)와 PACF(부분 자기상관 함수)를 분석해 잔차의 이분산성을 확인.

확장 모형

- **EGARCH (Exponential GARCH):** 분산이 로그 스케일로 표현되어 비대칭성을 반영.
- GJR-GARCH: 좋은 뉴스와 나쁜 뉴스의 영향을 다르게 반영.
- Multivariate GARCH: 다변량 데이터에 확장.

시계열 모형 : 변동성 모형

비선형 시계열 모형

EGARCH(Exponential Generalized Autoregressive Conditional Heteroskedasticity) 모형

- Nelson(1991)이 제안한 모형으로,조건부 분산의 비대칭성(asymmetry)과 **로그 척도**를 활용하여 GARCH 모형의 한계를 보완한 확장된 형태
- EGARCH는 특히 금융 시계열에서 "나쁜 뉴스"와 "좋은 뉴스"가 변동성에 미치는 영향을 다르게 모델링할 수 있다.

EGARCH 모형의 특징

조건부 분산의 로그 척도

- EGARCH는 조건부 분산을 로그 스케일로 표현하여 분산이 항상 양수가 되도록 보장한다. → 따라서 GARCH와 달리 αi , βi 등의 계수에 양수 제약 조건이 필요 없다.

비대칭 효과 모델링 (Leverage Effect) :

- "나쁜 뉴스"(음의 충격)가 "좋은 뉴스"(양의 충격)보다 변동성에 더 큰 영향을 미칠 수 있는 비대칭성을 반영한다.

메모리 효과(Long Memory)

- EGARCH는 과거 충격이 장기적으로 변동성에 영향을 미칠 수 있는 현상을 효과적으로 모델링한다.

EGARCH 모형의 수학적 표현

$$\log(\sigma_t^2) = \omega + \sum_{i=1}^{q} \beta_i \log(\sigma_{t-i}^2) + \sum_{j=1}^{p} \alpha_j \frac{\epsilon_{t-j}}{\sigma_{t-j}} + \sum_{j=1}^{p} \gamma_j \left(\left| \frac{\epsilon_{t-j}}{\sigma_{t-j}} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$\log(\sigma_t^2)$): 현재 시점의 조건부 분산(변동성)의 로그 값.

ω: 상수항.

βi : 이전 로그 분산(조건부 분산)의 효과 (GARCH 항).

αj : 표준화된 충격 ($\epsilon t-j/\sigma t-$j) 의 효과 (ARCH 항).

γj: 비대칭성을 반영하는 계수 (충격의 방향성).

시계열 모형 : 변동성 모형

비선형 시계열 모형

EGARCH와 GARCH의 차이점

특징	GARCH	EGARCH
조건부 분산의 표현	선형 형태	로그 형태 (비선형)
분산의 양수 조건	$\alpha_i, \beta_i \geq 0$ 제약 필요	제약 조건 없음
비대칭성 모델링	비대칭성 반영 불가	반영 가능 (Leverage Effect)
충격의 영향	대칭적	음/양의 충격을 다르게 반영
모형의 유연성	상대적으로 덜 유연	비대칭적 데이터 모델링에 더 적합

EGARCH 모형의 응용 분야

- 금융 시장 : 주식, 환율, 채권 등 금융 자산의 변동성 분석.
- "나쁜 뉴스"(예: 금융 위기)와 "좋은 뉴스"의 변동성에 대한 상이한 효과 분석.
- 리스크 관리 : VaR(Value at Risk) 계산에서 변동성 추정.
- 경제 데이터 : 거시경제 지표(예: 물가, 실업률)의 변동성 분석.

EGARCH 모형의 장점

- 조건부 분산이 항상 양수가 되도록 보장 (로그 스케일 사용).
- 비대칭 효과를 자연스럽게 모델링.
- 금융 데이터에서 자주 관찰되는 변동성 클러스터링과 비대칭성을 효과적으로 설명.

EGARCH 모형의 한계

- 복잡한 구조로 인해 추정이 더 어려울 수 있음.
- 과거 충격의 로그 척도를 사용하는 방식이 일부 데이터에서 비적합할 수 있음.

예측 분석

- **시계열 모형**을 이용한 예측
- **회귀 모형**을 이용한 예측
- **VAR 모형**을 이용한 예측
- **공적분분석 (VECM)**을 이용한 예측

*금융시계열 분석 이슈

시계열 데이터 잡음 제거(Denoising) 방법

금융시계열자료는 대부분 Non-Stationarity 하므로, 전처리 없이 머신러닝 알고리즘에 학습할 경우

→ **단순 후행 예측, 성능 저하, 잘못된 추론 등의 문제점 야기**

1. 이동 평균 (Moving Average)

- 방법: 데이터의 연속된 구간의 평균을 계산하여 잡음을 제거.
- 장점: 계산이 간단하며 단기적인 변동을 평활화.
- 단점: 급격한 변화(트렌드)가 있는 데이터에서는 정보 손실 가능.

2. 가우시안 필터 (Gaussian Filter)

- 방법: 데이터에 가우시안(정규 분포) 커널을 적용하여 평활화.
- 장점: 급격한 변화에 민감하지 않아 더 부드러운 결과 제공.
- 단점: 매개변수(표준 편차 σ) 설정이 필요.

3. 웨이블렛 변환 (Wavelet Transform)

- **방법**: 데이터의 신호와 잡음을 주파수 영역에서 분리하여 저주파 신호를 복원.
- **장점**: 비정상 시계열 데이터에서 잡음을 제거하기에 적합.
- **단점**: 파라미터(웨이블릿 종류, 임계값) 선택이 복잡.

- A new wavelet-based denoising algorithm for high-frequency financial data mining : Edward W. Sun , Thomas Meinl 2012
- Methods of Denoising Financial Data : Thomas Meinl and Edward W. Sun 2015
- Financial Time Series Forecasting Using Improved Wavelet Neural Network Master's Thesis 2014

*금융시계열 분석 이슈

시계열 데이터 잡음 제거(Denoising) 방법

4. 사비츠키-골레이 필터 (Savitzky-Golay Filter)

- 방법: 데이터의 로컬 구간에 다항식을 적합하여 잡음을 제거.
- 장점: 데이터의 패턴(예: 피크)을 유지하면서 잡음 제거.
- 단점: 윈도우 크기와 다항식 차수 선택이 중요.

5. 딥러닝 기반 Denoising (Autoencoder)

- 방법: 딥러닝 기반 오토인코더를 사용하여 노이즈 제거.
- 장점: 복잡한 패턴과 비선형 신호 잡음 제거에 적합.
- 단점: 모델 훈련에 많은 데이터와 시간이 필요.

6. 저주파 필터 (Low-Pass Filter)

- 방법: 신호의 주파수가 특정 임계값 이하인 부분만 통과시킴.
- 장점: 고주파 잡음을 제거하기에 적합.
- 단점: 급격한 변화나 고주파 성분의 신호 손실 가능.

7. 평균 필터와 중앙값 필터

- **평균 필터**: 단순히 일정 구간의 평균을 계산하여 평활화
- **중앙값 필터**: 일정 구간의 중앙값을 계산하여 이상치와 잡음을 억제

https://github.com/freejyb/KDT
시계열분석/ **Denoising** 방법

https://github.com/freejyb/Deep-Learning
Autoencoder

*금융시계열 분석 이슈

시계열 데이터 잡음 제거(Denoising) 방법

방법 비교 요약

방법	특징	장점	단점
이동 평균	단순 평균	간단하고 빠름	급격한 변화에서는 정보 손실
가우시안 필터	정규 분포 기반	부드러운 결과	파라미터 설정 필요
웨이블렛 변환	주파수 기반 분리	비정상 데이터에 적합	복잡한 설정 필요
사비츠키-골레이 필터	다항식 적합 기반	패턴 유지	윈도우와 차수 설정 중요
저주파 필터	주파수 임계값 기반	고주파 잡음 제거	신호 손실 가능
딥러닝 (오토인코더)	학습 기반	비선형 신호와 잡음 분리 가능	데이터와 훈련 시간 필요

- 단순한 잡음 제거: 이동 평균, 가우시안 필터
- 데이터 패턴 유지: 사비츠키-골레이 필터, 웨이블렛 변환.
- 복잡한 잡음 제거: 저주파 필터, 딥러닝 기반 방법.

시계열 예측 분석: 사전 검토

- **시계열의 분포는 ?** 정상성(Stationarity) 여부
- **자기상관계수(Autocorrelation)** : 시계열의 시차값 (Lagged Values) 사이의 선형 관계를 측정하는 개념
  → 륭(Ljung) & 박스(Box)의 검정(portmanteau 검정)을 통해 시계열의 일정 시차까지의 자기상관관계 존재 여부를 파악 가능
  → ACF(Autocorrelation Function)을 통해 시각화
- **자기상관함수 (ACF) : 확률과정의 평균, 분산과 함께 정상확률과정의 확률구조를 특정지어주므로, 시계열을 판단하는데 중요한 역할을 한다.**
- **부분자기상관계수(Partial Autocorrelation)**
  : 제3의 변수를 제거하고 시차에 따른 상관을 확인해보아야 하는데 (특정시차 만), 이를 부분상관계수를 의미

자기상관함수 (Autocorrelation Function, ACF)

- 자기상관함수(ACF)는 **시계열 데이터에서 시점 간의 상관관계**를 측정하는 함수로, 특정 시차(lag)를 기준으로 시계열 데이터의 값을 얼마나 잘 설명할 수 있는지를 나타낸다.
- 이는 시계열 분석에서 데이터의 패턴을 이해하고 모델링에 활용하기 위한 중요한 도구이다.

시차 k의 자기상관계수

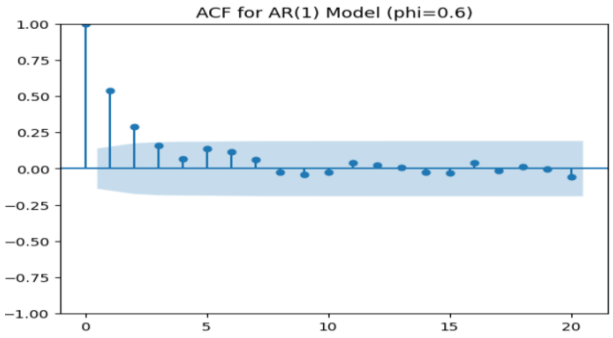
$$\rho_k = \frac{\mathrm{Cov}(X_t, X_{t-k})}{\sqrt{\mathrm{Var}(X_t) \cdot \mathrm{Var}(X_{t-k})}}$$

자기상관계수의 값은 [−1,1]사이에 있음.

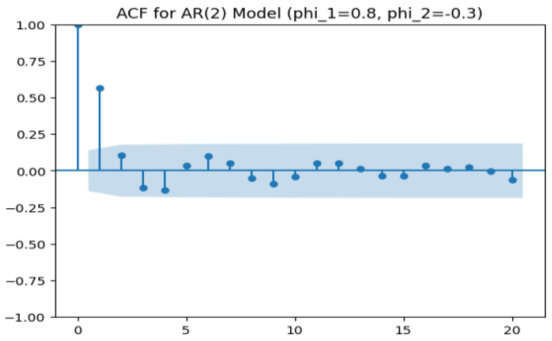
ACF의 시각화

- ACF는 시차에 따른 자기상관계수를 플롯하여 시각적으로 표현한다.
- 이 플롯은 ACF 플롯이라고 하며, 시계열 데이터의 패턴(주기성, 추세, 메모리 효과 등)을 분석하는 데 사용된다.

AR(1)



AR(2)



시계열 예측 분석: 사전 검토

자기상관함수 (Autocorrelation Function, ACF) 활용

- **시계열 데이터의 구조 이해:**

데이터가 특정 시차에서 강한 상관관계를 보이는지 확인.
예: 주기성(periodicity)이나 트렌드 탐지.

- **모형 식별:**

ARIMA 모델에서 AR(자기회귀) 및 MA(이동평균) 차수를 결정하는 데 사용.

- AR 모델: ACF가 지수적으로 감소하거나 특정 시점 이후 끊김.
- MA 모델: ACF가 특정 시차 이후 갑작스럽게 끊김.

- **잔차(residual) 검증:**

시계열 모델의 잔차가 백색잡음(white noise)인지 확인.
백색잡음일 경우 ACF는 시차 $k>0$에서 거의 0에 가까움.

S&P 사례

ACF와 PACF의 차이

- ACF: 전체 시계열 데이터를 고려하여 직접적으로 시차 k의 자기상관을 측정.
- PACF (부분 자기상관 함수): 중간에 있는 모든 시차의 효과를 제거하고, 순수하게 시차 k와의 상관관계를 측정.

ACF 플롯 (Autocorrelation Function of Residuals):

백색잡음 검증: 대부분의 시차(lags)에서 자기상관계수(ACF)가 신뢰구간(노란 영역) 안에 위치하고 있음.
이는 잔차(residuals)가 독립적이고, 백색잡음(white noise)의 특성을 가진다는 것을 의미.
만약 특정 시차에서 ACF가 신뢰구간을 벗어난다면, 해당 모델이 데이터의 구조를 완전히 설명하지 못했음을 나타냄.

잔차 플롯

- 잔차가 평균 0을 중심으로 랜덤하게 분포하고 있음(빨간 점선이 평균 0).
- 잔차의 분포가 특정 패턴 없이 무작위적으로 보이는 경우, 모델이 데이터를 잘 적합했다고 볼 수 있음.

시계열 예측 분석: 사전 검토

자기상관함수 (Autocorrelation Function, ACF) 추가적 검증

•**정규성 테스트**: 잔차가 정규분포를 따르는지 확인 (Shapiro-Wilk, Kolmogorov-Smirnov 테스트 등 사용 가능).

•**Ljung-Box 테스트**: 잔차가 백색잡음인지 여부를 통계적으로 검증하는 방법

Ljung-Box 테스트

- 시계열 데이터의 잔차(residuals)가 백색잡음(white noise)인지 검증하는 통계적 방법.
- 백색잡음은 시계열 모델이 데이터를 충분히 설명했음을 의미하며, 잔차가 독립적이고 자기상관이 없는 상태를 가정한다.

Ljung-Box 테스트의 목적

- 잔차가 특정 시차 범위 내에서 자기상관을 가지는지 확인한다.
- 잔차가 백색잡음이라면 시계열 모델이 데이터를 잘 적합했다고 볼 수 있다.

Ljung-Box 테스트의 가설

- 귀무가설 ($H0$): 잔차는 백색잡음이다 (자기상관 없음).
- 대립가설 ($H1$): 잔차는 백색잡음이 아니다 (자기상관 있음)

Ljung-Box 통계량

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\rho_k^2}{n-k}$$

n: 데이터의 샘플 크기.
m: 테스트할 최대 시차(lags).
ρk : 시차 k에서의 잔차의 자기상관계수.

.p-value 계산 Q 통계량은 자유도 m의 카이제곱 분포를 따른다.

- **p-value**:< 0.05 : 귀무가설을 기각 (잔차가 백색잡음이 아님).
- **p-value**:> 0.05 : 귀무가설을 기각하지 못함 (잔차가 백색잡음임).

결과 해석

- **Ljung-Box 테스트결과**: p-값 <0.05 : 잔차가 백색잡음(white noise)이 아닐 가능성을 시사.

→ 이는 모델이 데이터의 전체 패턴을 완전히 설명하지 못했을 수 있음을 의미 .

- **Shapiro-Wilk 테스트결과**: p>0.05 잔차가 정규분포를 따른다는 귀무가설을 기각 할수 없음.

→ 이는 잔차가 정규성을 만족한다고 해석할 수있음.

시계열 예측 분석: 사전 검토

부분 자기상관 함수 (Partial Autocorrelation Function, PACF)

- **PACF**는 시계열 데이터의 특정 시차(lag)에서 **직접적인 상관관계**를 측정하는 함수이다.
- 이는 자기상관 함수(ACF)와 달리, 중간에 있는 다른 시차의 영향을 제거한 상태에서 시차 k에서의 상관관계를 측정하는 것이다..

- **직접적 상관관계:** PACF는 특정 시차에서의 순수한 상관관계를 계산하기 위해, 모든 중간 시차의 간접적인 영향을 제거한다.
- **ACF와의 차이**: ACF는 특정 시차에서의 모든 상관관계를 포함(직접적 + 간접적).

PACF는 간접적 상관관계를 제거하고 순수한 상관관계만 계산하는 것.!!

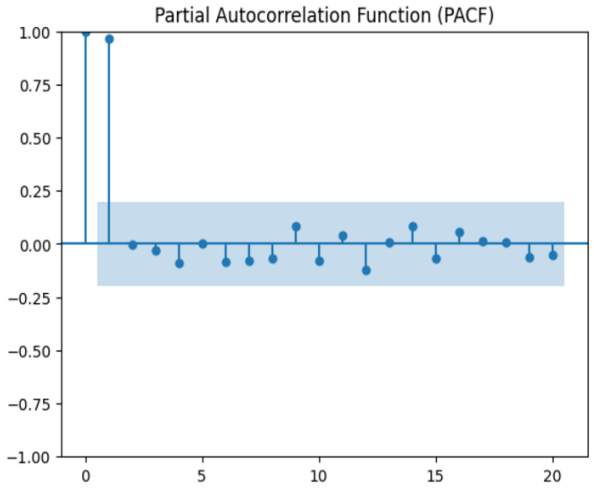
- **PACF의 값 범위**: PACF 값은 [−1,1] 범위에 속한다.

→ 1: 완전한 양의 상관관계. −1: 완전한 음의 상관관계.0: 상관관계 없음.

PACF와 ACF의 차이

특징	ACF (Autocorrelation Function)	PACF (Partial Autocorrelation Function)
계산 방식	특정 시차에서 전체 상관관계 측정	특정 시차에서 직접적인 상관관계 측정
포함 관계	직접적 + 간접적 상관관계 모두 포함	간접적 상관관계를 제거한 순수한 상관관계
모델링 활용	MA(이동 평균) 차수 q 결정	AR(자기회귀) 차수 p 결정
그래프의 패턴	지수적 감소 또는 특정 시차에서 끊어질 수 있음	특정 시차 이후 갑작스럽게 0에 가까워짐

PACF와 AR 모형
PACF는 AR(자기회귀) 모형의 차수 식별에 중요한 도구이다:
•AR(1): PACF가 시차 1에서만 유의미하고 이후 빠르게 감소.
•AR(2): PACF가 시차 1과 2에서 유의미하고 이후 감소.
•혼합 모델(ARMA): PACF와 ACF 모두 활용.



AR(1)

https://github.com/freejyb/KDT
시계열분석/ACF

시계열 모형을 이용한 예측

시계열의 추세변동 : 확정적 추세변동 + 확률적 추세변동

데이터의 움직임을 이해하고 예측하는 데 있어 중요한 역할

구분	확정적 추세 (Deterministic Trend)	확률적 추세 (Stochastic Trend)
정의	▪ 시간에 따라 일정한 방향성을 가지고 변화하는 추세 ▪ 고정된 함수나 모델에 의해 설명될 수 있다. ▪ 시간의 변화에 따라 예측 가능한 패턴을 보인다.	▪ 시간에 따라 **무작위적(trend)**이고 예측 불가능한 방식으로 변화하는 추세 ▪ 랜덤 워크(random walk)나 다른 확률적 과정에 의해 설명될 수 있다. ▪ 시간의 변화에 따라 예측하기 어려운 패턴을 보인다.
가정	▪ 추세의 기울기가 시간에 따라서 변하지 않을 것	추세의 기울기가 시간에 따라 변할 수 있고 , 추정된 증가량에는 과거 기간 동안 평균 증가만 갖는다.
특징	▪ 시간의 흐름에 따라 일정한 패턴이나 경향성을 보임. ▪ 선형적(linear) 또는 비선형적(non-linear)일 수 있음. ▪ 선형 회귀 모델을 사용하여 데이터의 확정적 추세를 모델링할 수 있음	▪ 시간의 흐름에 따라 불규칙적이고 예측 불가능한 변동을 보임. ▪ 시계열 데이터에서 노이즈(noise)나 충격(shock)으로 인한 변동을 포함할 수 있음. ▪ 주식 가격이 확률적 추세를 따를 수 있음.
적용 사례	▪ 장기적으로는 일정한 추세를 보이는 경향이 있을 때 적용 ▪ 경제 성장률, 인구 증가율 등	▪ 가격의 움직임이 무작위적이고 예측 불가능한 패턴을 보이는 경우 적용 ▪ 금융 시장의 주가 변동, 기후 변화 데이터 등

- **확정적 추세변동** : $Y_t = \alpha + \beta t + \varepsilon_t$ **(ARMA과정)**
- **확률적 추세변동** : $Y_t = Y_{t-1} + \varepsilon_t$ (ARIMA 과정)

$\boldsymbol{Y_t = \alpha + \beta t + Y_{t-1} + \varepsilon_t}$

시계열 모형을 이용한 예측

AR(1) 모형 | **AR(1) 모형: 단위근 검정**

$$X_t = \emptyset_0 + \emptyset_1 X_{t-1} + \varepsilon_t$$

- 현 시점의 자료를 자기의 p-시차 전의 과거의 값으로 나타낼 수 있는 모형
- (가정)
- **백색잡음(white noise) 과정** : 오차항 ε_t의 평균이 0이고, 분산이 σ^2 일정한 iid 가정
- 회귀모형"이므로, 설명변수(과거변수)

- $\emptyset_1 = 0$ 이면, X_t는 **White noise**
- $\emptyset_1 = 1$이면, $\emptyset_0 = 0$, X_t는 확률보행 (Random walk) 모형

→ 모형은 정상성을 띠지 않는다(시간 변경에 따라 분산이 증가하여 분포가 일정하지 않는다.)

→ 이 모형은 확률론적 추세 (Stochastic trend)가 존재하는 가장 단순한 모형

- **$\emptyset_1 = 1$이면, $\emptyset_0 \neq 0$, X_t는 Random Walk with Drift 모형**

→ **시**간이 지남에 따라 평균적으로 값이 증가하거나 감소하는 형태, 상수 $\emptyset_0$ = 표류 (drift)

→ **AR 모형은 정상성data에만 사용한다.**

결국 AR(1)모형 : -1< $\emptyset_1$<1 , AR(2)모형 : -1< $\emptyset_2$<1 , $\emptyset_1 + \emptyset_2 < 1, \emptyset_2 - \emptyset_1 < 1$

정상성을 여부를 알아보기 위한 단위근 검정 방법을 해야 한다. !!!!

시계열 모형을 이용한 예측

AR(1) 모형: 단위근 검정

단위근 (Unit root) 검정 → $\emptyset_1 = 1$

- 시계열 데이터가 정상성(Stationarity)을 만족하는지 검증하는 통계적 방법
- 비정상성을 가진 데이터는 평균이나 분산이 시간에 따라 변하며, 분석 및 예측 모델링에 적합하지 않을 수 있다.
- 단위근 검정은 데이터가 정상 시계열인지, 비정상 시계열인지 판단하는 데 필수적이다.

정상성과 단위근

- **정상성:** 시계열 데이터의 평균과 분산이 시간에 따라 일정하고, 자기상관이 시차(lag)에만 의존.
- 단위근: AR(1) 모델에서 자기회귀 계수 ϕ=1인 경우 단위근을 가졌다고 한다.
- 데이터가 단위근을 가지면, 정상성을 만족하지 않으며, 데이터가 시간에 따라 랜덤 워크(Random Walk)를 따른다.

단위근 검정 방법

1) Augmented Dickey-Fuller Test (ADF Test):
- 가장 널리 사용되는 단위근 검정 방법.
- 데이터가 단위근을 가지는지 검증하기 위해 시차(lag)를 추가하여 확장된 Dickey-Fuller 테스트.

2) Phillips-Perron Test (PP Test)
ADF 검정과 유사하지만, 오차항의 이분산성(heteroskedasticity)을 보정.

3) Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS Test):
KPSS 검정은 정상성 여부를 직접 검증. → 귀무가설: 데이터가 정상성을 만족한다 (정상성).

4) Zivot-Andrews Test:
데이터에 구조적 변화를 고려한 단위근 검정.

단위근 검정 후 처리 방법

1) 비정상 데이터인 경우:
- 차분(Differencing): 데이터에서 트렌드 제거.
- 로그 변환(Log Transformation): 데이터의 분산을 안정화.

2) 정상 데이터인 경우:
ARIMA 모델 등 정상성을 가정한 분석 기법을 바로 적용 가능.

시계열 모형을 이용한 예측

단위근 (Unit root) 검정 방법

1. ADF (Augmented Dicky-Fuller) 검정

→ 1차 차분한 시계열이 정상시계열인지 비정상시계열인지 검정

Dicky-Fuller 검정통계량

$$DF = \frac{\widehat{\emptyset_1} - 1}{std(\widehat{\emptyset_1})}$$

$$\hat{\phi}_1 = \frac{\sum_{t=1}^{N} X_{t-1} X_t}{\sum_{t=1}^{N} X_{t-1}^2}$$

- **귀무가설(H_0) :** 시계열자료에 단위근 **(Unit root)**이 존재한다 (비정상성 시계열)
  → $\emptyset_1 = 1$, P-value > 유의수준(0.05)
- **대립가설 (H_1) :** : 시계열자료가 정상성을 만족한다. → **$|\emptyset_1| < 1$,**
  P-value < 유의수준(0.05) 귀무가설 기각

2. Phillips-Perron Test (PP Test)

ADF 검정과 유사하지만, 오차항의 이분산성(heteroskedasticity)을 보정.
단위근 검정의 가설

- **귀무가설(H_0)** : 데이터가 단위근을 가진다 (비정상 시계열). , P-value > 유의수준(0.05)
- 대립가설 (H_1): 데이터가 단위근을 가지지 않는다 (정상 시계열)

3. KPSS(Kwiatkowski-Phillips-Schmidt-Shin Test) 검정

KPSS 검정은 정상성 여부를 직접 검증.

- **귀무가설(H_0)** : 데이터가 정상성을 만족한다 (정상성시계열 , P-value > 유의수준(0.05).
- **대립가설(H_0)** : 데이터가 단위근을 가진다 (비정상 시계열)

https://github.com/freejyb/KDT
시계열분석/ **Unit root test**

시계열 모형을 이용한 예측

AR(1) 모형: 단위근 검정

Dicky-Fuller 검정통계량

$$DF = \frac{\widehat{\emptyset_1} - 1}{std(\widehat{\emptyset_1})}$$

정상성 조건 $|\emptyset_1| < 1$ 이유? 수학적 이해

$X_t = \emptyset_0 + \emptyset_1 X_{t-1} + \varepsilon_t$

정상성의 정의으로부터

임의의 시점인 t에서의 기대값이 모두 동일해야 한다.

$E(X_t) = E(X_{t-1}) = \ldots\ldots = \mu$

AR모형의 정의에 따르면 오차항 ε_t 의 기대값 = 0

$E(X_t) = E(\emptyset_0 + \emptyset_1 X_{t-1} + \varepsilon_t)$

$= \emptyset_0 + \emptyset_1 E(X_{t-1})$ (E(ε_t) = 0 이므로)

$E(X_t) = E(X_{t-1}) = \mu$ 를 양변에 대입하여 정리하면

$$\mu = \emptyset_0 + \emptyset_1 \mu$$

$$\mu = \frac{\emptyset_0}{1-\emptyset_1}$$

$\emptyset_0 = \mu(1-\emptyset_1)$이므로 이를 AR(1)모형에 대입하면

$X_t = \emptyset_0 + \emptyset_1 X_{t-1} + \varepsilon_t$

$X_t = \mu(1-\emptyset_1) + \emptyset_1 X_{t-1} + \varepsilon_t$

$X_t - \mu = \emptyset_1(X_{t-1} - \mu) + \varepsilon_t$

위 식에서 분산을 구하고 임의의 시점 t 에서 분산이 동일해야 한다는 정상성 정의를 이용하면

$X_t - \mu = \emptyset_1(X_{t-1} - \mu) + \varepsilon_t$

$Var(X_t) = \emptyset_1^2 Var(X_{t-1}) + Var(\varepsilon_t)$

$Var(X_t) = \frac{Var(\varepsilon_t)}{1-\emptyset_1^2}$ ($Var(X_t) = Var(X_{t-1})$이므로)

➔ 분산은 항상 ≥ 0 이므로, $\emptyset_1^2$은 1보다 작아야 한다.

이는 **결국 $-1 < \emptyset_1 < 1$ 되어야 한다는 의미 !!!!!**

시계열 모형을 이용한 예측 : ARIMA

모형의 식별 → 모형의 추정 → 모형진단

p, q 차수 결정

1. 차분 후의 plot을 보고 1차 차분이 적절할지? 2차 차분이 적정할지? 여부를 확인
2. ACF와 PACF의 시각화 했을 때 모양을 통해 ARIMA(AR, MA, ARMA) 모델의 hyperparameter p와 q를 결정

ACF(Auto Correlation Function)

y_t와 y_{t+k}사이의 자기상관

$$ACF(k) = \frac{\sum_{t=1}^{N-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{N}(y_t - \bar{y})^2}$$

- k시간 단위로 구분된 시계열의 관측치 간 **상관계수 함수**
- k(시차)가 커질수록 ACF는 0에 수렴 (시간차이가 많아지면 상관관계가 없는 것!!)

ACF를 통해서 정상성을 여부 확인

→ **일정한 패턴이 없는 경우, 갑자기 값이 감소하는 패턴 → Stationary**

• 일정하게 패턴을 갖고 변화하거나, 천천히(서서히) 감소하는 패턴 → Nonstationary

PACF(Partial ACF)

$$PACF(k) = \text{Corr}(e_t, e_{t-k})$$

$$e_t = y_t - (\beta_1 y_{t-1} + \ldots + \beta_{k-1} y_{t-(k-1)})$$

- 시차가 다른 두 시계열 관측치 간 상관관계 함수
- 시차 k에서의 k단계 만큼 떨어져 있는 두 데이터 점들간의 상관이므로 두 점 사이의 시차가 다른 데이터 영향은 제거 된 순수한 상관관계

시계열 모형을 이용한 예측 : ARIMA

p, q 차수 결정

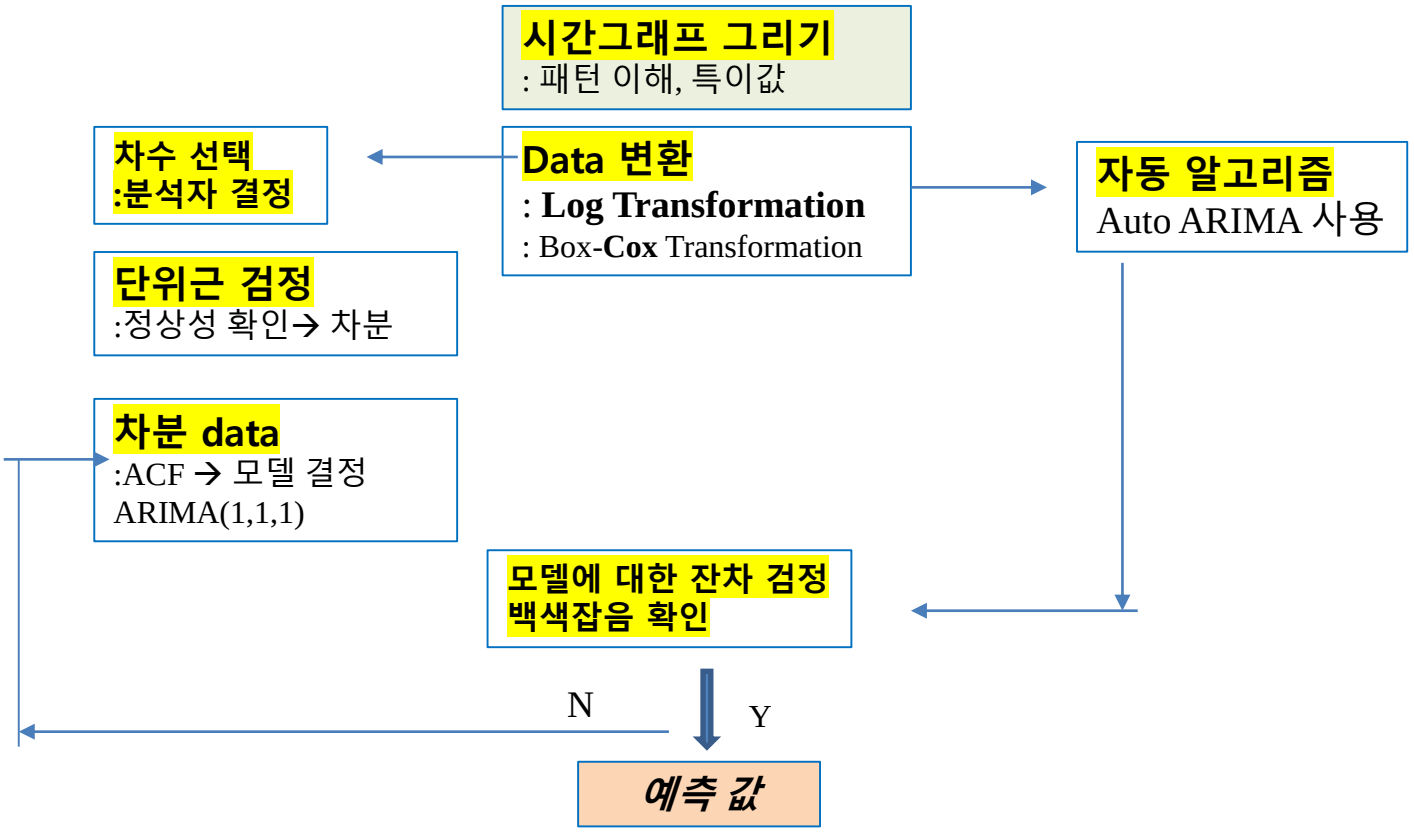
구분	ACF	PACF
AR(p)	시차(k)커질수록 점차 감소하여 0에 접근	P시차 이후 0에 접근
MA(q)	q시차 이후에 0 에 접근	시차(k)커질수록 점차 감소하여 0에 접근
ARMA(p,q)	시차(K)커질수록 점차 감소하여 0에 접근 (시차 q이 후 0에 접근)	시차(K)커질수록 점차 감소하여 0에 접근 (시차 p이후에 0에 접근)

ARIMA (p,d,q)로 표현

- ARIMA (1, 1,1)
- ARIMA (1,1,0)
- ARIMA (2,1,2)
- ARIMA (2,1,1)

시계열 모형을 이용한 예측

ARIMA 일반적 절차



https://github.com/freejyb/KDT/예측분석

ARIMA 확장

1. ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables)

- **ARIMAX**는 ARIMA 모델의 확장형
- 외생 변수(Exogenous Variables)를 추가하여 시계열 데이터를 모델링한다.
- 외생 변수는 시계열 데이터의 결과에 영향을 미치지만, 모델이 설명하려는 목표 변수에 의존하지 않는 독립 변수이다.

(ARIMAX 사용 예시)

•외부 요인(독립 변수)이 대상 데이터(종속 변수)에 영향을 미칠 때.
•경제학, 환경 데이터 분석, 마케팅 캠페인 효과 분석 등.

ARIMAX 모델의 구성 요소

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \beta X_t$$

Y t : 목표 시계열 데이터.
Xt : 외생 변수.
ϕi : AR(자기회귀) 계수.
θi : MA(이동평균) 계수.
β: 외생 변수의 계수.
d: 차분 횟수

ARIMA 확장

2. SARIMA (Seasonal ARIMA):

- **SARIMA**는 **ARIMA** 모델에 계절성(Seasonality)을 추가한 확장형 모델로, 계절적 패턴이 있는 시계열 데이터를 분석하고 예측하는 데 사용
- SARIMA는 시간 패턴 내에서 반복되는 계절적 변화를 효과적으로 처리할 수 있다.
- 계절적 패턴(예: 매년, 매달 반복되는 패턴)을 효과적으로 모델링.

(SARIMA의 구성)

1) ARIMA 구성 요소 (p,d,q)
비계절적 데이터의 자기회귀(AR), 차분(I), 이동 평균(MA) 모델링.

2) 계절성 구성 요소 (P,D,Q,s)
계절성을 설명하는 자기회귀, 차분, 이동 평균 및 계절 주기 s

p: 비계절적 자기회귀(AR) 차수.
d: 비계절적 차분(I) 차수.
q: 비계절적 이동 평균(MA) 차수.
P: 계절적 자기회귀(AR) 차수.
D: 계절적 차분(I) 차수.
Q: 계절적 이동 평균(MA) 차수.
s: 계절 주기 (예: 12개월, 7일 등).

(SARIMA 모델의 일반적인 표기법)

- : $(p,d,q)\times(P,D,Q,s)$ 여기서 s는 계절 주기.

(SARIMA의 필요성)

- ARIMA 모델은 계절적 패턴을 명시적으로 처리하지 못한다.
- 계절성을 처리하지 않고 모델링하면 예측 오류가 커질 수 있다.
- SARIMA는 계절적 주기와 일반적인 시계열 패턴을 동시에 처리한다.

SARIMA의 주요 단계

1.데이터 탐색: 데이터가 계절성을 가지는지 확인 (예: ACF/PACF 분석).
계절 주기 s를 식별.
2. **차분(Differencing)**:계절적 차분(D)과 비계절적 차분(d)을 사용해 정상성을 확보.
3. 모수 추정: (p,d,q) 및 (P,D,Q,s) 값을 선택.
4. 모형 적합 : SARIMA 모델 적합 및 검증.
5. 예측 및 평가:테스트 데이터에 대한 예측 수행.성능 지표(RMSE, MAE 등)로 모델 평가.

- **SARIMA의 장점** : 계절적 패턴과 비계절적 패턴을 동시에 처리. 다목적이고 다양한 데이터에 적용 가능.
- **SARIMA의 단점** :모수 설정 (p,d,q,P,D,Q,s)이 복잡., 계산 비용이 ARIMA보다 높음.

ARIMA 확장

2. SARIMA (Seasonal ARIMA):

잔차 진단 지표

- Ljung-Box (L1) (Q) 잔차의 자기상관을 검증 :
- Prob(Q) p>0.05, 잔차가 백색잡음임
- Jarque-Bera (JB) 잔차의 정규성을 검증 : Prob(JB) p>0.05 잔차가 정규분포를 따름
- Heteroskedasticity (H) 잔차의 이분산성을 검증 : Prob(H) p>0.05 잔차가 이분산성을 가지지 않음

SARIMA vs. ARIMA

특징	ARIMA	SARIMA
계절성 처리	계절성을 명시적으로 처리하지 못함.	계절적 패턴을 효과적으로 모델링.
적합성	비계절적 데이터에 적합.	계절적 주기가 명확한 데이터에 적합.
구성 요소	(p, d, q)	(p, d, q) × (P, D, Q, s)

ARIMA 확장

3. SARIMAX (Seasonal ARIMA with Exogenous Variables):

- **SARIMAX**는 **SARIMA** 모델에 외생 변수(Exogenous Variables)를 추가한 확장형 모델
- SARIMAX는 계절적 패턴, 비계절적 패턴, 외생 변수의 효과를 모두 포함하여 시계열 데이터를 모델링하고 예측한다.

SARIMAX 모델의 구성

1. SARIMA 구성 : $(p,d,q)\times(P,D,Q,s)$
- SARIMA의 계절성과 비계절적 패턴을 설명.
- s: 계절 주기 (예: 월별 데이터에서는 s=12).

2. 외생 변수 (Xt):
- SARIMAX는 시계열 데이터 Yt 를 설명하기 위해 외부 변수(독립 변수)를 포함한다.
- 외생 변수는 시점별로 변화하며, 목표 변수에 영향을 미치는 요인이다.

SARIMAX의 수식:

$$Y_t = \phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)\epsilon_t + \beta X_t$$

- Y t : 목표 시계열 데이터.
- Xt : 외생 변수.
- β: 외생 변수의 계수.
- $\phi(B)$, $\Phi(Bs)$): 비계절적 및 계절적 자기회귀(AR) 계수.
- $\theta(B)$, $\Theta(Bs)$: 비계절적 및 계절적 이동 평균(MA) 계수.
- B: 시차 연산자 (backshift operator).

SARIMAX의 주요 특징

- 계절적 및 비계절적 패턴 처리:SARIMA의 계절성과 비계절적 모델링을 그대로 포함.
- 외생 변수의 추가:외부 요인이 목표 변수에 미치는 영향을 모델링 가능.
- 예: 주택 가격 예측에서 금리(외생 변수) 포함.
- 정규성 가정 불필요:시계열 데이터의 특성(정규성, 등분산성)에 유연하게 대응.
- 유연성:SARIMAX는 외생 변수 없이 SARIMA로 동작 가능.

SARIMAX의 활용

- 경제학/금융: 예: 주가 예측에서 외부 경제 지표(금리, 환율 등)를 포함.
- 마케팅: 예: 매출 예측에서 광고 비용, 할인 프로모션 효과를 포함.

ARIMA 확장

3. SARIMAX (Seasonal ARIMA with Exogenous Variables):

SARIMAX 모델링 단계

1. 데이터 탐색
: 계절성 확인 (s) 및 ACF/PACF 분석.
외생 변수와 목표 변수 간의 상관관계 분석.
2. 모델 구성
: 비계절적 차수 (p,d,q), 계절적 차수 (P,D,Q,s), 외생 변수 포함.
3. 모수 추정 및 검증
: 각 계수의 통계적 유의성 검토 (p-값 확인).
4. 예측 및 평가:
테스트 데이터를 사용해 예측 성능 평가 (RMSE, MAE 등).

SARIMAX vs. SARIMA

특징	SARIMA	SARIMAX
외생 변수 포함 여부	외생 변수를 포함하지 않음	외생 변수를 포함하여 추가적 영향 모델링 가능
복잡성	더 간단한 구조	외생 변수로 인해 추가적인 복잡성 도입
적용 데이터	계절적 패턴이 뚜렷한 데이터	계절적 패턴과 외생 변수 영향을 받는 데이터

SARIMAX 모델의 장.단점

SARIMAX 모델의 장점
- 유연성:계절성, 비계절성, 외생 변수까지 포함한 복합적 데이터 모델링 가능.
- 외생 변수의 효과 측정:외생 변수의 계수를 통해 종속 변수에 미치는 영향 평가 가능.

SARIMAX 모델의 단점
- 복잡성 증가:외생 변수 선택과 모델 구성에 따른 복잡성.
- 계산 비용: SARIMA보다 높은 계산 비용.

회귀모형을 이용한 예측

- **Y : forecast variable**
- **X: predictor variables**

- 상관계수 → 시차 상관계수 : 변수 간 선행. 후행 확인
- 두 시계열의 산점도 : 두 변수 관계 파악

모형 설정 → 모형의 추정 → 모형진단/예측

- 다중회귀분석
- 설명변수 선택
- 다중공선성 문제
- 모형 간결의 원칙
(효율성과 편의의 상충 고려)
- scatterplot matrix

- 회귀계수 추정
: 최소제곱법(least square),
최대우도추정법.
- 유의성 검정 : p-value < 유의수준
- 편회귀계수 성격
- **Goodness-of-fit**
: R^2 , $adjR^2$

- 자기상관여부 검정
: Durbin-Watson 모형 (DW ≅ 2)
- 잔차(residual)의 자기상관
: 잔차의 ACF (AutoCorrelation Function)
잔차(residual) → **Breusch-Godfrey)** 검정 (LM (Lagrange Multiplier) 검정)
- **잔차 정규성**
: 잔차의 히스토그램
- **적합값에 대한 잔차 그래프**
: 오차에 heteroscedasticity

회귀모형을 이용한 예측
: 설명변수 값을 예측하고
모형에 대입하여 종속변수 예측 함

일반화된 시계열 회귀모형

- 안정적 시계열을 바탕으로 함.
- 불안정시계열인 경우 차분을 통해 안정적 시계열로 전환 후 모형 작성, 또는 공적분관계 모형을 이용
- 회귀계수와 ARMA 계수 추정의 반복적 과정을 통해 오차항(ε_t)이 백색잡음이 될 때 회귀계수 추정

VAR모형을 이용한 예측

- **ARIMA 모형** : **단변량** 시계열 모형 → **정상성** 만족해야 함.
- **VAR 모형** : **다변량** 시계열 모형 → 정상성 + **비정상성 시계열에도 사용 가능**, 시계열의 벡터를 예측하는 단변량(univariate) 자기회귀 모델의 일반적인 형태

VAR (Vector Auto Regression)

- Sims (1980) : 내생변수와 외생변수를 구분하지 않고 계수값에 대한 제약을 고려하지 않으면서 시계열이 나타내는정보만 이용하는 VAR모형 개발
- 과거 회귀분석, 계량모형이 이론적 근거나, 선험적 판단 등을 토대로 내생변수 및 외생변수(종속변수 및 설명변수)를 결정하여 구조모형 선정
- **시계열 변수 사이에서 나타나는 동태적 관계 분석 : 각 시계열 변수가 서로 영향을 주며 이를 고려해 각 변수의 미래값을 전체 시계열 변수의 과거값으로부터 예측하므로 양방향 모형**
- 변수 사이의 이론적 관계를 고려하지 않고 간단히 예측 실행 가능 → 이에 대한 비판으로 구조적(Structural) VAR 모형 개발

벡터자기회귀모형 : $x_t \sim VAR(1)$

$$x_t = \mu + Ax_{t-1} + \varepsilon_t$$

$$\begin{pmatrix} m_t \\ y_t \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{pmatrix} \begin{pmatrix} m_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

벡터자기회귀모형 : $x_t \sim VAR(p)$

$$x_t = \mu + A_1x_{t-1} + A_2x_{t-2} \; .. + + .. A_Px_{t-p} \; \varepsilon_t$$

$$\begin{pmatrix} m_t \\ y_t \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} m_{t-1} \\ y_{t-1} \end{pmatrix} + .. \; ... + \begin{pmatrix} \alpha_{1p} & \beta_{1p} \\ \alpha_{2p} & \beta_{2p} \end{pmatrix} \begin{pmatrix} m_{t-p} \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

차수 p 결정

- **모형 선정 기준** : **AIC와 BIC 기준값을 최소화하는** 차수p 를 최적 차수로 선정 함.

  → 차수가 커지면 모형 적합도는 높아지지만, 모형의 모수가 많아져 예측력이 저하되는 경향이 있음.

https://github.com/freejyb/KDT

예측분석/**VAR모형을 이용한 예측**

VAR모형을 이용한 예측

VAR 모형 활용

구분	내용	방법
인과관계분석	VAR 모형의 계수에 대한 가설검정을 통하여 VAR 모형에 포함된 변수 사이에 인과관계 존재 여부 를 파악하는 방법 **Granger Causality Test** ➔ 단일 시계열보다 다중 시계열로 설명력을 더 얻을 수 있는가? 를 확인해 보는 테스트 귀무가설: X가 Y에 영향을 미치지 않는다	F통계량 or Wald통계량으로 검정
충격반응분석	VAR 모형에 포함된 여러 변수 중 한 변수에 충격이 나타났을 때 시간이 경과함에 따라 다른 구성변수에 어떤 영향을 미치는가를 파악하는 방법	
예측오차 (분산) 분해	VAR모형에 구성변수의 상대적 중요성의 정도를 파악하는 방법 → VAR 모형의 각 변수에 대해 미래 값에 대한 예측오차를 각 변수에 의해 발생하는 비율로 분할	벡터 y_t를 VMA 모형으로 나타내고, 오차항 ε_t를 직교화된 오차항 u_t로 나타낸 후 MSE 최소화
예측	VAR 모형은 모든 변수가 내생변수 → 특정한 조건 없이 예측값 추정 → 최적예측값 : MSE 최소값	

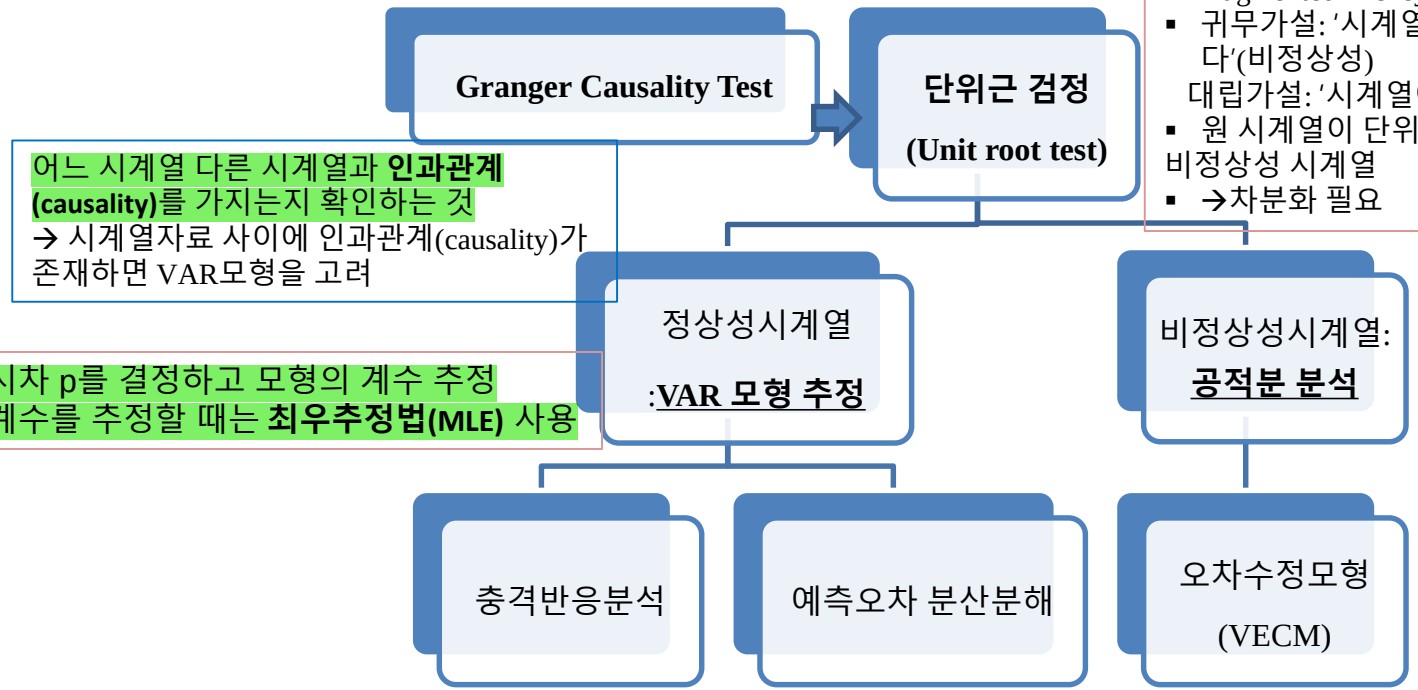
VAR 모형 문제점

- 변수 설정, 배열순서, 시차길이 등에 따라 변수사이의 파급효과에 대한 분석 결과 상이
- 추정해야 할 모수가 많아 예측력 이 저하 될 수 있음.

VAR모형을 이용한 예측

경제시계열 데이터 분석법

VAR(Vector Autoregression) 모형 분석



- **단위근 : 근 중에서 크기가 1인 근이 있으면 그 근을 단위근(unit root)**
- **시계열이 정상성 (Stationarity)여부 판단**
- Augmented Dickey Fuller 검정
- 귀무가설: '시계열에 단위근이 존재한다'(비정상성)
  대립가설: '시계열이 정상성을 만족한다'
- 원 시계열이 단위근을 가지고 있으면 비정상성 시계열
- →차분화 필요

어느 시계열 다른 시계열과 **인과관계 (causality)**를 가지는지 확인하는 것
→ 시계열자료 사이에 인과관계(causality)가 존재하면 VAR모형을 고려

시차 p를 결정하고 모형의 계수 추정
계수를 추정할 때는 **최우추정법(MLE)** 사용

예측오차의 분산분해 (forecasting error variance decompositions)란?

- 미래값을 예측하고 미래예측 오차분산을 시계열별로 분해
- 한 변수의 변화에 관한 예측오차를 각 변수들에 의해서 발생하는 비율로 분할하는 것으로 이를 이용하여 한 변수의 변화를 설명함에 있어 모형내 각 충격의 상대적중요도를 측정할 수 있음

kdigital/재무/VAR을 활용한 다변량 예측 모델링.ipynb

공적분 분석을 이용한 예측 : VECM

공적분 관계 (Cointegration relationship) 란?

- ***공적분 관계가 존재하지 않는 경우:차분(d) 변환하여 VAR모형으로 추정***
- ***공적분관계가 존재하는 경우: VECM모형으로 추정***

- 공적분 관계가 있는 시계열 변수 사이의 통계적 성질

➔ **단기적으로는 다를 수 있지만 장기적으로는 서로 일정한 관계(균형관계) 를 가지는 것**

- 시계열의 적분 차수가 모두 *d*일 때, 시계열의 선형결합(linear combination)의 적분 차수가 *d*보다 낮아질 때 시계열 사이에 공적분 (cointegrated) 관계가 존재한다고 한다.
- **$(n \times 1)$ 차원 벡터 x_t 를 구성하는 시계열이 불안정한 시계열이라도, 1차 적분 과정 I(1) 이며, $(n \times 1)$차원 벡터 $\beta(\neq 0)$가 존재하고 x_t 선형결합** (linear combination) **인 $\beta' x_t$가 안정적인 I(0) 과정 일 때**

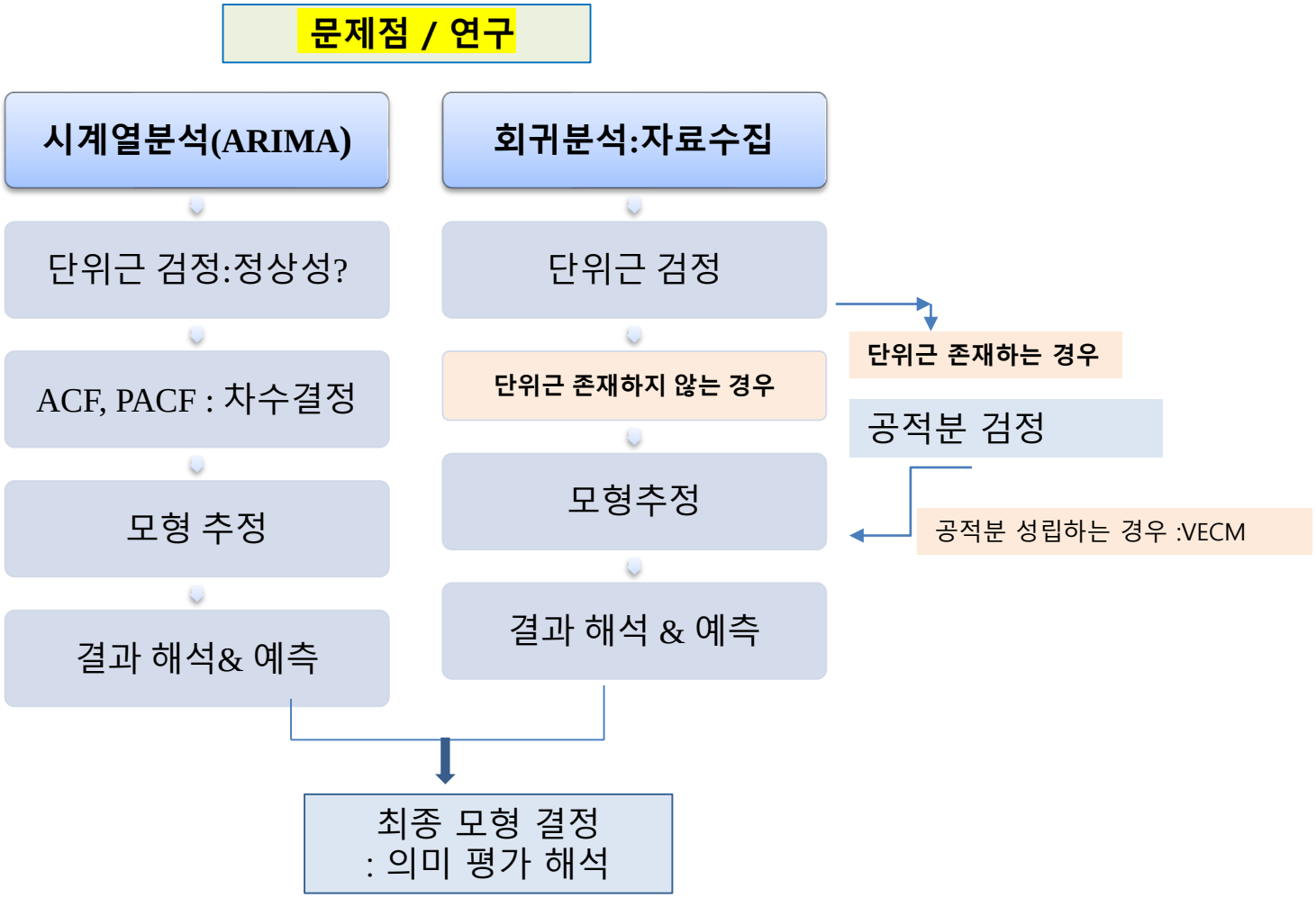
- **벡터 x_t 가 공적분 관계에 있다고 한다.**
- **벡터 β를 공적분 벡터라고 함.**

- 불안정시계열을 대상으로 회귀분석을 하는 경우 일반적인 가설검정기법을 적용하면, 실제로는 변수 사이에 아무 상관관계가 없는데도 불구하고 회귀식 추정결과가 외견상 유의성이 높은 상관관계가 있는 것처럼 나타나는 **가성적 회귀 (spurious regression) 현상이 발생한다**. → 특히 대부분 경제 시계열의 경우 가성적 회귀를 갖는 불안정적 시계열 임.
  - 두 개의 불안정시계열 사이에 균형관계가 나타나고, 이러한 균형관계로 부터의 괴리가 안정적일 경우 두 변수 사이에는 공적분 관계가 존재한다.

C.W.J. Granger
공적분검정 은 가성회귀 상황을 피해갈 수 있는 사전검정(pre-test)으로 간주될 수 있다"고 지적 함.
적분 차수(Order of integration) : 정상성 상태의 시계열을 얻기 위해 필요한 차분 횟수

공적분 분석을 이용한 예측

공적분검정 일반적 절차



공적분 분석을 이용한 예측 : VECM

VECM 적용하는 이유 ?

VAR 모형은 각 시계열이 안정성 조건을 만족하지 않아도 사용할 수 있다.

→ ***그러나 일반적으로 불안정성 시계열의 경우 차분을 하거나 변수간 장기적 관계에 대하여 정보를 상실할 수 있다는 단점이 있다.***

→ ***불안정적인 시계열 변수사이에도 공적분 관계가 존재한다고 판정되면 VECM 이용 가능 함.***

→ ***VECM 은 변수간 공적분 관계에 있는 시계열은 차분을 거치지 않고 원 데이터를 써서 모형에 적합 시킬 수 있다는 점에서 장점을 가진다.***

→ 두 개 이상의 경제시계열의 장기적 관계가 안정적일 때(공적분 관계가 존재할 때) 차분방정식에 장기적 관계에 대한 균형오차를 추가한 **오차수정모형**을 만들 수 있다.

공적분 분석을 이용한 예측

공적분 검정(Cointegration Tests)

Cointegration 검정은 다중 시계열이 장기적 기간을 두고 안정적인 연관성을 보이는가? 를 확인하는 것 !!

Engel & Granger 검정

- -회귀분석 결과의 잔차항에 대해 검정

-N개의 비정상시계열 사이에는 일반적으로 N-1개까지의 공적분 관계가 존재할 수 있다

-Engel & Granger 공적분검정 : 세 개이상의 비정상 시계열 사이의 공적분검정부터 한계를 가짐

• Johansen 검정 (Johansen test) 1991

– 벡터 형태로 검정, Engel & Granger 공적분 검정의 한계를 극복하는 검정

→ 공적분관계의 수와 모형의 파라미터들을 최우추정법(MLE)으로 추정 검정하는 방법

→ 모든 변수를 내생변수로 간주 한다는 점에서 종속변수를 선택할 필요가 없으며 여러 개의 공적분관계를 식별해 낼 수 있음.

→ 최우추정법을 이용하여 VAR 모형으로 공적분관계를 추정하는 한편 우도비검정(likelihood ratio test)을 바탕으로 공적분계수를 결정 할 수 있도록 함.

- **귀무가설(H_0) : 공적분의 관계가 없다. P-value > 유의수준 (0.05)**
- **대립가설(H_1) : 공적분의 관계가 있다. P-value < 유의수준 (0.05)**

다변량 시계열분석에 의한 Johansen's cointegration test이 다른 공적분 검정법 보다 우월한것으로 인정되어 널리 사용되고 있다.

공적분 분석을 이용한 예측 : VECM

오차수정모형(ECM: Error Correction Model , Vector Error Correction Model, VECM)

비정상시계열을 대상으로 회귀분석을 하는 경우에 차분하여 정상시계열로 만들어 회귀식을 추정

→ 통계적 문제는 해결되나 **장기적인 관계에 대한 정보는 손실 발생한다.**

-공적분관계가 존재하는 경우 오차수정모형(ECM) 을 이용하면 장기적 균형관계에 대한 정보와 함께 단기적 움직임도 동시에 파악할 수 있음

비정상적인 시계열 변수들의 선형결합이 안정적인 과정이 될 때, 즉 공적분 관계가 존재하는 변수들에 대해 적용할 수 있다. (활용) 벡터오차수정 모형은 변수들의 장기적 균형관계와 단기적 동적관계를 동시에 이해하는 데 유용하기에 다양한 금융시장 실증분석에서 활용

오차수정모형(ECM)의 구조

2개의 **비정상시계열** Y_t , X_t 사이에 **공적분 관계가 존재하는** 경우

1) 회귀분석을 통해 회귀계수 (α , β) 를 계산하고 잔차항 구함. $y_t = \alpha + \beta X_t + \varepsilon_t$

(β : Y 와 X 시계열간의 장기관계를 나타내는 계수)

2) 위 식에서 도출된 잔차를 이용하여 다음 식 추정함. $\Delta Y_t = \gamma_0 + \sum_{i=0}^{k} \gamma_0 \Delta X_{t-i} + \delta \varepsilon_{t-1}$

(δ : 장기 균형점에서 이탈했을때 장기균형점으로의 복귀 속도, 만약 $\delta < 0$ 일 때 Y 는 균형점에 안정적으로 접근)

→ 모든 변수들이 안정적인 시계열이기 때문에 가성적 회귀문제는 발생하지 않음

→ 잔차항 ε_t 를 통해 수준변수가 갖고 있는 정보 ($\varepsilon_t = y_t - \alpha - \beta X_t$)를 반영

→ 동시에 차분변수 (ΔY_t , ΔX_t) 가 갖고 있는 정보를 하나의 모형 내에 포함하는 구조

공적분 분석을 이용한 예측 : VECM

공적분을 활용한 Pair Trading (Cointegration-Based Pair Trading)

두 개별주식의 주가가 일시적으로 격차(spread) 가 벌어졌거나 , 격차가 좁혀 졌을때,
이 현상이 시간에 따라 정상 수준으로 회귀할 것으로 예측되는 경우 활용할 수 있는 투자 전략

Pair Trading이 가능하려면 **시간 경과에 따른 비율의 기대값이 평균에 수렴해야 한다.**

→ 즉, 공적분 관계가 성립되어야 함을 의미.

공적분의 의한 방법론은 주가가 비정상 시계열임을 가정하고 스프레드의 정상성을 추구한다.

A 주식의 가격 = 추세성분 + 잔차성분

B 주식의 가격 = 추세성분 + 잔차성분

A의 추세성분과 B의 추세성분이 선형적인 관계에 있다면 B의 추세성분에 일정한 상수 n를 곱하여 A에 값에서 빼줌으로써 A의 주가 - n(B의 주가)의 추세성분을 제거 (long + short)

두 자산의 스프레드에서 추세성분이 사라지고 잔차성분만 남게된 경우, 두 자산이 공적분 관계에 있다고 한다.

→ 이때 , 잔차성분이 정상성을 띄게 되면, 통계적으로 평균으로 회귀하는 성질을 활용하여,벌어진 스프레드에 투자하여 차익거래가 가능하게 된다.

➔ 공적분 계수 (Cointegration Coefficient)

이자율 기간 구조 분석에 관한 연구

공적분 분석을 이용한 예측 : VECM

공적분을 활용한 Pair Trading (Cointegration-Based Pair Trading)

공적분 계수 (Cointegration Coefficient)

- 두 자산 간에 공통추세가 존재한다면
- 공적분 계수를 비율로 롱-숏을 구성하여 공통추세를 제거할 수 있고 (시장중립모형 구축),
- 손익 결정 : 롱-숏 포트폴리오의 Pay Off는 추세가 없는 잔차 성분만 남게 된다.
- 공적분 계수로 공통추세가 제거된다면, 남은 잔차 성분은 정상성 (Stationary)이 높아지게 된다.
- **공적분 계수는 투자 비율을 결정짓는 요소이기 때문에 매우 중요한 요소**

공적분계수 추정 Cointegration Coefficient (CC)

a종목의 Cointegration 계수 = $\frac{Cov(R_a, R_b)}{a\text{종목 수익률의 분산}}$

➔ ***b종목에 대한 a종목의 상대적 베타 β 계수***

https://github.com/freejyb/kdigital
재무
공적분활용 예측.ipynb

LSTM을 활용한 주가 예측

RNN- LSTM- GRU(Gated Recurrent Unit)

Long Short-term Memory 1997 Sepp Hochreiter Johannes Kepler University Linz

https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

A deep learning framework for financial time series using stacked autoencoders and longshort term memory
Wei Bao1 , Jun Yue2 *, Yulei Rao1 1 Business School, Central South University, Changsha, China, 2 Institute of Remote Sensing and Geographic Information System, Peking University, Beijing, China 2017

Understanding LSTM Networks

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

https://www.kaggle.com/taronzakaryan/predicting-stock-price-using-lstm-model-pytorch?scriptVersionId=59632795

Predicting Stock Price using LSTM model, PyTorch

kaggle.com/purplejester/a-simple-lstm-based-time-series-classifier

A Simple LSTM-Based Time-Series Classifier

https://curiousily.com/posts/time-series-forecasting-with-lstm-for-daily-coronavirus-cases/

Time Series Forecasting with LSTMs for Daily Coronavirus Cases using PyTorch in Python

* Prophet (2018)

시계열 예측을 위한 Facebook Prophet 사용하기

https://facebook.github.io/prophet/docs/installation.html#python
https://facebook.github.io/prophet/docs/quick_start.html

Forecasting at Scale

Sean J. Taylor*†
Facebook, Menlo Park, California, United States
sjt@fb.com

and

Benjamin Letham†
Facebook, Menlo Park, California, United States
bletham@fb.com

Abstract

Forecasting is a common data science task that helps organizations with capacity planning, goal setting, and anomaly detection. Despite its importance, there are serious challenges associated with producing reliable and high quality forecasts – especially when there are a variety of time series and analysts with expertise in time series modeling are relatively rare. To address these challenges, we describe a practical approach to forecasting "at scale" that combines configurable models with analyst-in-the-loop performance analysis. We propose a modular regression model with interpretable parameters that can be intuitively adjusted by analysts with domain knowledge about the time series. We describe performance analyses to compare and evaluate forecasting procedures, and automatically flag forecasts for manual review and adjustment. Tools that help analysts to use their expertise most effectively enable reliable, practical forecasting of business time series.

Keywords: Time Series, Statistical Practice, Nonlinear Regression

PROPHET

step is installing Rtools before attempting to ins

If you have custom Stan compiler settings, insta

Installation in Python

Prophet is on PyPI, so you can use `pip` to install

```
python -m pip install prophet
```

- From v0.6 onwards, Python 2 is no longer su
- As of v1.0, the package name on PyPI is "pr
- As of v1.1, the minimum supported Python

After installation, you can get started!

* Prophet (2018)

시계열 예측을 위한 Facebook Prophet 사용하기

- **Prophet**은 시계열 데이터를 분석하고 미래를 예측하기 위해 Facebook에서 개발한 오픈 소스 도구로, 특히 불규칙한 계절성을 가진 시계열 데이터를 처리하는 데 탁월
- Prophet은 데이터의 주기성, 추세, 불규칙성을 자동으로 학습해 시계열 모델링을 쉽게 할 수 있는 도구.
- Prophet은 시계열 예측을 위해 **추세(Trend)**, **계절성(Seasonality)**, 휴일/이벤트 효과(Holiday/Event Effects)를 기반으로 데이터를 분석하고, 이를 조합하여 예측하는 모델.

1. 추세(Trend) 모델링
Prophet은 **추세 변화 지점**(changepoints)을 찾아 시계열 데이터의 전반적인 추세를 포착한다.
데이터가 시간이 지나면서 증가하거나 감소하는 경향을 분석하는데, Prophet은 주로 두 가지 추세 모델을 지원한다.
•**선형 추세 모델(linear growth)**: 시간에 따라 직선 형태로 증가 또는 감소하는 추세를 나타낸다.
•**비선형 추세 모델(logistic growth)**: 성장에 한계가 있는 경우, S자 형태의 추세를 반영하여 데이터가 점차 포화 상태에 다다르는 과정을 모델링한다.
추세 변화 지점을 통해 Prophet은 데이터의 변화 패턴이 급격하게 달라지는 시점을 자동으로 탐지하고, 해당 시점마다 추세를 조정할 수 있습니다. 사용자는 이 지점을 직접 설정하거나 Prophet이 기본적으로 탐지한 지점을 사용할 수 있다.

2. 계절성(Seasonality) 모델링
Prophet은 데이터 내 **주기적인 패턴**을 찾고 이를 기반으로 예측에 반영합니다. 계절성은 크게 다음과 같은 주기로 나뉜다:
•**연간 계절성**: 매년 같은 시기에 반복되는 패턴(예: 겨울철의 매출 감소)
•**주간 계절성**: 매주 반복되는 패턴(예: 주중과 주말의 판매량 차이)
Prophet은 **푸리에 시리즈**(Fourier series)를 이용해 연간과 주간 계절성을 모델링합니다. 푸리에 시리즈는 주기적인 패턴을 여러 주파수 성분의 조합으로 표현하며, 이를 통해 복잡한 계절성을 효과적으로 반영할 수 있다.

3. 휴일 및 이벤트 효과(Holiday/Event Effects)
Prophet은 공휴일이나 특정 이벤트가 데이터에 미치는 영향을 모델링할 수 있습니다. 이를 통해 주기적인 공휴일(예: 연말연시)이나 특별한 이벤트(예: 블랙프라이데이)로 인해 발생하는 데이터 변동을 반영할 수 있다.
•사용자가 정의한 공휴일이나 이벤트 데이터를 추가로 입력하여 Prophet이 이를 예측에 반영하도록 설정할 수 있다.
•이 효과는 예측에 큰 영향을 줄 수 있으며, Prophet은 특정 휴일의 이전과 이후에 발생하는 변화를 반영해 예측을 보다 정교하게 만든다.

* Prophet (2018)

Prophet의 기본 수식

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- g(t): 추세 함수를 나타내며, 시간에 따른 전반적인 증가/감소를 반영
- $s(t)$: 계절성 함수를 나타내며, 연간 또는 주간 패턴
- $h(t)$: 공휴일이나 이벤트의 영향
- ϵt : 데이터의 불규칙한 변동을 나타내는 오차.

Prophet의 장점

- **해석 용이성**: 예측 결과가 각각의 요소(추세, 계절성, 이벤트)로 나뉘어 해석하기 쉽다.
- **사용 편의성**: 간단한 인터페이스와 최소한의 파라미터 조정으로 쉽게 예측 모델을 설정할 수 있다.
- **유연성**: 공휴일 및 이벤트를 예측에 반영할 수 있으며, 추세와 계절성의 패턴을 사용자 정의할 수 있다.

파이썬
딥러닝(내부) prophet

* Decomposition-based Linear Model : DLinear

- DLinear 모델은 시계열 데이터의 예측을 위해 설계된 딥러닝 모델 중 하나로, 주로 롱 시퀀스 예측(Long Sequence Forecasting) 문제에서 사용
- 시계열 데이터를 분해하여 보다 단순한 선형 회귀를 통해 예측하는 접근 방식
- 특히 복잡한 모델이 오히려 예측 성능을 저하할 수 있는 긴 시계열 문제에서 효과적인 성능을 보이는 것으로 알려져 있다.

DLinear의 주요 개념

- **Decomposition**: 입력 시계열 데이터를 분해하여, 전체 시계열을 두 가지 이상의 독립된 시계열로 나누는 과정이다.
- **Linear Layer**: 분해된 각 시계열을 예측하기 위해 선형 레이어를 사용하며, 이는 추세와 계절성을 개별적으로 학습하고 이후 합치는 방식으로 최종 예측을 만든다.

DLinear 모델의 특징

- 시계열 분해 (Time Series Decomposition) :
  - DLinear는 시계열 데이터를 추세(Trend)와 계절성(Seasonality)으로 분해하는 접근을 취한다.
  - 각각의 구성 요소를 따로 모델링하고 나중에 결합함으로써 더 나은 예측 성능을 달성한다.
- 선형 모델 기반 (Linear Model Based)
  - DLinear는 복잡한 비선형 신경망 대신 각 분해된 시계열에 대해 선형 모델을 사용하여 예측한다.
  - 이로 인해 모델의 학습 속도가 빠르고, 과적합(overfitting) 가능성이 줄어든다.
- 단순함과 효율성 (Simplicity and Efficiency)
  - 딥러닝 모델들이 종종 매우 복잡하고 많은 데이터로 인해 학습이 오래 걸리는 것에 비해, DLinear는 단순한 모델링으로 효율성을 높였다.
  - 긴 시계열에 대해 상대적으로 간단한 구조로도 높은 성능을 보이는 것이 특징이다.

추세 (Trend):
추세는 데이터의 장기적인 변화 방향을 나타내며, 시간이 지남에 따라 증가하거나 감소하는 경향을 의미
DLinear 모델에서는 trend_linear 레이어가 입력 데이터의 추세를 학습한다. 이를 통해 장기적인 패턴을 포착하고, 전체 데이터의 증가 또는 감소 방향을 예측할 수 있다.
계절성 (Seasonality):
계절성은 데이터가 일정 주기를 가지고 반복되는 패턴을 나타낸다.
seasonality_linear 레이어는 이러한 반복적인 변화를 학습하는 역할을 한다. 이를 통해 주기적인 패턴을 정확히 모델링하여 예측에 반영할 수 있다.

* Decomposition-based Linear Model : DLinear

DLinear 모델의 학습 과정

- DLinear 모델은 입력 데이터 x를 각각의 선형 레이어에 통과시켜 추세(Trend)와 계절성(Seasonality)을 따로 계산합니다.
- forward() 함수: trend_linear와 seasonality_linear를 각각 호출하여 추세와 계절성을 계산하고, 최종 예측 값은 이 둘을 더하는 방식으로 만들어진다.
- 이를 통해 데이터의 장기적인 변화와 반복적인 패턴을 모두 고려하여 더 정확한 예측 값을 생성한다.

왜 시계열 분해를 사용하는가?

- 시계열 데이터를 추세와 계절성으로 나누면, 각각의 패턴을 더 명확하게 학습할 수 있다.
- 복잡한 시계열 데이터를 단순화하여 각 부분을 별도로 학습하게 되면 과적합을 방지할 수 있고, 더 빠르고 정확한 예측이 가능하다.
- 이처럼 두 가지 요소를 따로 분해해서 학습한 다음 결합함으로써, 복잡한 비선형 모델 대신 단순한 선형 모델로도 좋은 성능을 낼 수 있게 된다.

사용 사례

- DLinear 모델은 특히 전력 소비 예측, 재고 수요 예측, 금융 시장 데이터 분석 등 긴 시간에 걸친 시계열 데이터를 다룰 때 유용
- 이 모델은 데이터의 패턴을 학습하고 비교적 단순한 방법으로 추세와 계절성을 분리하여 미래 값을 예측한다.

장점

- 모델 단순성: 모델이 복잡하지 않아 학습과 해석이 용이하다.
- 빠른 학습: 다른 복잡한 딥러닝 모델에 비해 학습 속도가 빠르다.
- 긴 시계열에 적합: 긴 시계열 데이터에 대해 효과적으로 학습할 수 있다.

https://github.com/freejyb/KDT
시계열분석/Dlinear모델

* TimesNet (2023)

TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis

시계열 데이터를 더 정확히 분석하기 위해 1D 데이터를 2D 형태로 변환하여 모델링하는 새로운 접근법을 소개한다.

이 모델은 **다중 주기성**을 고려해, 시간 패턴을 더욱 정밀하게 분석하고자 설계되었다.

TimesNet의 주요 구성 요소

1.시계열 데이터의 2D 변환:

1. 기존의 1D 시계열 모델은 주파수와 주기가 복잡하게 얽힌 패턴을 포착하는 데 한계가 있습니다. TimesNet은 주파수 성분을 분석하는 고속 푸리에 변환(FFT)을 사용해 시계열 데이터를 2D 텐서 형태로 변환한다.
2. 변환된 2D 텐서의 행(row)은 **기간 간 변이(interperiod variation)**를, 열(column)은 기간 내 변이(intraperiod variation)**를 나타냅니다. 이를 통해 지역적 패턴과 전체적 패턴을 모두 포착할 수 있다.

2.TimesBlock 모듈:

1. TimesNet의 핵심 모듈인 **TimesBlock**은 인셉션 구조에서 영감을 받은 파라미터 효율적인 다중 스케일의 2D 필터를 사용하여 2D 텐서를 처리합니다. 이를 통해 각 2D 텐서 내의 다양한 스케일의 특징을 효율적으로 포착한다.
2. TimesBlock은 여러 주기에서 추출된 정보를 적응적으로 통합하여 풍부한 표현을 제공한다.

3.잔차 연결(Residual Connection):

1. TimesNet은 각 TimesBlock에 잔차 연결을 적용해 학습 안정성을 높이고 과적합을 방지한다.
2. 이 구조 덕분에 큰 데이터셋에서도 효율적으로 학습할 수 있다.

TimesNet의 성능 및 응용 분야

TimesNet은 단기 및 장기 예측, 분류, 이상 탐지, 데이터 결측치 보정 등 다양한 시계열 분석 작업에서 일관된 우수한 성능을 보이며, 특히 여러 시간 해상도를 동시에 분석할 수 있다는 점에서 다른 최신 모델보다 뛰어난 성능을 자랑한다

* TimesNet (2023)

TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis

TimesNet은 몇 가지 핵심 요소를 포함한다:

Multi-Scale Temporal Module (MST): 다양한 시간 스케일을 통해 데이터를 처리한다.
이 모듈은 시계열 데이터에서 발생할 수 있는 다양한 시간 주기를 효과적으로 인식하여 장기 및 단기 패턴을 모두 포착할 수 있도록 한다.

Fourier Transform Block: 주기적 패턴을 강조하기 위해 Fourier 변환을 활용합니다. 이 방식은 일반적인 시계열 데이터에서 발생하는 주기성을 잘 포착해 모델의 성능을 높인다.

Dynamic Filter Layer: 각 시간 단계에서 다른 필터를 적용할 수 있는 동적 필터 레이어가 있다. 이를 통해 시계열 데이터의 시점에 따라 변하는 특성에 더 잘 대응할 수 있다.

Multi-Head Attention Mechanism: Transformer의 다중 헤드 어텐션 구조를 차용하여 다양한 주파수 대역을 강조한다. 이를 통해 다양한 시간 패턴과 주파수를 동시에 반영할 수 있게 된다.

Residual Connection: 원본 시계열 신호와 모델의 출력을 잇는 잔차 연결(residual connection)을 포함하여, 정보 손실을 줄이고 모델의 안정성을 높인다.

* 환율 예측 모형

1. 시간 시계열 분석 (Time Series Analysis)

•**ARIMA 모델**: ARIMA(Autoregressive Integrated Moving Average) 모델은 과거의 환율 데이터를 기반으로 미래의 환율을 예측하는 데 사용된다. 이 모델은 데이터의 자기상관과 차분, 이동평균을 포함하여 데이터를 분석한다

(**예시)**: 미국 달러(USD)와 원화(KRW) 환율을 예측하기 위해 ARIMA 모델을 사용하면, 과거의 USD/KRW 데이터 패턴을 분석하여 향후 몇 개월의 환율을 예측할 수 있다.

2. 머신러닝 및 딥러닝 모델

1) LSTM (Long Short-Term Memory)
: LSTM은 순환 신경망(RNN)의 한 종류로, 시간에 따라 변화하는 데이터를 처리하는 데 뛰어난 성능을 보인다. 특히 장기 의존성을 학습할 수 있어 환율 예측에 효과적이다.
(예시) LSTM 모델을 사용하여 EUR/USD 환율을 예측하면, 과거 데이터로부터 학습한 패턴을 바탕으로 미래 환율을 예측할 수 있다.
→ 주로 **장기 예측에 활용**
2) Random Forest:
: 여러 개의 결정 트리를 사용하여 예측 성능을 향상시킨다. 이 모델은 다양한 입력 변수들을 고려하여 예측을 수행한다.
(예시) 주요 경제 지표(예: GDP, 금리, 인플레이션 등)와 과거 환율 데이터를 입력 변수로 사용하여 환율을 예측할 수 있다.

3. 통계 및 경제 모형

•**Vector Autoregression (VAR)**
: VAR 모델은 여러 시계열 데이터 간의 상호 관계를 분석하여 예측을 수행한다.
이 모델은 다양한 경제 변수와 환율의 상관관계를 고려하여 미래 환율을 예측한다.
(예시) 미국의 금리, 물가상승률, 무역수지 등의 데이터를 사용하여 EUR/USD 환율을 예측할 수 있다.

https://github.com/freejyb/KDT
예측분석/환율예측

이상 탐지(Anomaly Detection) 모델 : 금융시계열

- 이상 탐지(Anomaly Detection)는 데이터에서 정상 패턴과 다른 비정상적이거나 드문 패턴(이상치)을 식별하는 작업입니다. 이상 탐지는 다양한 응용 분야에서 중요한 역할을 하며, 예기치 않은 이벤트나 문제를 빠르게 탐지하고 대응할 수 있도록 돕는다.
- 금융 시계열 데이터에 대한 이상 탐지 과정은 주가, 거래량, 거래 패턴 등과 같은 금융 데이터에서 비정상적인 패턴을 식별하여 시장 이상, 사기 거래, 급격한 가격 변동 등을 조기에 감지하기 위한 것이다.

[일반적인 금융 시계열 이상 탐지 과정]

1. 데이터 수집 및 이해
2. 데이터 전처리
결측치 처리,스케일링,노이즈 제거, 차분 변환

3. 이상 탐지 방법 선택
1) 통계적 접근 방법
이동 평균과 표준 편차 기반 탐지: 이동 평균이나 이동 표준 편차를 기준으로 정상 범위를 설정하고, 특정 범위를 벗어나는 값들을 이상치로 간주한다.
분산 및 IQR: 데이터의 분산 또는 사분위 범위(IQR)를 활용해 데이터가 정상 범위에서 벗어난 경우를 이상으로 판단한다.
2) 기계 학습 기반 방법
- 지도 학습: 정상과 이상 데이터를 라벨링할 수 있는 경우, SVM, 결정 트리, 랜덤 포레스트 등 지도 학습 알고리즘을 활용하여 새로운 데이터의 이상 여부를 예측한다.
- 비지도 학습: 이상 데이터에 대한 라벨링이 부족할 경우, 클러스터링 기법(K-means, DBSCAN)이나 주성분 분석(PCA)을 사용해 정상 범위에서 벗어나는 클러스터를 이상으로 간주한다.

3) 딥러닝 기반 방법
- LSTM: 시계열 데이터의 순차적 패턴을 학습하여 예측하고, 예측 오차가 특정 임계값을 넘는 경우를 이상으로 판단합니다. 금융 데이터의 시간적 상관성을 학습하는 데 유리하다.
- 오토인코더(Autoencoder): 정상 패턴을 학습하여 입력을 재구성한 후, 재구성 오류(reconstruction error)를 기반으로 이상을 탐지한다. → **Denoising Autoencoder**
- 변이형 오토인코더(VAE): VAE를 사용해 정상 데이터를 학습하고, 새로운 데이터가 학습된 분포에 벗어나는 경우 이상으로 간주한다.
- GAN 기반 모델: GAN을 사용해 정상 데이터를 생성하고, 판별자를 통해 이상 데이터를 구분합니다. 이상치와 정상 패턴을 동시에 구분할 수 있어 복잡한 금융 데이터에 적합하다.

이상 탐지(Anomaly Detection) 모델 : 금융시계열

[일반적인 금융 시계열 이상 탐지 과정]

4. 모델 학습 및 검증

훈련 데이터와 테스트 데이터 분리 모델 학습 모델 검증

5. 임계값 설정 및 이상탐지

임계값 설정

실시간 이상 탐지: 금융 데이터는 실시간으로 업데이트되므로, 실시간 이상 탐지 시스템을 통해 이상 거래나 급격한 가격 변동을 감지하여 경고 시스템을 운영할 수 있다.

6. 결과 분석 및 조정

이상 탐지 결과 분석 ,모델 튜닝

7. 피드백 루프 및 시스템 개선

피드백 루프: 이상 탐지 시스템의 예측 결과를 계속해서 피드백 받아, 실제 이상 데이터와 비교하고 모델을 업데이트한다.

지속적 학습: 금융 시장은 변화가 빈번하므로, 새로운 데이터를 주기적으로 추가하여 모델이 최신 패턴을 반영할 수 있도록 지속적으로 학습시킨다.

별도 한글 자료 참고 !!

예측모델 평가 / 성능 개선

예측력 평가 지표

- **Mean Squared Error**

: 실제값과 예측값의 차이의 제곱을 합하여 예측기간수로 나눈 값

- **Root Mean Square Error(RMSE)**

· 실제 값과 모델의 예측 값의 차이

- **MAE : Mean Absolute Error**

: 실제값과 예측값의 절대값 차이의 합을 예측기간 수로 나눔

- **MAPE : Mean Absolute Percent Error**

: 실제값에 대한 MAE 비율

예측모델 평가 / 성능 개선

예측모델에 대한 지속적인 성능 추적 필요

→ 예측오차가 지속적으로 감소 또는 증가하는지 여부 확인

→ **추적지표 (TS: Tracking Signal)** : 예측치의 평균이 일정한 진로를 유지하고 있는지를 나타내는 척도

: 예측오차들의 합(누적예측오차)를 평균절대편차 (MAD : Mean Absolute Deviation)로 나눈 것

: **평균절대편차** (MAD): 각 측정치에서 전체 평균 값을 뺀 값의 절댓값으로 표시되는 편차들의 합에서 산술평균

- 추적지표의 값이「0」에서 크게 이탈한다면 Bias 발생하고 있음을 의미. 0 근처에 있는 것이 정상
- 일반적으로 **-4~4 사이에 있는 경우 정상으로 판단**

- 예측오차 = 실제 값 - 예측값
- MAD = $\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

$$\text{Tracking Signal} = \frac{\text{예측오차의 합}}{MAD}$$

성능 개선 검토

- **TS 상한(4)이나 하한(-4)을 벗어나는 경우 성능 개선 필요**
- **새로운 데이터로 재학습 → 수정된 분석모델 도출**
- **다른 분석 기법 적용 시도**

예측모델 성능 개선 방안

LSTM Hyper-parameter 선택

구분	내용
learning rate	▪ 그라디언트의 방향으로 얼마나 이동하면서 학습할 것인지를 결정 학습진도율이 작으면 학습의 속도가 느리고, 크면 학습이 되지 않고 진동할 수 있다. 보통 0.0001에서 0.1 사이의 값으로 지정한다.
Optimization 방법	아담(ADAM) (Kingma와 Ba, 2014) : 최적화 방법론들 중 일반적으로 좋은 추정 값을 찾아준다고 알려져 있음
Drop out 비율	▪ **Drop-out (Hinton 등, 2012)** 노드간의 상관관계를 줄이기 위해 역전파(back propagation) 알고리즘으로 계산할 때 의도적으로 일부 노드를 사용하지 않고 학습하는 방법
Epoch	반복적인 학습 알고리즘을 사용할 때, 모든 학습 데이터를 한 번씩 사용하는 것을 의미 → 예를 들어, 1,000개의 학습데이터가 있고 미니배치 크기를 50개로 정하였을 때 알고리즘을 20번(= 1,000개/50개) 반복하면 1번 반복(1 epoch)하였다고 한다. → 몇 가지 조합을 통해 가장 낮은 MSE를 갖는 반복횟수와 히든유닛 개수를 사용하였으며 방법

[참고] 증권시장 예측이 가능한가?

Paul Samuelson

- 도박판, 경마장, 증권사객장에서 부자가 되기는 힘들다. 주가예측가능성 효시

Louis Bachelier(the theory of speculation '13)

- 증권가격 예측 불가능

Charles Dow

- 주가는 예측 가능 , 다우이론, 다우지수(1884)

J.B Williams

- The theory of investment value : discounted cash flow

. Keynes , Gerald Loeb

- 어느 종목에 확신이 있다면 분산투자는 바람직 하지 않다.

H. Markorwitz : Portfolio selection ('52) Mean-Variance Analysis('59)

- '리스크가 모든 투자에서 핵심이다' 계란을 한 바구니에 담지 말라

James Tobins

- Liquidity Preference as Behavior Toward risk(58), Separation Theorem

[참고] 증권시장 예측이 가능한가?

W. Sharpe

- Simplified Model for Portfolio Analysis('61)

Eugene Fama

- EMH 평균수익률 이상 올릴 수 있는가 (전문가 확률 50%)

Modigliani/ Miller

- '58 자본구조와 기업가치

Sharpe/Lintner/Mossin/Treynor

- CAPM('64) 리스크가 기업평가를 좌우

Fischer Black, Myron Scholes, Robert Merton

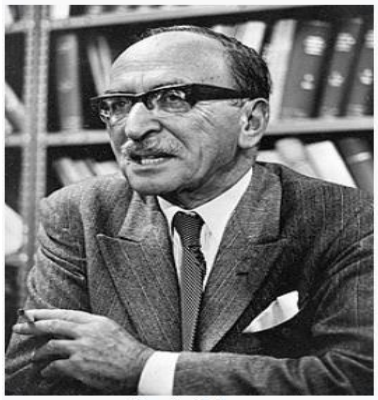
- "The Pricing of Options and Corporate Liabilities"
- "Theory of Rational Option Pricing" 파생상품

Barr Rosenberg

- 수익률과 리스크 측정모델

Mark Rubinstein Tobins

- Portfolio insurance


Dennis Gabor

Dennis Gabor(1900~1979)

"The future can't be predicted, but future can be invented."