빅데이터 분석을 위한 파이썬

이 현 창
디지털혁신실 디지털신기술반

2022.7.20-22

Outline

1. 데이터 분석
2. 빅데이터
3. 파이썬과 주피터노트북
4. 머신러닝
5. 분석 예제

데이터 분석

데이터 분석과 내러티브

행동경제학의 주요 발견(유의할 점):

- 사람들은 무엇에서건 '항상' 이야기(인과관계)를 찾아낸다.
- 데이터에서 '그럴듯한' 이야기를 발견한 연구자는 (그 결과가 우연히 얻어진 것일지라도) 그 이야기에 놀라울 정도로 강한 확신을 갖게 된다.
- 사람들은 메시지가 '퍼센트나 비율'과 같은 추상적인 단위보다 '만명당 x명'과 같은 구체적인 단위로 주어질 때 쉽게 받아들인다.

단계별 데이터 분석

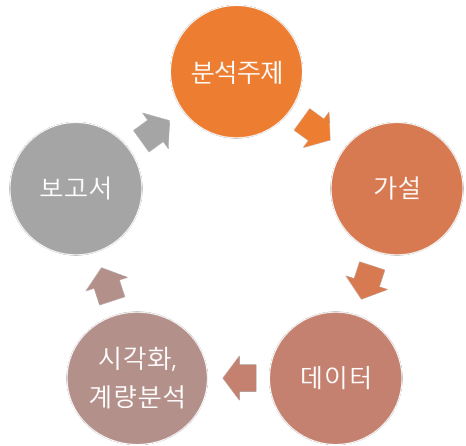
- 분석주제(질문)
- 가설(변수 간 관계에 대한 이론·모형)
- 데이터 준비(입수, 전처리, 정제, 가공)
- 분석(기술통계, 상관/회귀계수, 전망, 머신러닝)
- 시각화(선, 막대기, 히트맵/네트워크)
- 보고서(내러티브, 기승전결)

애자일agile 프로세스

'An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.' John Tukey

'It is better to be roughly right than precisely wrong.' John Maynard Keynes

데이터 분석 프로세스를 애자일하게 반복 수행하며 정확한 질문(분석주제)과 모형(가설), 데이터, 분석방법을 탐색



애자일 프로세스 - bottleneck

- 애자일 프로세스의 가장 큰 걸림돌(병목 지점)은 데이터 준비 과정

  (국민계정)

- 분석주제·가설, 분석방법론 등은 주로 본격적인 프로세스 시작 전 완료
  (부서별 수요(정책·조사연구, 모니터링 등), 데이터, 기존 연구 및 학술지식 등에 의해 결정)
- 입수·전처리·가공 및 분석·시각화를 위한 두 가지 방법: Excel v. Pipeline

Excel

- 엑셀에서의 데이터 작업은 손쉽고 직관적
  1. 누구나 엑셀 프로그램을 열고 숫자를 입력하거나 읽을 수 있음
  2. 현재 작업중인 데이터를 눈으로 확인
- 데이터가 커지고 분석 프로세스가 길어지면 처리 시간이 급격하게 증가
  1. 아무리 모니터가 커도 데이터가 한 눈에 안들어옴
  2. 여러 시트에 산재된 데이터가 어떻게 생성되었는지, 오류는 없는지 확인하기 어려움

Pipeline

- 데이터 분석 파이프라인(pipeline)은

  1. 데이터 속성에 맞도록 데이터를 구조화하고
  2. 독립적 기능을 수행하는 하위 프로세스들로 구성되어

  데이터 준비, 분석·시각화를 효율적으로 수행

- 정확한 연구주제(질문)에 대한 강건한(robust) 분석결과(답)를 얻기 위해서는 다양한 데이터 설정하에서 반복하여 파이프라인을 실행

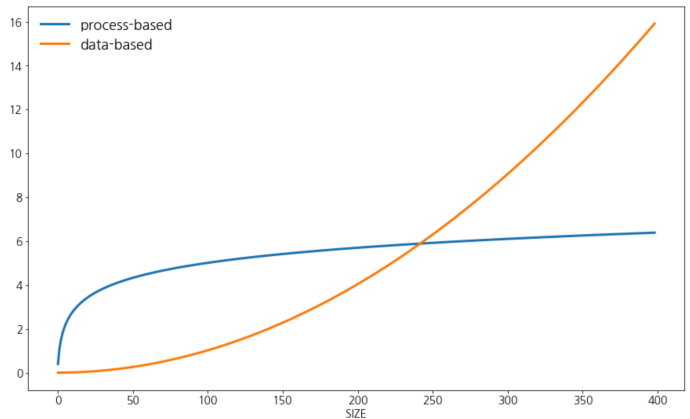
  데이터 설정에 따라 일부 프로세스 변경이 필요한 경우, 해당 프로세스만 수정후 전체 파이프라인을 일괄 수행

Pipeline - robustness check

- 단 하나의 변수 및 표본 구성, 계량모형에서 얻은 분석 결과는 동 표본에서만 유효할 가능성이 매우 높음
  - 변수 예측을 최우선 목적으로 하는 머신러닝에서는 표본내 예측력은 높지만 표본외 예측력이 낮은 과적합(overfitting) 방지가 최우선 과제
  - 일반적인 계량분석에서는 분석결과가 이론에 부합하고, 추정계수가 유의하면 강건성을 문제삼는 경우가 적음
- (어쩌다 얻어 걸린) 분석결과를 과도하게 신뢰하여 오랜기간 정성들여 보고서를 작성하고 나면 이후 되돌이킬 수 없으므로, 보고서 작성 전에 다양한 변수 조합과 변환, 표본기간에 대해 유사한 분석 결과가 얻어지는지 확인

Pipeline v. excel

엑셀기반 데이터 분석에 소요되는 시간은 데이터 크기에 따라 급격하게 증가하지만,
파이프라인기반 데이터 분석의 경우 데이터 크기의 영향을 거의 받지 않음



통합데이터플랫폼

- 디지털혁신실은 2024년 완료를 목표로 통합데이터플랫폼 개발을 진행 중
- 통합데이터플랫폼은

  1. 데이터 포털 BReiT
  2. 데이터 카탈로그 (데이터 조회, 메타데이터 및 권한 관리)
  3. 데이터 레이크 (빅데이터 저장, 전처리)
  4. 데이터랩 (고성능 연산, MLOps)

  등으로 구성

데이터 분석 라이브러리

- BIS는 직원들의 데이터 분석 수요에 대응하여 medau 라는 데이터 분석 라이브러리를 구축하였으며, fred, world bank 등의 경우 open api를 이용하여 데이터 조회, 입수를 자동화하는 라이브러리가 구축되어 있음
- 디지털신기술반은 데이터 조회부터 시각화까지 당행 데이터 분석에 자주 사용되는 기능을 모아 데이터 분석 라이브러리를 구축할 계획

1. BReiT Model Hub
2. 자생적 연구모임
3. 외부 전문업체 용역

데이터

- Data formats and Table
- Aggregate and granular/micro data
- Data sources
- Preprocess, cleansing, process
- Data analysis methods
- Data tools/programming languages

데이터 포맷

- xlsx, csv, dta, mat, txt, html, json, xml, parquet, pickle, ...

  download from webpages, web crawling, api, libraries, ...

- Structured: time series, cross-sectional, panel

  $\Rightarrow$ Table 형태로 주어짐

- Unstructured: text, audio/video, images

  $\Rightarrow$ Table 로 변환하여 분석

Table

ID	Time	GDP	Credit	Country	⋯
1	2021-03-31	0.362	5.62	US	⋯
2	30 Jun 2021	0.581	9.10	US	⋯
3	2021-03-31	0.183e	-	South Korea	⋯
4	2021-06-30		2.96	KR	⋯
5	2021Q1	0.476	4.11	CN	⋯
⋮	⋮	⋮	⋮	⋮	

Aggregate v. granular/micro data

To infer the relation between income and consumption from aggregate and granular data

t	C	I
2010	1.32	0.81
2011	2.51	0.93
2012	1.13	1.02
⋮	⋮	⋮

$\Downarrow$

$E[C|I] = \beta_0 + \beta_1 I$

V.

t	i	C	I
2010	1	1.45	0.85
2011	1	2.82	0.94
2012	1	1.45	1.03
⋮	⋮	⋮	⋮
2010	2	0.87	0.79
2011	2	2.31	0.90
2012	2	1.03	1.01
⋮	⋮	⋮	⋮

$\Rightarrow$

$E[C\|I^*]$	I^*
1.45	(0.7, 0.8]
1.51	(0.8, 0.9]
1.55	(0.9, 1.0]
⋮	⋮

Aggregate v. granular/micro data

- Aggregate data
  - 한국은행의 주요 분석대상인 거시경제변수를 직접 분석
  - 분석결과 해석이 연구자의 명시적, 암묵적 가정에 크게 의존
- Granular/micro data
  - 분석결과 해석에 연구자의 명시적, 암묵적 가정의 영향이 작음
  - 분석결과를 거시경제적 시사점으로 연결하기 위한 분석 필요

Sources

- ECOS, MDSS, BReiT, FAIRS, FISS, 가계부채DB, 외환전산망
- KOSIS, FISIS, R-ONE, KIS-value, NICE, KCB, CRETOP, KRX, 신용정보원
- IMF, BIS, FRED, OECD, UN, World Bank, WTO, WIOD, 각국 중앙은행, 미 상무부
- 노동연구원, 재정연구원, 무역협회, 고용노동부, 한국생산성본부, 서울시
- Bloomberg, Refinitiv, Infomax, haver, CEIC, Markit, Eikon
- Google, 온라인 쇼핑몰, 인터넷 뉴스, 위성사진, 한국전력, 한국도로공사, 카드사

Import, preprocess, cleansing

- 데이터 입수
- 날짜포맷, 데이터타입, 개체명 등 표준화
- 통계량, 결측치, 이상치 제거
- 데이터 선택, 결합
- 변수 생성, 통계량 계산, 시계열 빈도 변환
- 데이터 구조 설정
- 시각화, 노이즈 제거, 계절조정

Analysis/visualization

- Descriptive statistics
- (panel) regression, VAR, local projection, state-space
- DSGE, HA(NK), ABM, OLG
- Machine learning, NLP
- Bar, line, pie, mixed charts
- Network, heatmap, word cloud

Analysis - regression

- Q: x와 y의 관계는?
- 여타 요인(Z)을 통제한 두 변수 간 상관계수(β)는 충분히 좋은 답

$$y = \alpha + \beta x + \gamma Z + \varepsilon$$

- 인과관계(causality)를 이야기하기 위해서는
  1. y, x, z (x, y와 관련있는 주요 변수)를 포함한 설득력있는 모형을 찾고
  2. 동 모형에 근거한 계량모형을 이용하여 회귀계수(β)를 추정
- 한은칼럼 '상관과 회귀' (한은소식 2022년 6월호)

Analysis - machine learning

<참 고>

계량경제모형과 머신러닝

- 계량경제모형은 변수의 확률분포에 대한 가정하에 변수 간 관계 추정에 초점
  - 변수 간 관계는 통상 선형회귀식으로 표현
  - 데이터를 모두 사용하여 회귀계수를 추정
  - (모형검증) 추정결과가 이론에 부합하는지 여부, R^2, AIC, BIC 등으로 유효성 판단
- 머신러닝 알고리즘은 확률분포에 대한 가정없이 변수 예측에 초점
  - 특정 모형구조(하이퍼파라미터) 하에서 변수 간 관계를 포착하는 패턴(파라미터)을 알고리즘적으로 탐색
  - 일부 데이터에 대해 패턴을 학습한 다음 나머지 데이터를 이용하여 표본외 예측력 평가
  - (모형검증) 표본외 예측력을 기준으로 모형의 유효성을 판단

Analysis - machine learning

- 변화에 대처하는 슬기로운 한국은행(한은소식 2021년 8월호)
  - 2016년 알파고 파동, '머신러닝은 블랙박스라서 안돼'
  - (표본외) 예측력만 좋다면 ... interpretability vs. explainability
- 계량경제모형을 이용한 데이터 분석의 어려움
  - 분석 결과에 대한 신뢰도가 보고받는 사람(평가자)의 해당분야 이해도에 크게 좌우됨
  - 연구자의 암묵적 가정과 평가자의 암묵적 이해가 어긋나면...

Data tools/programming languages

tool	data process	analysis	visualization	community	accessibility
Stata	◑	●	◑	◑	◑
Matlab	◑	●	●	◑	◑
Gauss	○	◑	○	○	◑
E-views	○	○	○	○	◑
SAS	◑	◑	○	○	○
C/Fortran	○	◑	○	◑	○
R	●	◑	●	◑	◑
Python	●	●	●	●	●
Julia	◑	●	◑	◑	●

Note: ● is better than ○.

빅데이터

빅데이터

- 5 Vs (velocity, volume, value, variety and veracity)
- BIS working paper ‘Big data and machine learning in central banking’
- 과학기술부 통합 데이터 지도
  - 16개 빅데이터 플랫폼(금융, 유통, 지역경제 등)
  - 감염병 대응 빅데이터 플랫폼 구축 중(한국은행 참여)
- BReiT - BigData Hub(예정)
- (가로 또는 세로가 길어서) 한 화면에 모두 담을 수 없는 데이터

주요 빅데이터 및 활용 사례

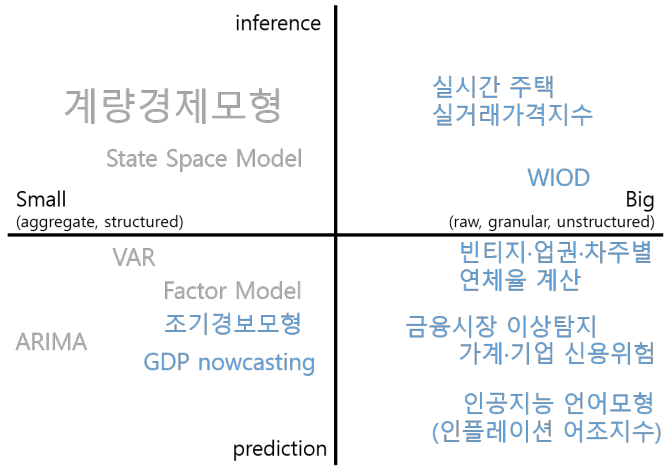
- 대규모 경제금융 시계열(GDP nowcasting, 조기경보모형)
- 거래 데이터(금융시장 이상탐지, 주택시장 실거래가격지수 등)
- 가계부채DB(신용위험 평가, 업권·차주·빈티지별 연체율)
- 신용카드, 전력사용량, 교통 등(경기 모니터링)
- 텍스트, 비디오/이미지, 오디오 등 비정형 데이터(전체 웹 데이터의 80-90%)
    - AI transcribe, 인공지능 언어모형(인플레이션 어조지수 등)

빅데이터와 데이터 분석

- 가로가 긴 빅데이터
  - 여러 변수의 공통요인을 추출하거나 설명(정보)력이 낮은 변수 제거
  - 계량경제모형(PCA, Factor Model), 머신러닝(벌점회귀모형)
- 세로가 긴 빅데이터
  - 거래 데이터, 대체 데이터(전기, 수도, 통신 등)
- 산업계에서 유용성이 검증된 빅데이터는 대부분 가로와 세로가 모두 긴 빅데이터
  - 넷플릭스 추천, 이미지 인식, 음성인식 비서 등
  - 가계부채DB

데이터 규모 및 분석 목적별 활용 사례

- 계량경제분석은 소규모 데이터를 이용하여 변수 간 (선형)관계 추론
- 머신러닝은 변수 예측, 변수 간 비선형, 상호의존성 포착
- 거의 모든 데이터 분석에 파이썬이 항상 유용



파이썬과 주피터노트북

디지털 신기술과 데이터 분석

- 다양한 형태의 빅데이터는 보다 효율적인 데이터 수집·전처리·가공 및 분석·시각화 역량을 요구
- 새로운 데이터 환경 및 분석 수요에 효과적으로 대응하기 위해서는

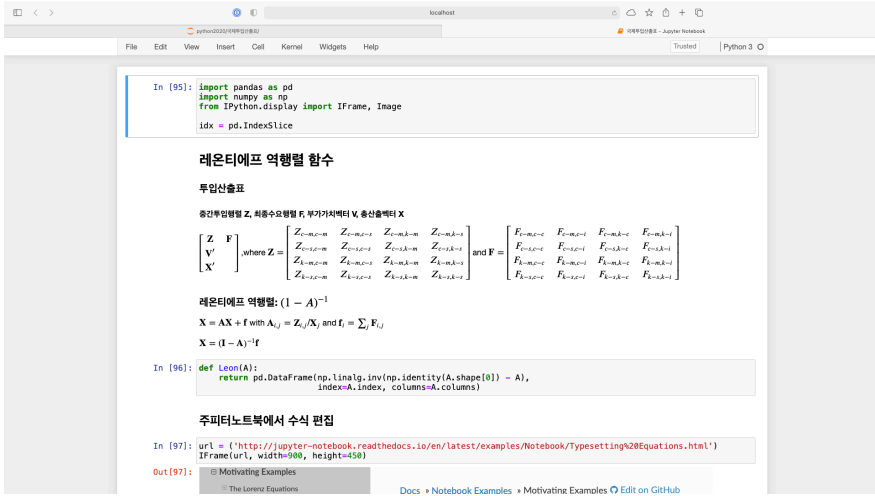
  1. 탐색적 데이터 분석과 협업이 용이한 환경에서
  2. 여러 데이터 도구를 효과적으로 결합하여 활용할 수 있는
  3. 사용자 인터페이스와 데이터 분석 언어를 선택

  $\Rightarrow$ 주피터노트북과 파이썬

주피터노트북

- 직관적이고 편리한 웹브라우저 기반 인터페이스
  - 일련의 입력과 출력, 텍스트와 코드 셀로 구성
  - 데이터 입수부터 가공, 분석, 시각화, 보고서/주석 전 과정 관리
- 다양한 분석환경에서 이용 가능
  - BReiT, PC, Google Colab
- 효율적인 파이프라인 구축 및 공유
  - 재현가능연구(reproducible research)
  - GDP nowcasting, 조기경보모형 등

주피터노트북 - 화면 예시


localhost
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
In [95]:
import pandas as pd
import numpy as np
from IPython.display import IFrame, Image
idx = pd.IndexSlice
레온티에프 역행렬 함수
투입산출표
중간투입행렬 Z, 최종수요행렬 F, 부가가치벡터 V, 총산출벡터 X
where
and
레온티에프 역행렬:
with
In [96]:
def Leon(A):
return pd.DataFrame(np.linalg.inv(np.identity(A.shape[0]) - A),
index=A.index, columns=A.columns)
주피터노트북에서 수식 편집
In [97]:
url = ('http://jupyter-notebook.readthedocs.io/en/latest/examples/Notebook/Typesetting%20Equations.html')
IFrame(url, width=900, height=450)
Out[97]:
Motivating Examples
The Lorenz Equations
Docs » Notebook Examples » Motivating Examples Edit on GitHub

Python supports

- 접근성(open source)
- 간결성(pythonic way of coding, philosophy)
- 객체지향(intuitive programming)
- 편리한 인터페이스(Jupyter notebook/lab)
- 방대한 사용자 커뮤니티(stackoverflow, github, ...)
- 다양한 라이브러리(collection, analysis, AI/ML, visualization)

Python can do

- Matlab 대체
- Stata 보완
- Agent-based model
- Machine learning
- 파이프라인 구축(MLOps)
- Contextual data analysis

Python - matlab, stata

- 매틀랩으로 작성된 코드는 대부분 파이썬으로 변환하여 보다 효율적으로 실행할 수 있음
- 파이썬으로 작성된 코드가 매틀랩에 비해 느리게 실행되는 경우 Cython을 이용하여 매틀랩 보다 빠르게 실행할 수 있음
    - kalman filter
- Dynare를 사용하고 있다면...
- 계량경제모형을 이용할 경우 stata와 파이썬을 보완적으로 사용

Python - ML and MLOps

- 대부분 머신러닝 라이브러리가 파이썬을 지원
  - scikit-learn
  - pytorch
  - tensorflow
  - huggingface
- 데이터 입수, 전처리, ML, 시각화 전 과정을 효율적으로 관리

Python - contextual data analysis

- Table 형태 데이터의 각 행과 열에 의미있는 이름(label)을 부여하여

  전 데이터 분석 프로세스(파이프라인)를 효율적으로 관리
  - 데이터 선택, 결합, 연산 프로세스를 오류없이 빠르게 수행
  - 가독성이 높아지므로 분석 결과 업데이트, 공유가 용이
- 행과 열 label은 날짜, 문자열, 숫자 등으로 지정
  - Python의 강력한 날짜, 문자열 처리 기능을 이용하여 손쉽게 라벨을 생성
  - 데이터분석 라이브러리 Pandas의 계층라벨은 보다 구조화된 분석을 지원

Python - contextual data analysis

		Country A			Country B		
		Sector 1	Sector 2	Final Consumption	Sector 1	Sector 2	Final consumption
Country A	Sector 1	0	20	5	0	30	10
	Sector 2	0	0	10	0	0	10
Country B	Sector 1	0	0	0	0	20	20
	Sector 2	0	0	20	0	0	30

실습

github.com/hyunchangyi/python101

- lecture_note.pdf
- intro.ipynb, pandas.ipynb, preprocess.ipynb

머신러닝

AI/ML 박기영·고정원(2019)

- AI : 인간의 학습, 추론 능력을 갖춘 컴퓨터 시스템 또는 관련된 과학기술
  - (strong AI) 사람처럼 행동하고 다양한 업무를 수행
  - (weak AI) 좁은 범위, 또는 단일한 업무를 처리하는 인공 지능
- ML : 인공 지능을 구현하기 위한 기술
  - “T(task)라는 작업을 수행하는 어떤 컴퓨터 프로그램의 성능 P(performance measure)가 E(experience)를 통해 향상된다면 이 프로그램은 E를 통해 학습한다” Mitchell(1997)
  - “명시적인 프로그래밍이나 지시 없이 데이터의 패턴을 인식하는 기법”

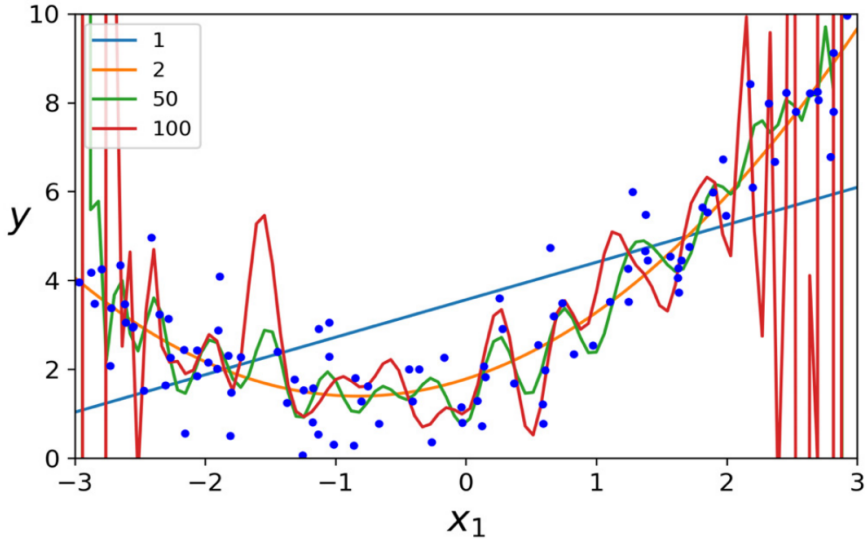
Learning

- 지도학습 supervised learning
  - 입력변수와 라벨label로 구성된 데이터를 통해 학습
  - 예측(분류, 회귀)
  - 인공지능 언어모형 fine-tuning
- 비지도학습 unsupervised learning
  - 인공지능 언어모형 pre-training
- 강화학습 reinforcement learning
  - 알파고

Hyperparameter

- 하이퍼파라미터는 머신러닝 알고리즘이 데이터의 패턴을 학습하는 방식, 파라미터는 머신러닝 알고리즘이 패턴을 인식한 결과를 의미
  - AR 모형의 하이퍼파라미터는 모형의 최대 시차, 파라미터는 AR 모형의 시차별 계수
- 하이퍼파라미터 최적화를 통해 머신러닝 알고리즘의 표본외 예측력을 극대화
  1. 전체 데이터를 학습 데이터와 테스트 데이터로 구분
  2. 학습 데이터에 대해 하이퍼파라미터를 바꿔가며 머신러닝 알고리즘을 학습
  3. 테스트 데이터를 이용하여 각 (하이퍼파라미터별) 알고리즘의 예측력을 계산하고 예측력이 가장 높은 하이퍼파라미터 선택

Overfitting 박기영·고정원(2019)



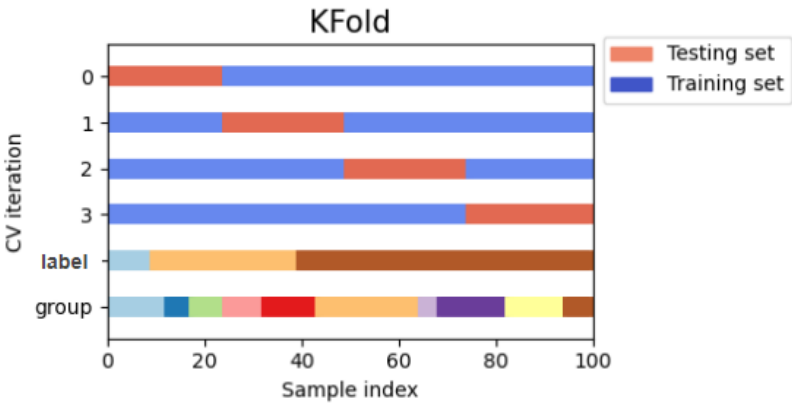
Model validation techniques

- 경제이론, R^2, AIC/BIC
- Holdout (train/test split) validation
- Leave-one-out cross-validation.
- K-fold cross-validation
- Walk-forward (time series split) validation

$\Rightarrow$ 데이터 속성(그룹)을 고려하여 과적합을 최소화하는 기법 적용
(인접한 관측치가 상관되어 있을 때 $\rightarrow$ 같은 그룹(학습/테스트)에 배정)

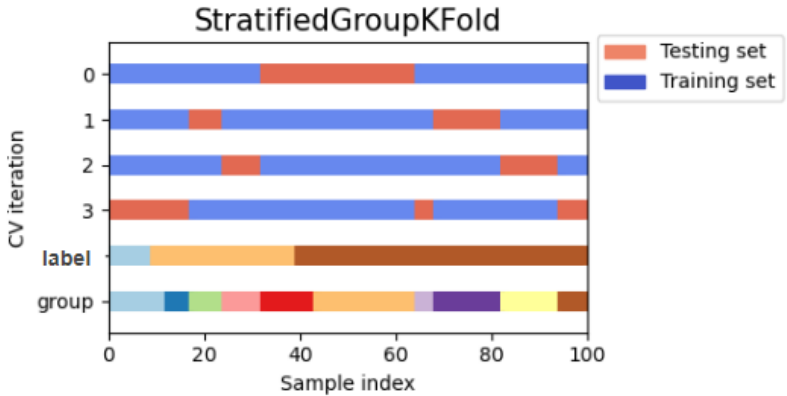
K-fold cross-validation

- 100개 관측치
- 3개 label
- 10개 group
- 관측 순서대로 fold(training/ test datasets) 구성



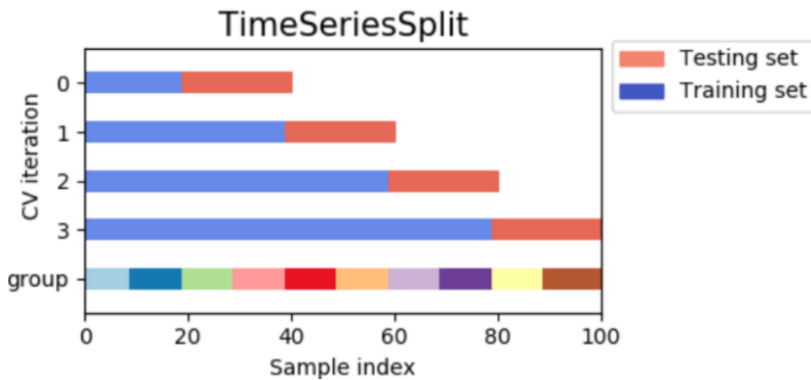
Stratified Group K-fold cross-validation

- 100개 관측치
- 3개 label
- 10개 group
- group단위로 각 label을 골고루
- 조기경보모형



Walk-forward validation

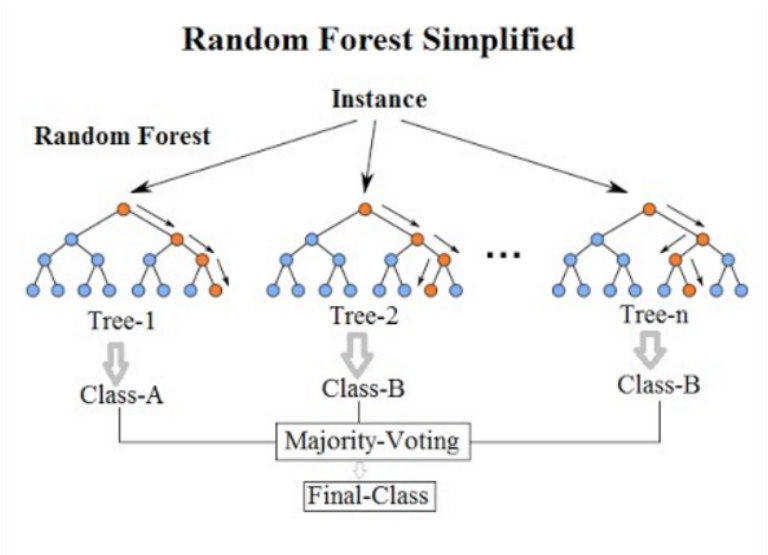
- 100개 관측치
- 10개 group
- 시간순서대로 training set구성
- GDP nowcasting



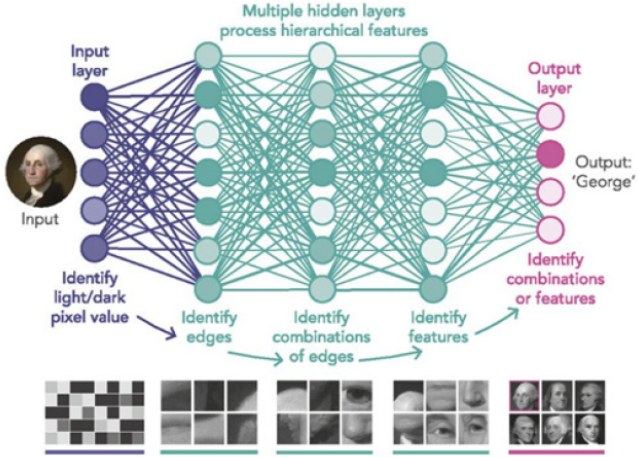
Unsupervised learning algorithms

- Penalized regression
  - Lasso, ridge regression, elastic-net
- SVM, Decision Tree
- Ensemble
  - 51% 정확도의 분류기 1,000개로 75%의 정확도 달성(colab)
  - Random Forest, Extremely Randomized Trees
- Deep learning
  - NN, RNN/LSTM, CNN

Random forest (colab)

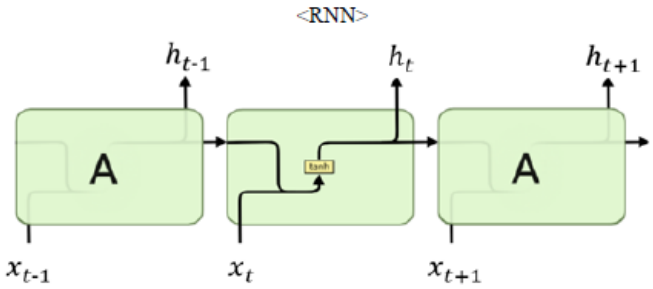


Deep learning (BOK이슈노트2022-7)



Note: This figure exemplifies a deep learning network for image recognition. The network consists of an input layer, three hidden layers, and an output layer. Input layer consists of nodes (the blue circles) storing the RGB values extracted from each pixel of the

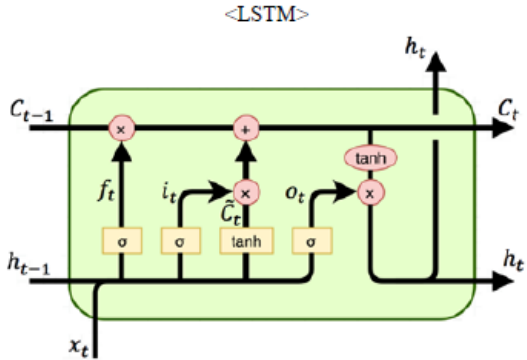
Deep learning - RNN/LSTM



Note: RNN cell(square in the diagram) takes input x_t and hidden layer h_{t-1}(containing past information) and generates h_t. Output at time t is calculated by a linear combination of nodes in h_t

Source: Amidi (2021)

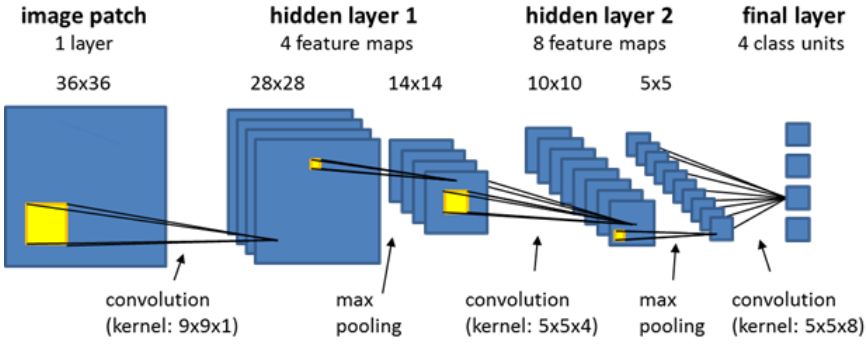
Deep learning - RNN/LSTM



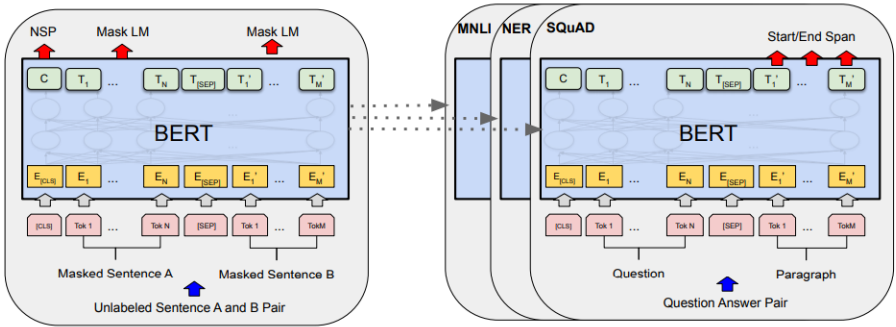
Note: LSTM cell (square in the diagram) takes input x_t and hidden layer h_{t-1} and generates hidden state h_t and cell state C_t at time t. Long-term information of LSTM cell at time t is passed on to the next LSTM cell by cell state C_t. Output at time t is calculated by a linear combination of h_t.

Source: Olah (2015)

Deep learning - CNN



인공지능 언어모형


NSP
Mask LM
Mask LM
BERT
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
BERT
Question
Paragraph
Question Answer Pair
Fine-Tuning

분석 예제

분석 예제

- WIOT
- 실시간 아파트 실거래가격지수
- GDP nowcasting
- 데이터 기반 조기경보모형
- 인공지능 언어모형

끝