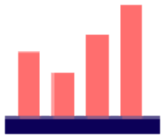
2019.10.11



데이터사이언스

- 열린 데이터세상과 미래의 플랫폼 -

이혜선

포항공과대학교 산업경영공학과

e-mail: hyelee@postech.ac.kr

web : http://www.postech.ac.kr/~hyelee

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

1. 데이터사이언스와 데이터

1-1 데이터사이언스 (데이터과학)

데이터 (Data)

데이터분석
(Data Analytics)

데이터 큐레이션
(data curation)
데이터추출, 변환
- SQL,R, Python

통계모형, 인공지능
(t-검정, 회귀분석, 머신러닝)

데이터 시각화
(data visualization)
그래픽 (R의 ggplot)

인사이트 창출
(Insight)

1-2 데이터화 (Datafication)

Activity, Unvisible ➡ Data화 되는것 (Jose van Dijck, 2014)

- 기계가 읽어들일수 있는 모든것 (숫자, 영상, 이미지, 텍스트) 데이터로 변환

이미지(영상)

➡ Data
(숫자, 벡터로 변환)

1-2 데이터화 (Datafication)

2016.12.03

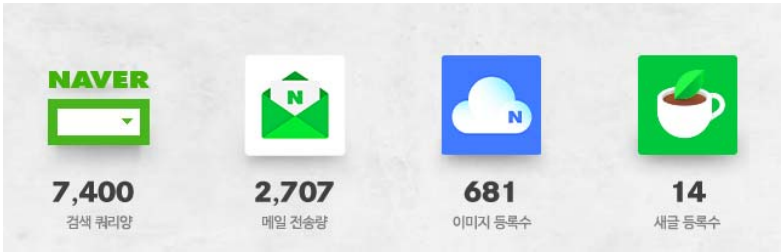
- How many people?

(i) count # passenger of subway gate (nearby 3 subway stations)

(ii) count # of people in image (using deep learning technique)

1-2 데이터화 (Datafication)

1초마다 새롭게 생성되는 온라인 데이터



1-2 데이터화 (Datafication)

• 데이터센터 '각' – Naver

" 당신은 남기고
우리는 보관합니다"

- 데이터사이언스
- 데이터화 (Datafication)
- 데이터감시 (Datavelliance= Data + surveillance)
- 빅데이터의 활용

2. 데이터애널리틱스와 통계

2-1 데이터애널리틱스

데이터 큐레이션(data curation)
데이터추출, 변환
- SQL,R, Python

데이터 시각화 (data visualization)
그래픽 (R의 ggplot, python의 matplotlib)

통계모형, 인공지능
(t-검정, 회귀분석, 머신러닝)

2-2 통계적 통찰력 : 분산(편차)의 의미

- 2018년 우리나라의 '행복지수' 평균은 157개국 중 57위로 비교적 높은 편
- 표준편차로 측정한 '행복 불평등도'는 157개국 중 96위로 행복의 격차가 매우 심각한 사회로 조사됨

청년AI.Big Data아카데미 온라인 교육과정 [포스텍 MOOC]

[포스텍 MOOC] 빅데이터분석과 R 프로그래밍 (이혜선 교수)

[포스텍 MOOC] 머신러닝기법과 R프로그래밍 (이혜선교수)

https://pabi.smartlearn.kr/#courses

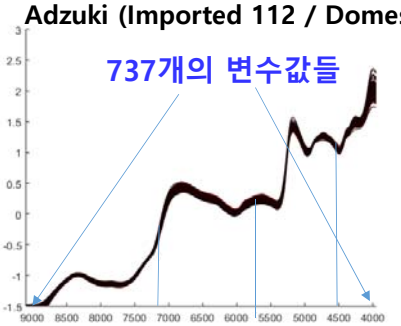
데이터사이언스를 위한 통계학입문 I
POSTECH

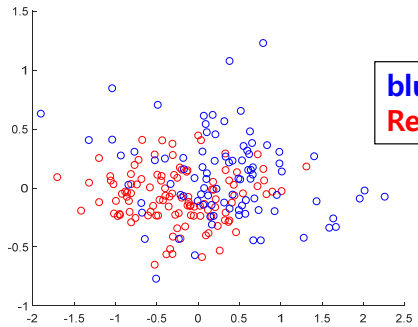
3. Auto-encoder 기반 분류분석

*세미나 축약본으로 세부과제 분석결과는 첨부되지 않았습니다.

3-1 스펙트럼데이터 – 원산지 판별

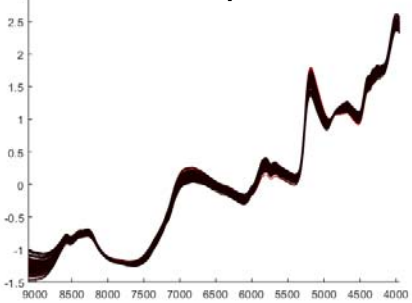
Adzuki (Imported 112 / Domestic 89) : 팥

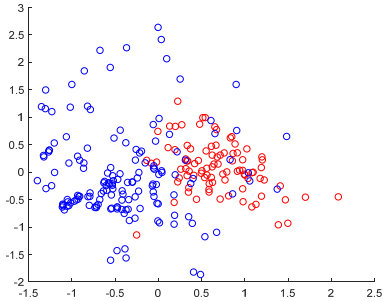




blue: 수입산
Red: 국내산

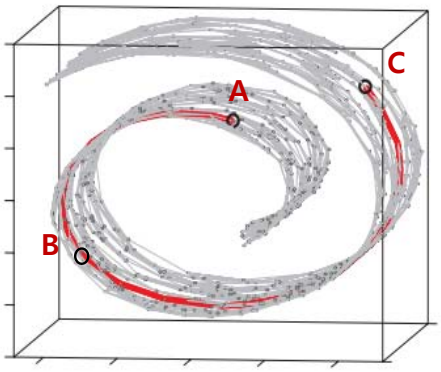
Perilla seeds (Imported 91 / Domestic 154) : 들깨



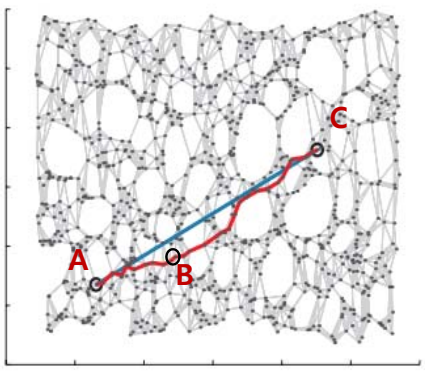


3-2 Feature learning : 매니폴드 러닝

▪ Linear method - Difficult finding non-linear manifold



3차원 공간에서의 거리
AB > AC

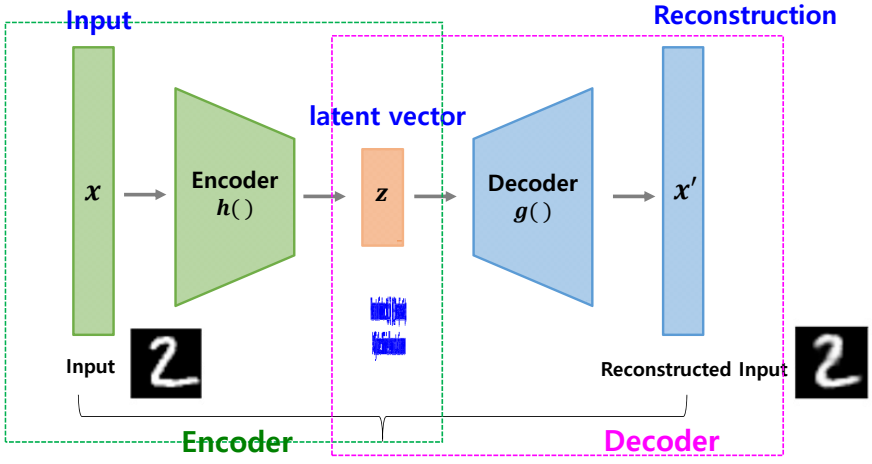


실제 매니폴드의 거리
AB < AC

3-3 Auto-encoder (딥러닝기법)

- compress the input layer(x) into latent variable(z) - Encoder

=> uncompress into something(x') closely with the original data(x) - Decoder

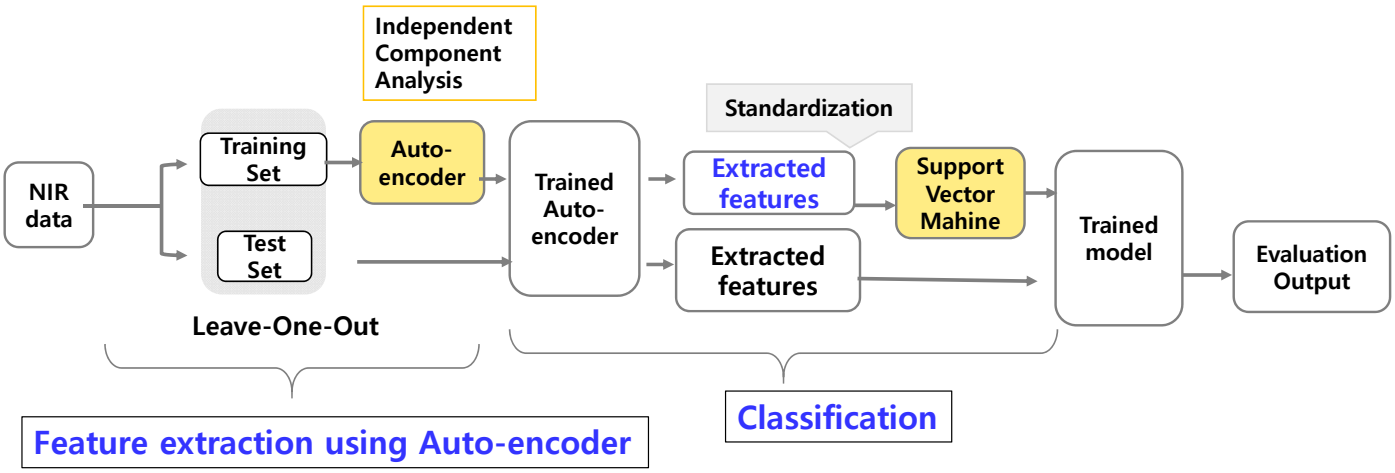


Reconstruction Error : $L(x, x')$ $(L(\cdot)$: $Loss\ function)$

→ Objective : Minimize the reconstruction error

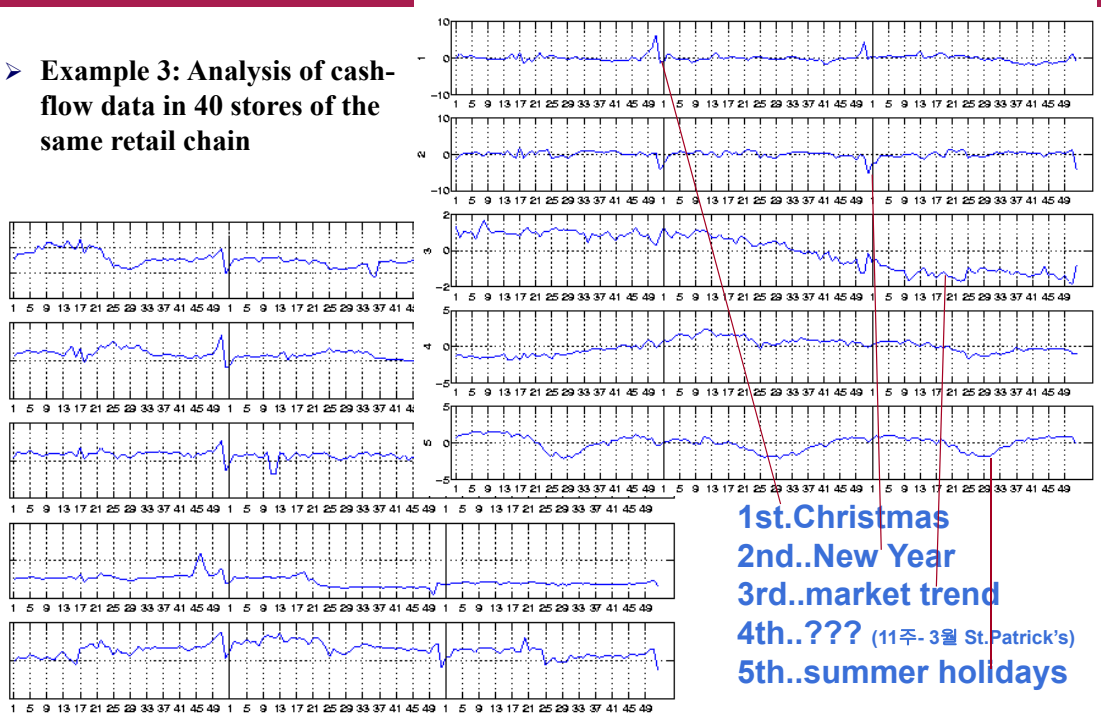
Reference : https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

3-3 Auto-encoder기반 분류 모형



3.4. Independent Component Analysis (ICA)

- **Example 3: Analysis of cash-flow data in 40 stores of the same retail chain**



1st.Christmas
2nd..New Year
3rd..market trend
4th..??? (11주- 3월 St.Patrick's)
5th..summer holidays

4. 빅데이터의 블랙홀

4-1 현업데이터의 현실

1. Masking problem – 실제는 이상치인데.. mask되는 경우
2. Swamping problem – 너무 많은 범위를 이상치로 진단하는 경우 실제는 아닌데 swamp되는 경우
3. Missing problem –결측치 문제

4-1 현업데이터의 현실

- A 화장품 브랜드의 고객맞춤형 프로모션

	내용	기타
	1. 통합고객마스터	
	2. 통합고객 구매정보	
	3. 매장 마스터	
	4. 통합고객_포인트내역	
	5. 웹 경로 고객정보 마스터	HADOOP 리스트 전달 시 재요청 예정
	6. VIP승급조회	
	8. 통합VIP 고객 등급 마스터	
	9. 뷰티포인트 종류	
	10. 통합고객 경로별 수신동의	
	11. 캠페인별집계	
	12. 고객별집계	
	13. 상담집계	
	14. 캠페인별 혜택별집계	
	15. 상품테이블	BRANDCD 추가예정

결측치 : 엇갈리게 발생하는 결측치 문제는 쉽지 않다!!

X_1	X_2	X_3	⋯	X_k
1.2	176		⋯	1.5
1.6			⋯	1.8
⋯	⋯	⋯	⋯	⋯
0.9	175		⋯	

Y_1	Y_2	Y_3
	176	90
⋯	⋯	⋯
0.9		93

4-1 현업데이터의 현실

Multi-task Learning : 서로 연관된 종속변수(task)들의 예측모형 학습

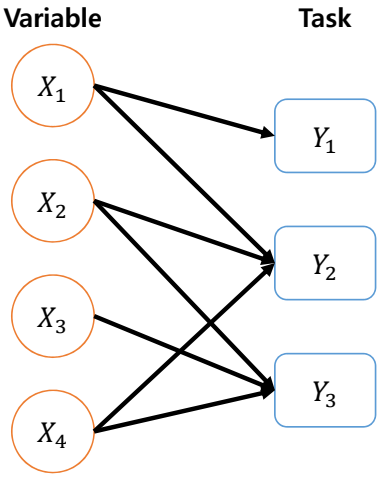
Task 1

X_1	X_2	X_3	X_4	Y_2
2	1	10	40	10
3	0	11	23	20
2	1	19	53	17

⋮

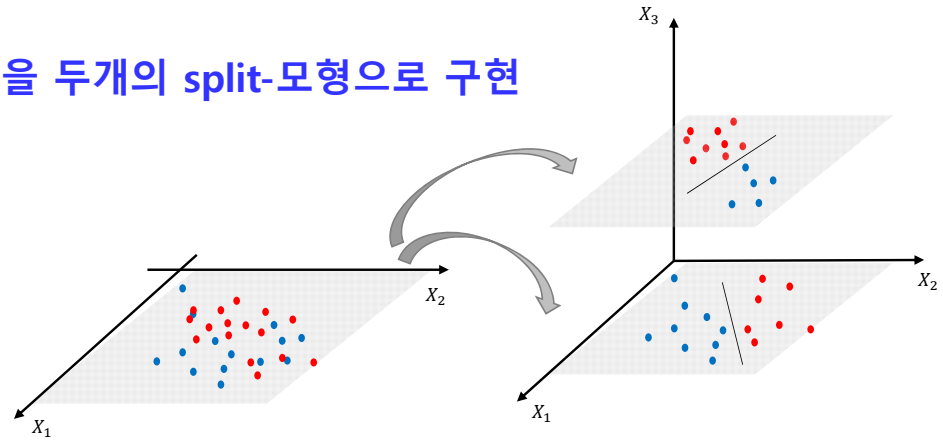
Task 3

X_1	X_2	X_3	X_4	Y_3
2	1	10	40	0.13
3	0	11	23	0.23
2	1	19	53	0.4



4-2 Model Tree (여러 개의 예측모형을 구현)

예측모형을 두개의 split-모형으로 구현



5. 미래의 플랫폼 - 열린 데이터 세상

5-1 열린 데이터 세상

• 핀란드의 의료데이터 프로젝트 (FinnGen)

- 핀란드인(Finnish)과 유전자(Genome)의 합성어.
- 자발적 참여자의 유전자정보를 수집하고 환자의 의료정보까지 통합구축.

- 2017년 50만명(국민의 10%) 유전자정보로 특정질병간 상관관계연구

- 6개월마다 데이터 업데이트 – 전세계 연구자와 공유
- 관절염/당뇨병 등 자가면역질환 연구 수행중 – 개인 맞춤형 약 개발 추진중

5-1 열린 데이터 세상

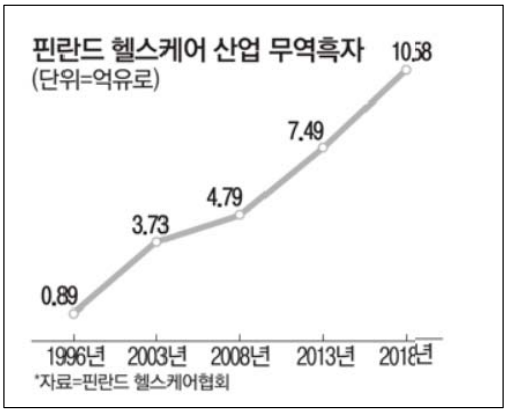
국민의 공공의료와 사회보장 정보를 자유롭게 활용할 수 있도록 규제를 없앰

핀란드 '국민의료·사회보장 데이터 2차 활용법' 공시 (2019.5)

➡ 의료정보 규제 푼 핀란드…글로벌 헬스케어 투자 몰려

⬇

GE헬스케어와 퍼킨엘머(Perkin Elmer), 바이에르, MSD 등 글로벌 헬스케어 기업들이 핀란드로 몰려들고 있다 !!



5-1 열린 데이터 세상

• 금융빅데이터 개방시스템 (2019.6) - 한국신용정보원

5000여개 금융기관의 4000만명의 정보를 통합운영 개방

ref) 금융위원회 배포자료

5-2 미래의 플랫폼-오픈소스

H2O

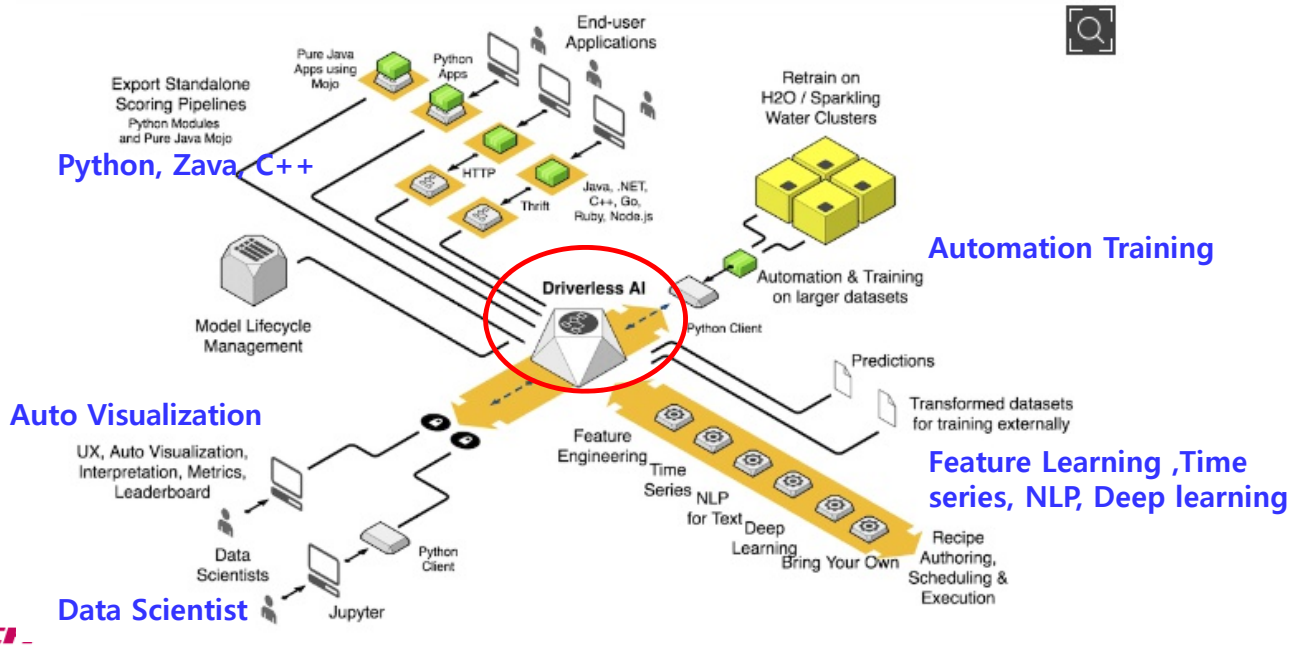
H_2O.ai

"AI to do AI"

AI : Artificial Intelligence

5-2 미래의 플랫폼 – Automation learning

• 자동화 머신러닝 (Automation Machine Learning)



Harmony

에너지-신재생 에너지/효율적 배분

데이터센터

환경 (전기, 미세먼지)

금융의 투명화

개인정보보호법에 의한 개인정보 제공 규제완화

의료산업발전

Datafication / Dataveillance