Introduction to Artificial Intelligence [AICS223]

AI, Data & Preprocessing (W03)

고려대학교

인공지능사이버보안학과

KOREA UNIVERSITY

Prof. Mee Lan Han (aeternus1203@gmail.com)

CONTENTS

1. Artificial intelligence (AI)
2. Learning method
   - ✓Supervised Learning
   - ✓Unsupervised Learning
3. Data
   - ✓Data Characteristic
   - ✓**Data Preprocessing**
   - ✓**Feature Engineering**

Review

■ Supervised Learning

□ 정답지 (Labeling)이 있는 데이터를 대상으로 학습하는 과정

□ 실제 데이터에서의 정답지 표현 (Class라고 기재된 레이블)

■ Unsupervised Learning

□ Labeling이 없는 데이터셋을 대상으로 비슷한 특징을 갖는 데이터셋이 군집화함으로써 새로운 결과를 추론하는 학습과정

□ 비지도학습은 정답을 맞히는 목적의 모델을 생성하지 않음

■ Data Characteristics

□ **정형 데이터**

- 관계형 데이터베이스 시스템의 테이블과 같이 고정된 컬럼에 저장되는 데이터와 파일
- 지정된 행과 열에 의해 데이터의 속성이 구별되는 스프레드시트 형태의 데이터

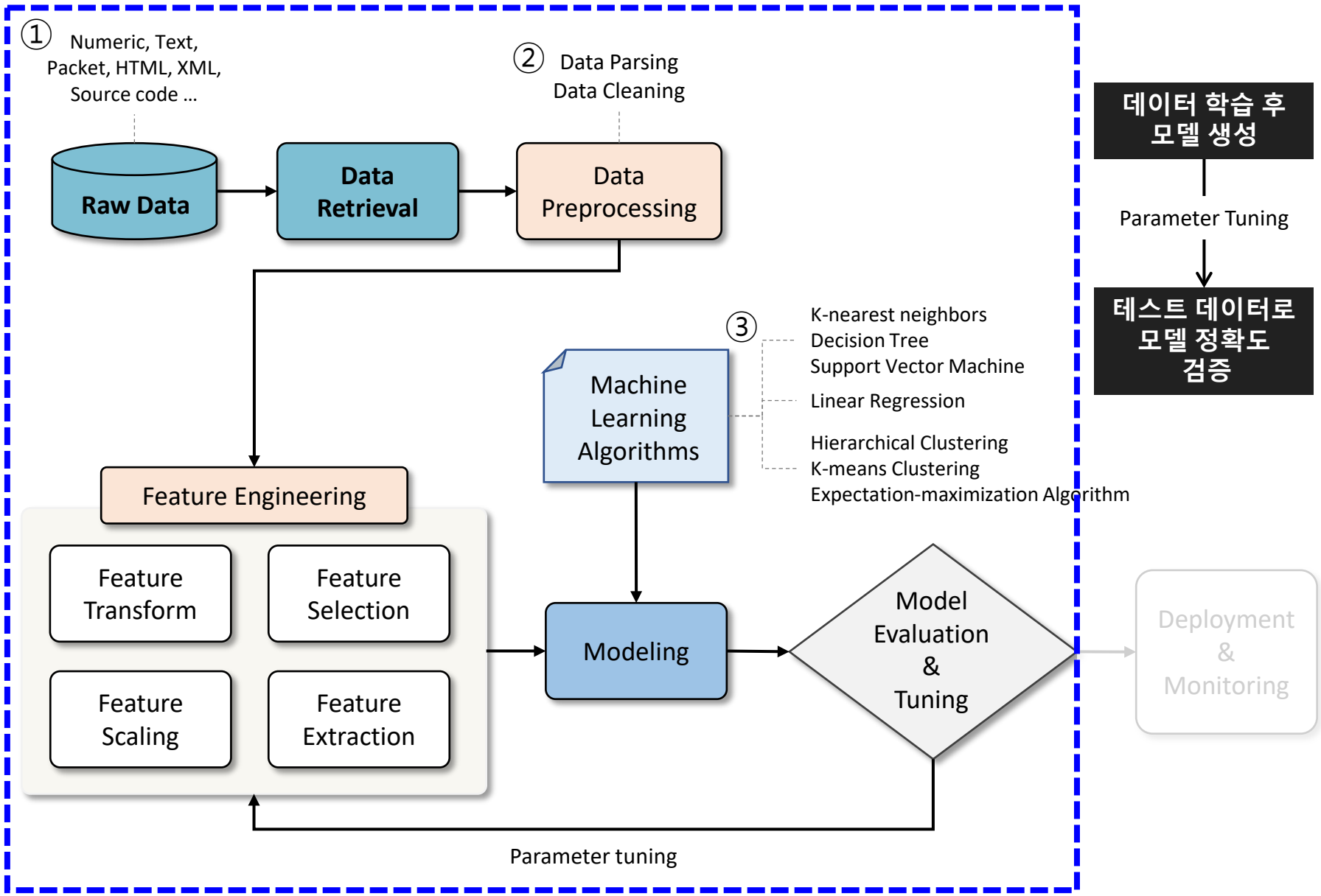
□ **반정형 데이터**

- 해당 파일 파싱 후 메타구조를 갖는 정형데이터 형태의 테이블 구조로 재생성
- 보통 API 형태로 제공되기 때문에 데이터 처리 기술이 요구
- HTML, XML, JSON 등

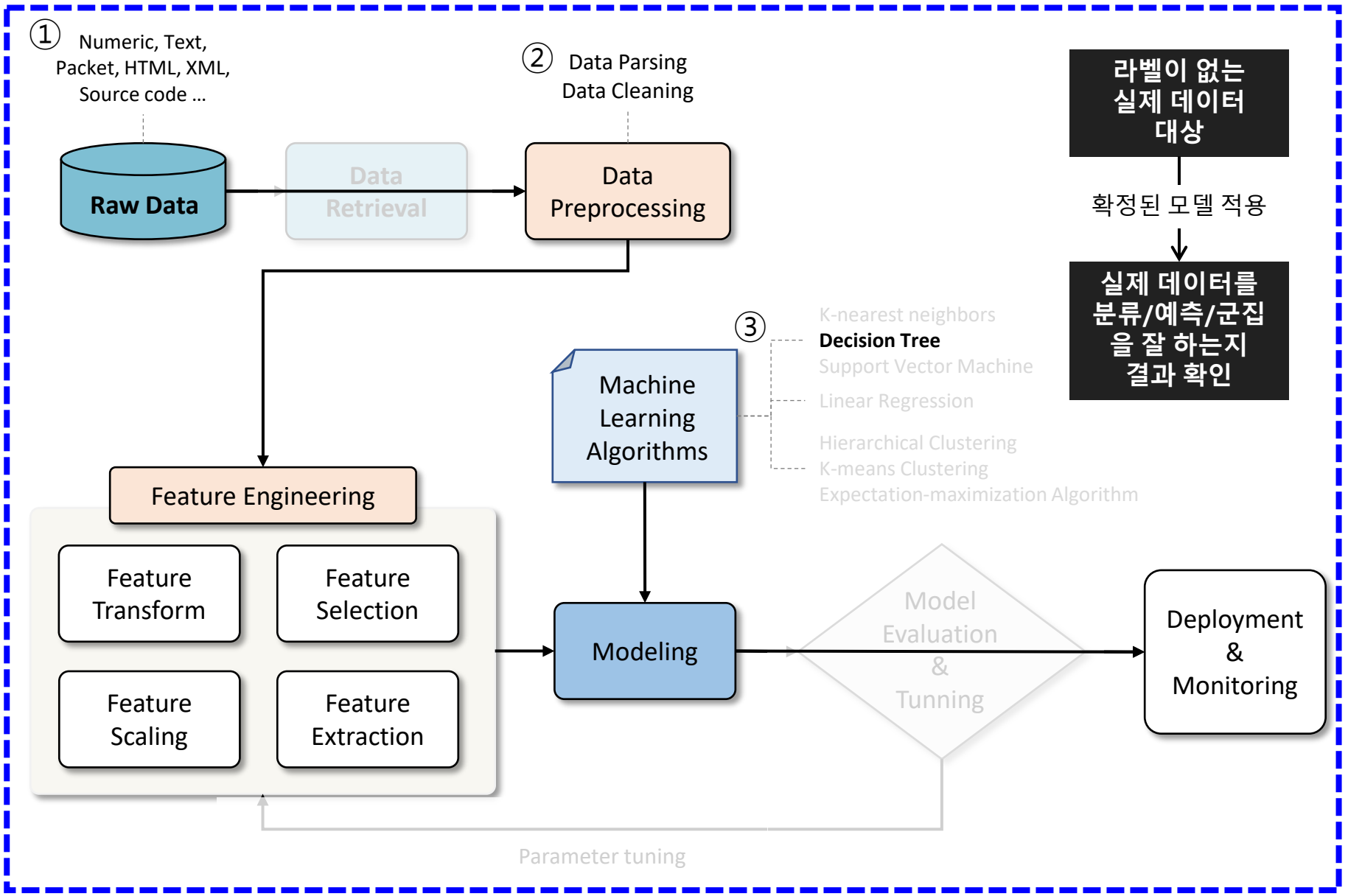
□ **비정형 데이터**

- 고정된 필드에 저장되어 있지 않은 데이터
- HEX (16진수), 이미지, 비디오 스트림, 오디오 데이터, 자연어
- 수집 난이도가 높음

Review - Machine Learning Pipeline



Review - Machine Learning Pipeline



Data Preprocessing

■ Data Preprocessing Methods

□ **Normalization** & **Standardization** & **One-Hot Encoding**

□ Missing Value Treatment

□ Outlier Treatment

■ Feature Engineering

□ Feature Treatment

□ Feature Engineering

□ Underfitting/Overfitting

Data Preprocessing

■ Normalization

□ **입력된 x 값들을 모두 0과 1사이의 값으로 변환하는 방법**

□ 칼럼(Feature)마다 해당 값의 범위(Scale, Range)가 크게 다른 경우의 전처리 방법

□ Min-Max Scaling

- 데이터의 최소값은 0, 최대값은 1로 변환

no	A	B
0	-3.075383	-0.448552
1	4.302209	-2.151012
2	-0.522519	-1.245304
3	1.394279	-1.169169
4	0.427537	-0.458080
5	13.250265	-0.734271
6	0.137697	-2.041710
7	2.614901	-3.057545
8	7.866735	-2.140529
9	0.297992	-0.503868

no	A	B
0	0.000000	1.000000
1	0.451902	0.347465
2	0.156371	0.694613
3	0.273782	0.723795
4	0.214565	0.996348
5	1.000000	0.890487
6	0.196812	0.389359
7	0.348549	0.000000
8	0.670241	0.351483
9	0.206630	0.978798

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$0.206630 = \frac{0.297992 - (-3.075383)}{13.250265 - (-3.075383)}$$

머신러닝에서 scale이 큰 feature의 영향이 비대해지는 것을 방지
딥러닝에서 Local Minima에 빠질 위험 감소 (학습 속도 향상)

Data Preprocessing

■ Standardization

- □ 데이터 단위가 불일치 할 때 대상 데이터를 같은 기준으로 볼 수 있도록 함
- □ 데이터가 정규분포를 따른다는 가정하에 실시
- □ **입력된 x 들의 정규 분포를 평균이 0 이고 분산이 1 인 표준 정규 분포로 변환하는 방법**

	no	기간	국가명	수출건수	수출금액	수입건수	수입금액	무역수지	기타사항
0	0	2015년01월	중국	116932	12083947	334522	8143271	3940676	NaN
1	1	2015년01월	미국	65888	5561545	509564	3625062	1936484	NaN
2	2	2015년01월	일본	54017	2251307	82480	3827247	-1575940	NaN
3	3	2015년02월	중국	86228	9927642	209100	6980874	2946768	NaN
4	4	2015년02월	미국	60225	5021264	428678	2998216	2023048	NaN
5	5	2015년02월	일본	48652	2000724	83320	3837614	-1836890	NaN
6	6	2015년03월	중국	117529	11868032	234321	7226911	4641121	NaN
7	7	2015년03월	미국	75789	6795064	491428	4055574	2739490	NaN
8	8	2015년03월	일본	60018	2127216	98373	4383839	-2256623	NaN
9	9	2015년04월	중국	118916	11765637	274021	7648402	4117234	NaN

$$z = \frac{x - \mu\,(\text{평균})}{\sigma\,(\text{표준편차})}$$

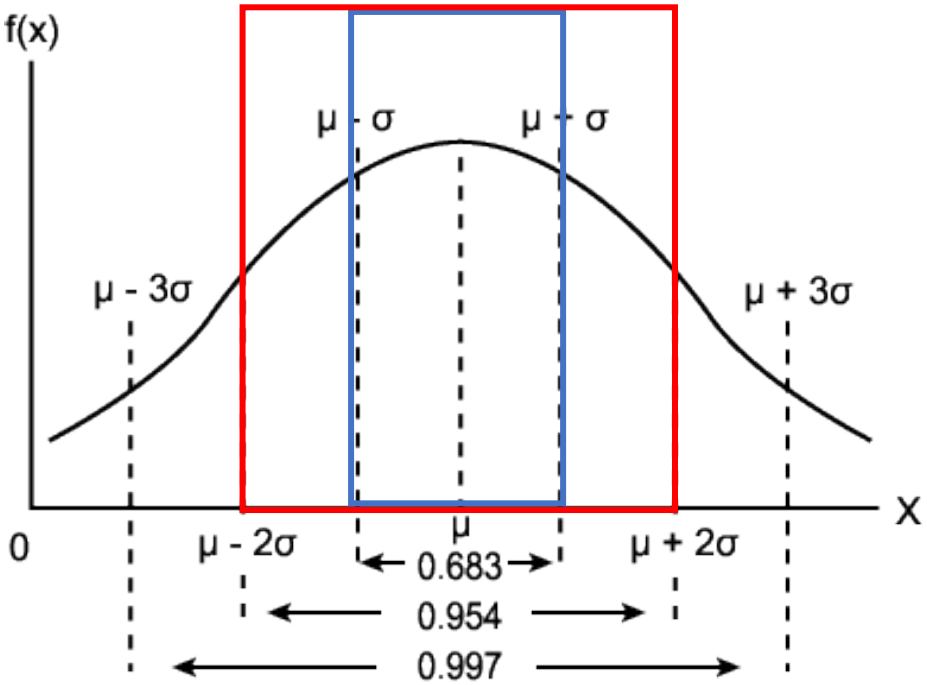
	수출건수	수출금액	수입건수	수입금액	무역수지
0	1.310266	1.227229	0.367049	1.457119	0.871746
1	-0.521464	-0.328940	1.438912	-0.836352	0.103141
2	-0.947458	-1.118725	-1.176322	-0.733721	-1.243869
3	0.208443	0.712760	-0.400969	0.867079	0.490584
4	-0.724682	-0.457845	0.943610	-1.154543	0.136338
5	-1.139983	-1.178511	-1.171179	-0.728459	-1.343943
6	1.331689	1.175715	-0.246529	0.991969	1.140366
7	-0.166163	-0.034637	1.327857	-0.617821	0.411093
8	-0.732110	-1.148331	-1.079002	-0.451192	-1.504911
9	1.381462	1.151284	-0.003427	1.205921	0.939456

머신러닝에서 scale이 큰 feature의 영향이 비대해지는 것을 방지
딥러닝에서 Local Minima에 빠질 위험 감소 (학습 속도 향상)

Data Preprocessing

■ Standardization

- □ 데이터 단위가 불일치 할 때 대상 데이터를 같은 기준으로 볼 수 있도록 함
- □ 데이터가 정규분포를 따른다는 가정하에 실시
- □ **입력된 x 들의 정규 분포를 평균이 0 이고 분산이 1 인 표준 정규 분포로 변환하는 방법**



머신러닝에서 scale이 큰 feature의 영향이 비대해지는 것을 방지
딥러닝에서 Local Minima에 빠질 위험 감소 (학습 속도 향상)

Data Preprocessing

Standardization

- 데이터 단위가 불일치 할 때 대상 데이터를 같은 기준으로 볼 수 있도록 함
- 데이터가 정규분포를 따른다는 가정하에 실시
- 입력된 x 들의 정규 분포를 평균이 0 이고 분산이 1 인 표준 정규 분포로 변환하는 방법

Normalization과 Standardization 수행 이유

- 데이터의 분포를 조정하여 모델 성능을 향상시키기 위해 사용
- 데이터의 스케일 차이로 인한 문제 해결, 모델 학습을 더 효율적으로 수행

머신러닝에서 scale이 큰 feature의 영향이 비대해지는 것을 방지
딥러닝에서 Local Minima에 빠질 위험 감소 (학습 속도 향상)

Data Preprocessing

■ One-Hot Encoding

- □ 머신러닝이나 딥러닝 프레임워크에서 범주형을 지원하지 않는 경우 사용
- □ 카테고리별 이진 특성을 만들어 해당하는 특성만 1, 그렇지 않은 경우 0으로 만드는 방법

파이썬의 pandas에서 **get_dummies** 함수를 이용

no	기간	국가명	수출건수	수출금액	수입건수	수입금액	무역수지	기타사항
0	2015년01월	중국	116932	12083947	334522	8143271	3940676	NaN
1	2015년01월	미국	65888	5561545	509564	3625062	1936484	NaN
2	2015년01월	일본	54017	2251307	82480	3827247	-1575940	NaN
3	2015년02월	중국	86228	9927642	209100	6980874	2946768	NaN
4	2015년02월	미국	60225	5021264	428678	2998216	2023048	NaN
5	2015년02월	일본	48652	2000724	83320	3837614	-1836890	NaN
6	2015년03월	중국	117529	11868032	234321	7226911	4641121	NaN
7	2015년03월	미국	75789	6795064	491428	4055574	2739490	NaN
8	2015년03월	일본	60018	2127216	98373	4383839	-2256623	NaN
9	2015년04월	중국	118916	11765637	274021	7648402	4117234	NaN

	미국	일본	중국
0	0	0	1
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	1	0
6	0	0	1
7	1	0	0
8	0	1	0
9	0	0	1

Data Preprocessing

■ One-Hot Encoding

□ 머신러닝이나 딥러닝 프레임워크에서 범주형을 지원하지 않는 경우 사용

□ 카테고리별 이진 특성을 만들어 해당하는 특성만 1, 그렇지 않은 경우 0으로 만드는 방법

no	기간	국가명	수출건수	수출금액	수입건수	수입금액	무역수지	미국	일본	중국
0	2015년 01월	중국	0.142372	0.794728	0.197014	0.700903	0.70832	0	0	1
1	2015년 01월	미국	0.035939	0.295728	0.332972	0.085394	0.496512	1	0	0
2	2015년 01월	일본	0.011187	0.042477	0.001249	0.112938	0.12531	0	1	0
3	2015년 02월	중국	0.078351	0.629759	0.099597	0.542551	0.603281	0	0	1
4	2015년 02월	미국	0.024131	0.254394	0.270146	0	0.50566	1	0	0
5	2015년 02월	일본	0	0.023306	0.001901	0.11435	0.097732	0	1	0
6	2015년 03월	중국	0.143617	0.77821	0.119186	0.576069	0.782345	0	0	1
7	2015년 03월	미국	0.056584	0.390099	0.318885	0.144042	0.581375	1	0	0
8	2015년 03월	일본	0.0237	0.032983	0.013593	0.188761	0.053373	0	1	0
9	2015년 04월	중국	0.146509	0.770376	0.150022	0.633488	0.726979	0	0	1

Data Preprocessing

■ Missing Value Treatment

□ 결측값이 있는 상태로 모델을 생성할 경우, 변수간의 관계가 왜곡
-> 모델의 정확성이 떨어짐

□ **NA or N/A (Not Available)** 라고 표시 (or **False, Unknown**)

User	Device	OS	Transactions
A	Mobile	Android	5
B	Mobile	Android	3
C	NA	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4
F	NA	Android	2
G	Tablet	Android	4

Data Preprocessing

■ Missing Value Treatment

1) 단순 삭제

- 결측값이 발생한 모든 관측치를 삭제
- 모델 학습에 사용되는 변수(Feature) 중 결측값이 발생한 관측치만 삭제

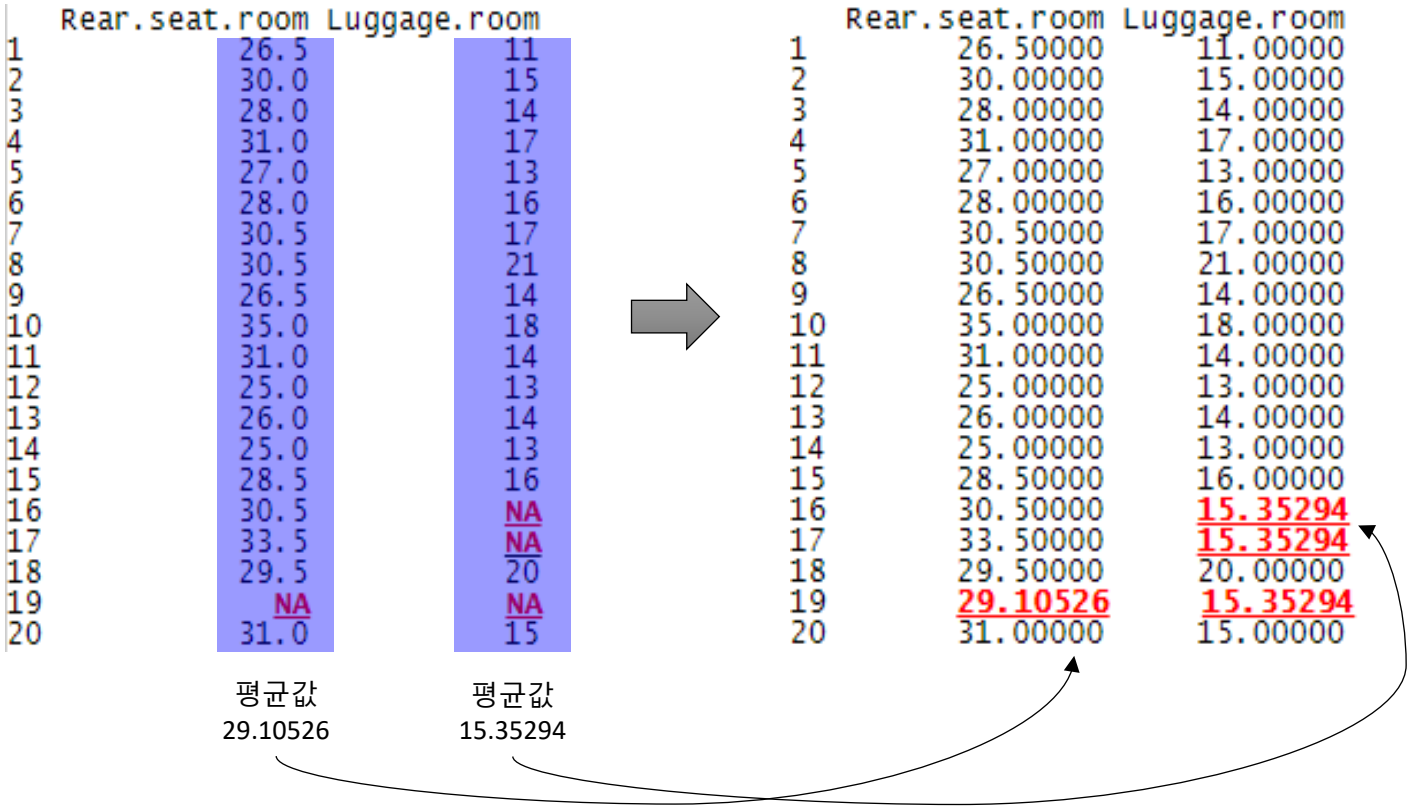
Date	Notify	URL	gTLD	ccTLD	OS	Encoding	IP	WebServer	Lang
1998-01-02	Team CodeZero	www.janet-jackson.com	com			utf-8			
1998-01-04	Optiklenz(LOU)	marin.k12.ca.us		us		iso-8859-1			Helvetica
1998-01-04	Team CodeZero	www.bolling.af.mil	mil			windows-1252			Times New Roman
1998-01-04	foxy man	hope.hinet.net	net			utf-8			
1998-01-04	Team CodeZero	www.dm.af.mil	mil			utf-8			
1998-01-05	net_	pr0n.prizon.net	net			utf-8			
1998-01-05	Team CodeZero	www.nic.ad		ad		utf-8			
1998-01-22	The No|d Crew	www.legislate.com	com			iso-8859-1			
1998-01-25	Giftgas	www.kleiber.com	com			utf-8			
1998-01-25	ToxicEdge	www.chem.umu.se		se		utf-8			
1998-01-26	Optiklenz	www.hartnet.org	org			utf-8			Helvetica
1998-01-28	Giftgas	www.datapark.net	net			utf-8			
1998-01-29	SpiritWalker	www.connectos.com	com			utf-8			
1998-01-29	alker-SisterMoon-	www.resopal.nl		nl		utf-8			
1998-02-20	SUiD-USA/Inet	www.stat-usa.gov	gov			utf-8			
1998-04-14	RaPtoR 666	www.koreamls.com	com			utf-8			
1998-04-19	str0" Fouk0" Luna	www.leonardodicaprio.com	com			utf-8			
1998-05-24	milw0rm	www.fantasyfootball.co.uk	com	uk		utf-8			
1998-05-24	milw0rm	www.michaelpowles.co.uk	com	uk		utf-8			
1998-05-25	milw0rm	www.leje.com	com			utf-8			Helvetica
1998-05-26	milw0rm	www.aps-plc.com	com			utf-8			
1998-05-26	milw0rm	www.dencom.army.mil	mil			utf-8			
1998-05-26	milw0rm	www.intelliprise.com	com			utf-8			
1998-05-27	milw0rm	www.dyervolvo.com	com			utf-8			
1998-05-27	milw0rm	www.tugmanufacturing.com	com			utf-8			
1998-06-01	The Resistance	www.rtnielson.com	com			utf-8			
1998-06-02	milw0rm	www.barc.ernet.in		in		iso-8859-1			
1998-06-02	milw0rm	www.cartoon.co.uk	com	uk		utf-8			
1998-09-13	HFG	www.nyt.com	com			utf-8			
1998-09-15	H4G1S	www.slashdot.org	org			utf-8			
1998-09-17	Team X-Ploit	www.sanpedro.gob.mx	gov	mx	Windows	utf-8			
1998-10-15	Fluxx	www.jackdaniels.com	com			windows-1252			Times New Roman
1998-11-16	0rE & Dem0l	hptesla.mi.infn.it		it		utf-8			
1998-11-19	HcV	www.hack-net.com	com			utf-8			Helvetica
1998-11-19	HotMan	www.asiaaccess.net.th	net	th		utf-8			

Data Preprocessing

■ Missing Value Treatment

2) 다른 값으로 대체

- 각 변수들이 특정한 확률분포를 따른다고 가정하고 분포의 모수들을 추정하여 대체를 실시하는 방법
- 평균값, 최빈값, 중간값 등으로 대체



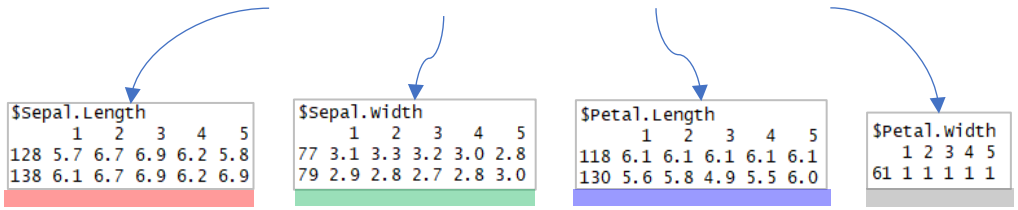
Data Preprocessing

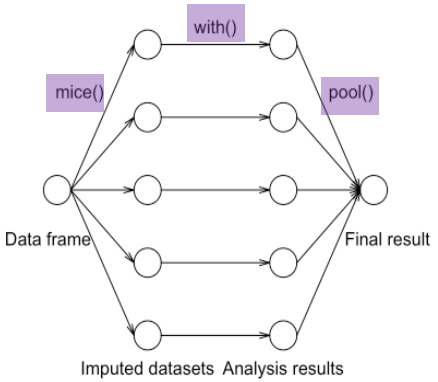
■ Missing Value Treatment

3) 예측값 삽입

- Data Frame 에 있는 다른 모든 값을 사용하여 (Training data로 사용) 누락된 값을 예측한 후, 예측된 값을 삽입하고 누락된 값을 처리함
- R의 mice() 패키지를 적용

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
44	5.0	3.5	1.6	0.6	<NA>
118	7.7	3.8	NA	2.2	virginica
61	5.0	2.0	3.5	NA	versicolor
130	7.2	3.0	NA	1.6	virginica
138	NA	3.1	5.5	1.8	virginica
7	4.6	3.4	1.4	0.3	<NA>
77	6.8	NA	4.8	1.4	versicolor
128	NA	3.0	4.9	1.8	virginica
79	6.0	NA	4.5	1.5	versicolor
65	5.6	2.9	3.6	1.3	<NA>





* dataset을 m개 생성 (디폴트는 5)

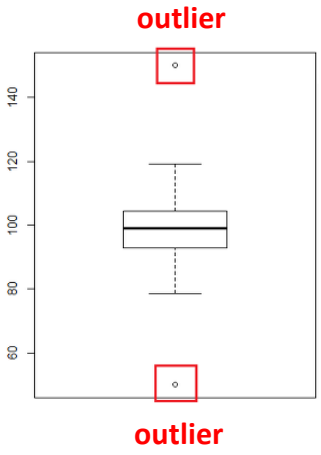
	1	2	3	4	5
7	setosa	setosa	setosa	setosa	setosa
44	setosa	setosa	setosa	setosa	setosa
65	versicolor	versicolor	versicolor	versicolor	versicolor

Species

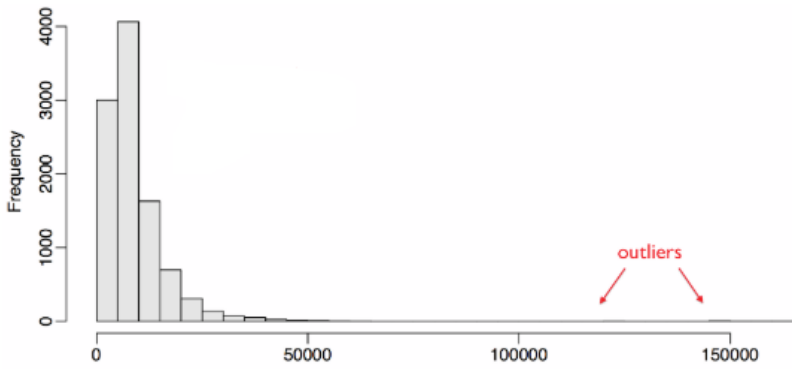
Data Preprocessing

■ Outlier Treatment

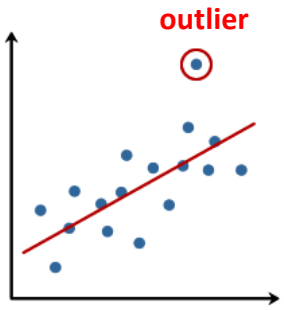
- □ 데이터와 동떨어진 (dissimilar) 관측치
- □ 모델의 결과를 왜곡할 가능성이 있는 관측치
- □ 변수의 분포 시각화를 통해 확인



Boxplot



Histogram



Scatter plot

Data Preprocessing

■ Outlier Treatment

1) 단순 삭제

- Human error에 의한 경우 해당 관측치를 삭제

2) 다른 값으로 대체

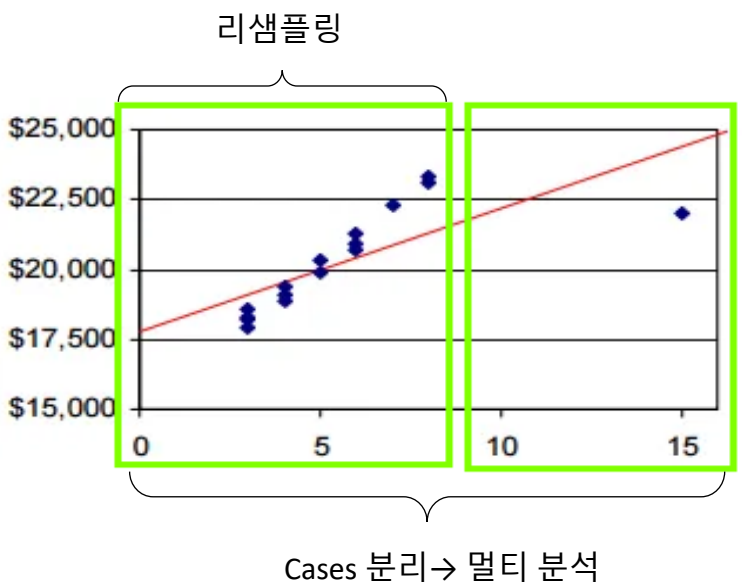
- 절대적인 관측치의 숫자가 작은 경우, 관측치를 삭제하는 대신 다른 값 (평균 등)으로 대체

3) 리샘플링

- 이상값을 분리하여 분석 범위를 수정하는 방식으로 모델을 만드는 방법

4) Case 분리하여 멀티 분석

- 이상값을 포함한 모델과 제외한 모델을 모두 만들고 각각의 모델로 분석



Data Preprocessing

■ Data Preprocessing Methods

- □ Normalization & Standardization & One-Hot Encoding
- □ Missing Value Treatment
- □ Outlier Treatment

■ Feature Engineering

- □ Feature Treatment
- □ Feature Engineering
- □ Underfitting/Overfitting

Feature Engineering

■ Feature Treatment

□ 기존의 변수를 사용해서 데이터에 정보를 추가하는 과정

□ **새로운 관측치나 변수를 추가하지 않고도 기존의 데이터를 보다 유용하게 만드는 방법**

① Scaling
② Binning (구간화)
③ Transform
④ Dummy

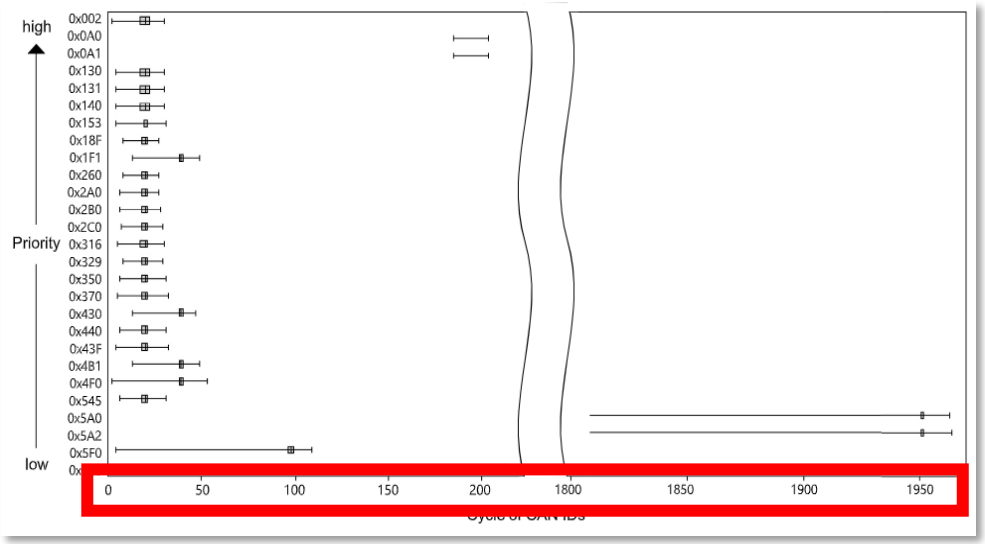
Feature Engineering

■ Feature Treatment

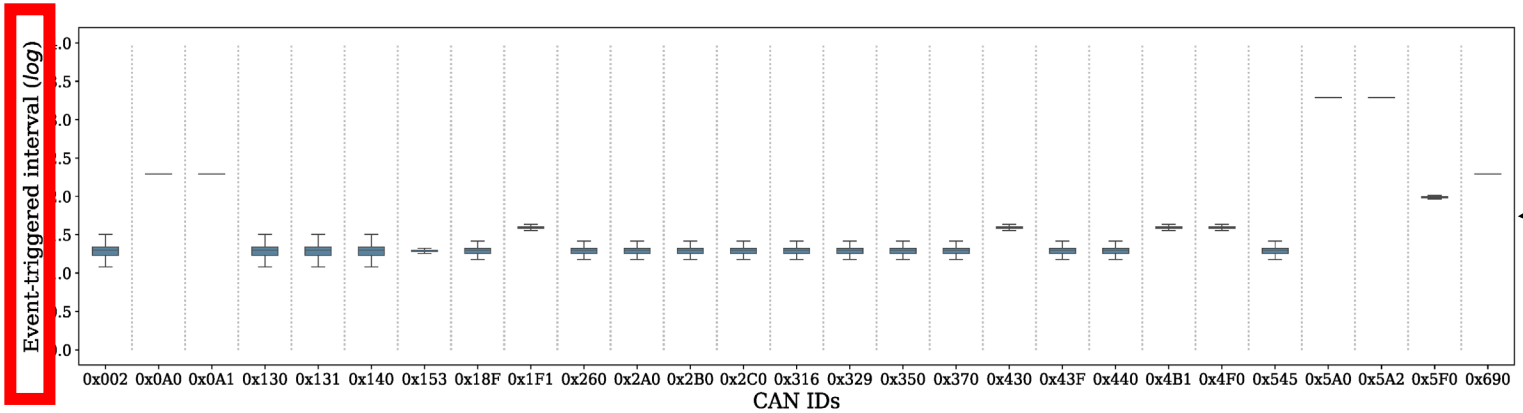
(1) Scaling

- 변수의 단위 변경을 원할 경우
- 변수의 분포가 편향되어 있는 경우
- 변수 간의 관계가 잘 드러나지 않는 경우

→ Normalization과 Standardization



* Log 함수, Square root를 취하는 방법



Feature Engineering

■ Feature Treatment

(2) **Binning** (구간화)

: 연속형 변수를 범주형 변수 또는 순위형 변수로 변환하는 방법

□ 변수 구간화를 사용하는 이유

(1) 이상치로 발생 가능한 문제를 완화

(2) 결측치 처리를 보다 간편하게 수행할 수 있음

(3) 과(소/대)적합을 완화 시켜주는 효과

Feature Engineering

■ Feature Treatment

□ 변수 구간화를 사용하는 이유

(1) 이상치로 발생 가능한 문제를 완화

Outlier

Height (키): **155, 167, 169, 170, 171, 178, 177, 172**, **300** => Mean value 184.3333333 **??**

평균값은 이상치에 상당히 민감 => 해당 데이터 셋의 대표값으로 부적절한 데이터

1구간 (x < 160) : 1명
2구간 (160 ≤ x < 170): 2명
3구간 (170 ≤ x < 180): 5명
4구간 (180 ≤ x) : 1명

구간화를 통한 대표값이 데이터인
학생들의 키를 더욱 잘 설명함

(2) 결측치 처리를 보다 간편하게 수행할 수 있음

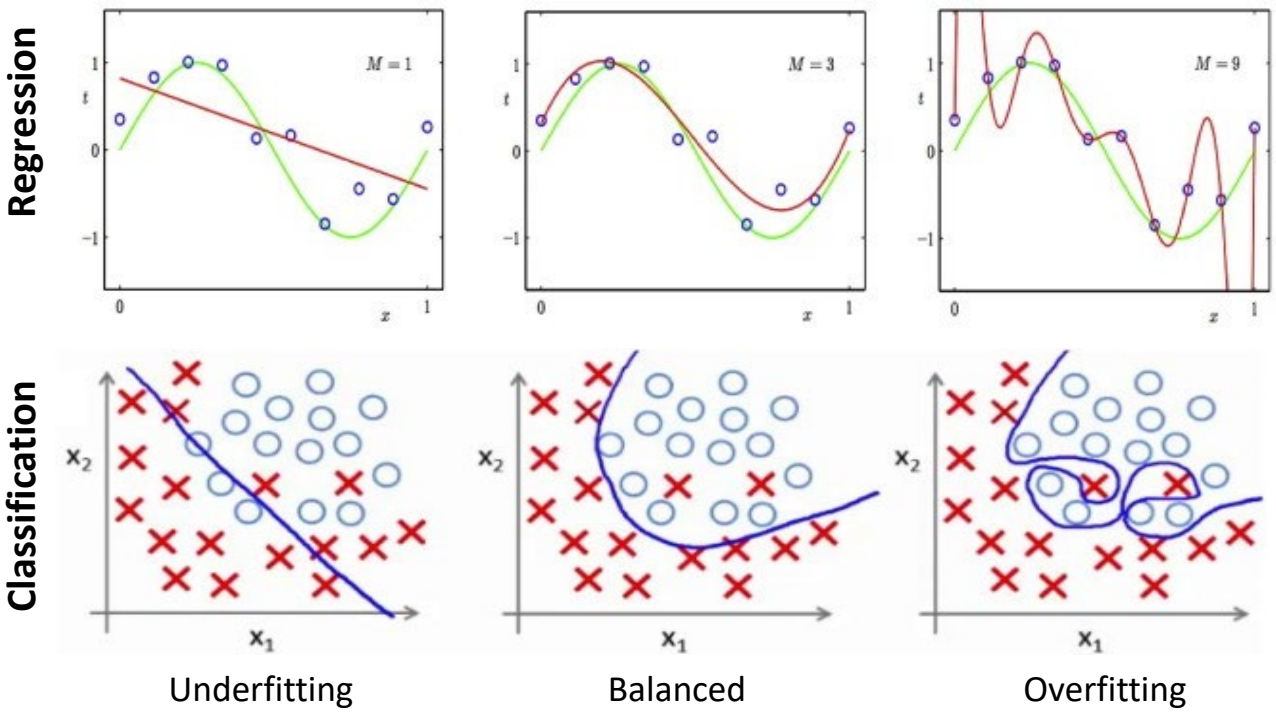
Feature Engineering

■ Feature Treatment

□ 변수 구간화를 사용하는 이유

(3) 과(소/대)적합을 완화 시켜주는 효과

- **과소적합 (Underfitting):** 모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못할 때 발생
- **과대적합 (Overfitting):** 모델이 훈련 데이터에 너무 잘 맞지만 테스트 데이터에는 잘 맞지 않음
  (일반성이 떨어진다는 의미 → 범용화하여 사용할 수 없는 모델)



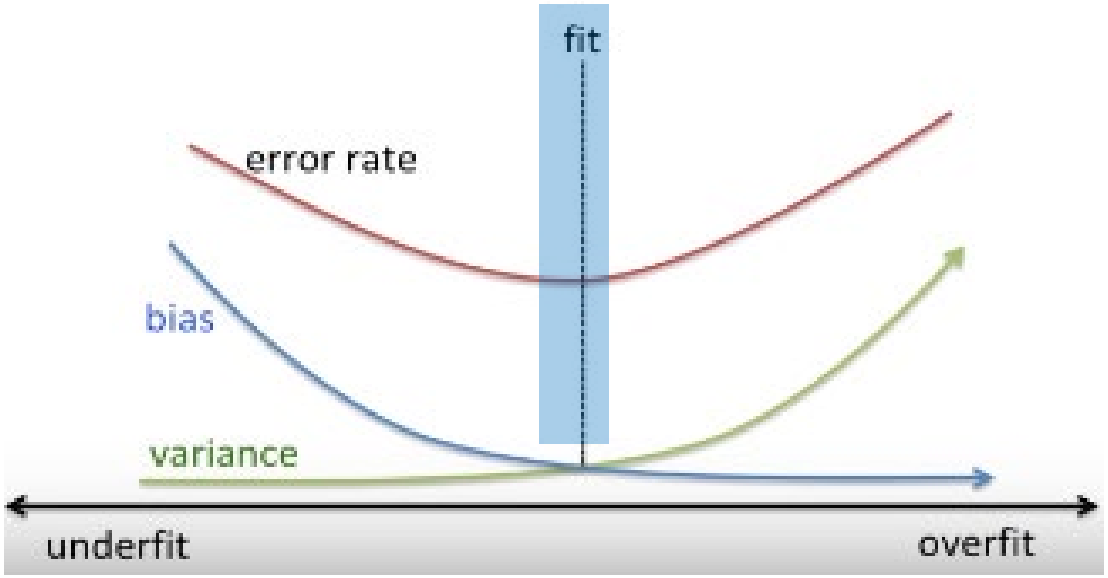
Feature Engineering

■ Feature Treatment

□ 변수 구간화를 사용하는 이유

(3) 과(소/대)적합을 완화 시켜주는 효과

- **과소적합 (Underfitting):** 모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못할 때 발생
- **과대적합 (Overfitting):** 모델이 훈련 데이터에 너무 잘 맞지만 테스트 데이터에는 잘 맞지 않음
  (일반성이 떨어진다는 의미 → 범용화하여 사용할 수 없는 모델)

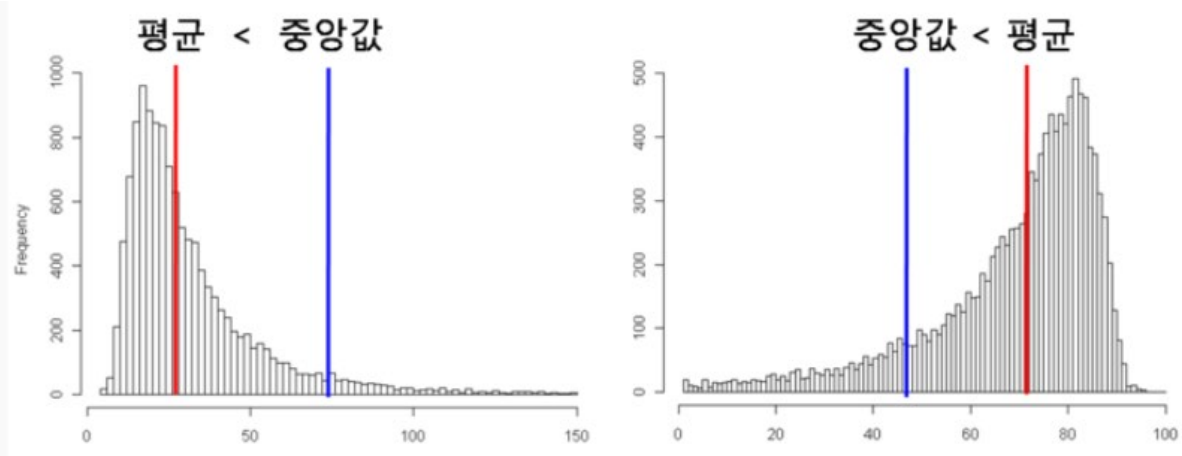


Feature Engineering

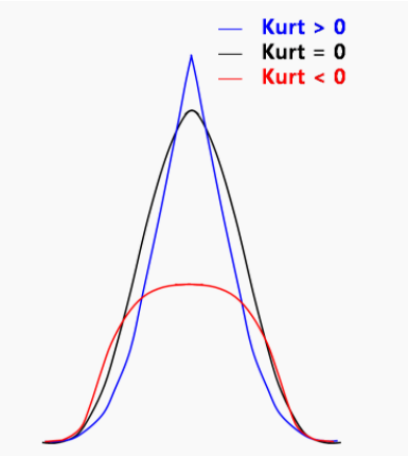
■ Feature Treatment

(3) Transform

- 기존에 존재하는 변수의 성질을 이용하여 다른 변수를 만드는 방법
- 날짜 데이터 => 주중/주말 or 오전/오후로 구분
- 특정 주기값 => 평균, 분산, Skewness (왜도), Kurtosis (첨도) 로 구분



Skewness
(비대칭도)



Kurtosis
(첨도)

Feature Engineering

■ Feature Treatment

(4) Dummy

- Binning 과는 반대로 범주형 변수를 연속형 (Numeric) 변수로 변환하기 위해 사용하는 방법
- 해당 범주 (Category)에 해당하는 경우 '1', 해당하지 않는 경우 '0'으로 값을 입력
- 통계나 기계학습 (머신러닝)을 수행할 때 컴퓨터가 범주형 자료값을 인식할 수 있도록 함

Emp_Code	Gender	Var_Male	Var_Female
A001	Male	1	0
A002	Female	0	1
A003	Female	0	1
A004	Male	1	0
A005	Female	0	1
A006	Male	1	0
A007	Male	1	0

Feature Engineering

■ Feature Engineering

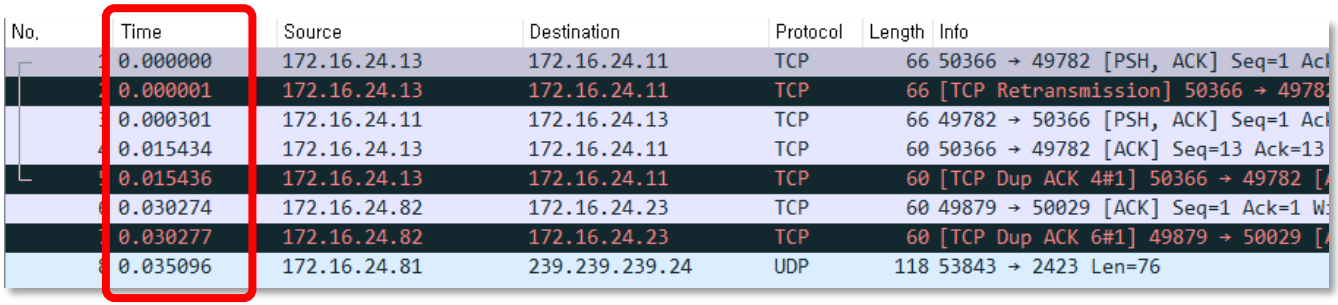
□ Classification 적용을 위한 Feature Engineering

Timestamp	CAN_ID	DLC	Data field
1466719494.987398	0x440	8	ff 00 00 00 ff 6f 08 00
1466719494.991160	0x2b0	5	d5 ff 00 07 2d
1466719494.992101	0x316	8	45 29 88 09 29 29 00 77
1466719494.992324	0x0a0	8	70 7e 88 09 00 26 02 00
1466719494.992567	0x0a1	8	80 83 00 00 2a 00 00 00
1466719494.992806	0x18f	8	fe 46 00 14 00 3c 00 00
1466719494.993066	0x002	8	00 00 00 00 00 03 07 8d
1466719494.993302	0x153	8	00 00 00 ff 00 ff 00 00
1466719494.993533	0x260	8	19 29 29 30 ff 89 63 00
1466719494.993782	0x2a0	8	00 00 74 1d 82 2d d5 0e
1466719494.994056	0x329	8	40 b3 7c 14 11 20 00 14

Feature Engineering

■ Feature Engineering

□ Classification 적용을 위한 Feature Engineering



No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	172.16.24.13	172.16.24.11	TCP	66	50366 → 49782 [PSH, ACK] Seq=1 Ac
2	0.000001	172.16.24.13	172.16.24.11	TCP	66	[TCP Retransmission] 50366 → 4978
3	0.000301	172.16.24.11	172.16.24.13	TCP	66	49782 → 50366 [PSH, ACK] Seq=1 Ac
4	0.015434	172.16.24.13	172.16.24.11	TCP	60	50366 → 49782 [ACK] Seq=13 Ack=13
5	0.015436	172.16.24.13	172.16.24.11	TCP	60	[TCP Dup ACK 4#1] 50366 → 49782 [
6	0.030274	172.16.24.82	172.16.24.23	TCP	60	49879 → 50029 [ACK] Seq=1 Ack=1 W
7	0.030277	172.16.24.82	172.16.24.23	TCP	60	[TCP Dup ACK 6#1] 49879 → 50029 [
8	0.035096	172.16.24.81	239.239.239.24	UDP	118	53843 → 2423 Len=76

- **Features**
  - 네트워크 통신량 단위 BPS, PPS
  - 정의된 시간단위 내 Source 와 Destination 사이의 Entropy
  - 동일한 Source 와 Destination 간의 Time interval

Timestamp	CAN_ID	DLC	Data field
1466719494.987398	0x440	8	ff 00 00 00 ff 6f 08 00
1466719494.991160	0x2b0	5	d5 ff 00 07 2d
1466719494.992101	0x316	8	45 29 88 09 29 29 00 77
1466719494.992324	0x0a0	8	70 7e 88 09 00 26 02 00
1466719494.992567	0x0a1	8	80 83 00 00 2a 00 00 00
1466719494.992806	0x18f	8	fe 46 00 14 00 3c 00 00
1466719494.993066	0x002	8	00 00 00 00 00 03 07 8d
1466719494.993302	0x153	8	00 00 00 ff 00 ff 00 00
1466719494.993533	0x260	8	19 29 29 30 ff 89 63 00
1466719494.993782	0x2a0	8	00 00 74 1d 82 2d d5 0e
1466719494.994056	0x329	8	40 b3 7c 14 11 20 00 14

- **Features**
  - 정의된 시간단위 내 CAN ID의 Entropy
  - 정의된 시간단위 내 CAN ID의 Survival rate
  - 정의된 시간단위 내 CAN ID의 Cosine similarity
  - CAN ID의 Sequences similarity
  - Data field 의 CAN ID 기준 Distance

Feature Engineering

■ Feature Engineering

□ **Clustering** 적용을 위한 Feature Engineering

id	date	notify	domain	ip	location	system	server
24783971	2015-09-01 8:34	hayder-sql	http://www.travelsitehub.com/ass.html	108.179.232.80	Unknown	Linux	Apache
24783970	2015-09-01 8:33	Theking	http://xn--mgbc5a2e.com	144.76.76.231	Unknown	Linux	nginx
24783954	2015-09-01 8:11	jangene_cak	http://grinandbakeit.com/a.txt	74.220.215.85	Unknown	Linux	Apache
24783945	2015-09-01 8:09	OmAax	http://lilleaamosemuseum.dk	212.97.132.146	WestEuro	Linux	LiteSpeed
24783883	2015-09-01 7:05	3xp1r3	http://myviewsinmywords.com/wp-login.php	67.227.144.7	Unknown	Linux	Apache
24783875	2015-09-01 6:52	AnonJoker	http://superbowltemplate.com/lolz.php	198.57.149.247	Unknown	Linux	nginx
24783874	2015-09-01 6:52	AnonJoker	http://superbowlsquaresonline.com/lolz.php	198.57.149.247	Unknown	Linux	nginx
24783831	2015-09-01 6:20	Angel Dot I	http://phone-sex-milf.com/angel.htm	198.252.74.35	Unknown	Linux	Apache
24783423	2015-09-01 5:16	4Ri3 60ndr0	http://tecnicoesquerda.com.br/xxx.htm	186.202.127.228	SouthAmeri	Linux	Apache
24783408	2015-09-01 4:43	Fallaga Tean	http://wbconstruction.co.za	41.193.5.45	Africa	Linux	Apache
24783262	2015-09-01 4:01	BlackHeart	http://highstyle.in	198.23.192.231	Asia	Linux	Apache
24783246	2015-09-01 3:23	NeT-Bug"	http://www.pacificunified.com	162.219.44.2	Unknown	Linux	Apache
24783222	2015-09-01 2:55	4Ri3 60ndr0	http://ofertapatos.com.br/xxx.htm	177.154.48.130	SouthAmeri	Linux	Apache
24783206	2015-09-01 2:29	dowoh	http://kabarkalteng.info/wp-login.php	202.43.182.28	Unknown	Linux	Apache
24783119	2015-09-01 0:57	josef parado	http://www.vakinankaratra.gov.mg	41.190.237.22	Africa	Linux	Apache
24783114	2015-09-01 0:47	4Ri3 60ndr0	http://valasysb2bmarketing.com/xxx.htm	66.199.227.122	Unknown	Linux	Apache
24783106	2015-09-01 0:40	BlueTornado	http://cercialani.com	72.29.72.141	Unknown	Linux	Apache
24783103	2015-09-01 0:37	4Ri3 60ndr0	http://shoppingsukh.com/xxx.htm	66.199.227.122	Unknown	Linux	Apache
24782717	2015-08-31 19:48	Nofawkx Al	http://dinderlocal.gov.sd	196.29.187.156	Africa	Linux	Apache
24781086	2015-08-31 7:54	dowoh	http://laporan.pta-makassarkota.go.id	49.50.8.67	Asia	Linux	Apache
24777453	2015-08-29 15:57	bl4ck-dz	http://training.onab.go.th/index.htm	202.44.53.251	Asia	Win 2008	IIS/7.0
24777452	2015-08-29 15:57	bl4ck-dz	http://tak.onab.go.th/index.htm	202.44.53.251	Asia	Win 2008	IIS/7.0
24777445	2015-08-29 15:56	bl4ck-dz	http://srn.onab.go.th/index.htm	202.44.53.251	Asia	Win 2008	IIS/7.0
24777441	2015-08-29 15:56	bl4ck-dz	http://snk.onab.go.th/index.htm	202.44.53.251	Asia	Win 2008	IIS/7.0

? ? ?

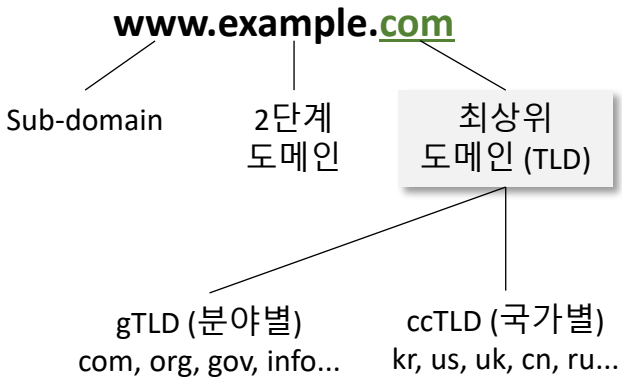
Feature Engineering

- gTLD: 일반 최상위 도메인
- ccTLD: 국가 코드 최상위 도메인

■ Feature Engineering

□ Clustering 적용을 위한 Feature Engineering

Domain	gTLD	ccTLD	ccTLD_Grouping
corp1ontheweb.com	com		
www.albemarlebulkheads.com	com		
www.kanggyeong.ms.kr		kr	EastAsia
apmab.ap.nic.in		in	SouthAsia
yll.loxa.edu.tw	edu	tw	EastAsia
m4.dpps.tcc.edu.tw	edu	tw	EastAsia
www.tcvhs.ylc.edu.tw	edu	tw	EastAsia
www.gcep.gov.jo	gov	jo	WestAsia
www.xyl.cc		cc	Australia
www.premyier.com	com		
www.duo-california.at		at	WesternEurope
bits.bpa.arizona.edu	edu		



OS	OS_Grouping
Win 2000	Window
Linux	Linux
Win NT9x	Window
Linux	Linux
Win 2003	Window
Linux	Linux
Linux	Linux
Win 2003	Window
Win 2000	Window
FreeBSD	Unix
FreeBSD	Unix
Win 2000	Window

Encoding	Encoding_Grouping
big5	Taiwanese
big5	Taiwanese
big5	Taiwanese
gb2312	Chinese
gb2312	Chinese
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope

- **Window**: Window 9x 기반, Window NT 기반
- **Linux**: RedHat, Fedora, CentOS, Tizen...
- **Unix**: BSD, Solaris, IRIX, HP-UX...

Character Encoding

- 문자 인코딩: 문자를 코드로 변환하는 것
  - ASCII, ANSI
  - Unicode (UTF-8, UTF-16)
  - Percent(%) Encoding

예) cp949, iso-2022-kr, ks_c_5601-1987, euc-kr
-> 한글 표현을 위한 문자 인코딩

Feature Engineering

- gTLD: 일반 최상위 도메인
- ccTLD: 국가 코드 최상위 도메인

■ Feature Engineering

□ Clustering 적용을 위한 Feature Engineering

Domain	gTLD	ccTLD	ccTLD_Grouping
corp1ontheweb.com	com		
www.albemarlebulkheads.com	com		
www.kanggyeong.ms.kr		kr	EastAsia
apmab.ap.nic.in		in	SouthAsia
yll.loxa.edu.tw	edu	tw	EastAsia
m4.dpps.tcc.edu.tw	edu	tw	EastAsia
www.tcvhs.ylc.edu.tw	edu	tw	EastAsia
www.gcep.gov.jo	gov	jo	WestAsia
www.xyl.cc		cc	Australia
www.premyier.com	com		
www.duo-california.at		at	WesternEurope
bits.bpa.arizona.edu	edu		

OS	OS_Grouping
Win 2000	Window
Linux	Linux
Win NT9x	Window
Linux	Linux
Win 2003	Window
Linux	Linux
Linux	Linux
Win 2003	Window
Win 2000	Window
FreeBSD	Unix
FreeBSD	Unix
Win 2000	Window

Encoding	Encoding_Grouping
big5	Taiwanese
big5	Taiwanese
big5	Taiwanese
gb2312	Chinese
gb2312	Chinese
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope

- **Window**: Window 9x 기반, Window NT 기반
- **Linux**: RedHat, Fedora, CentOS, Tizen...
- **Unix**: BSD, Solaris, IRIX, HP-UX...

Character Encoding

- 문자 인코딩: 문자를 코드로 변환하는 것
  - ASCII, ANSI
  - Unicode (UTF-8, UTF-16)
  - Percent(%) Encoding

예) cp949, iso-2022-kr, ks_c_5601-1987, euc-kr
-> 한글 표현을 위한 문자 인코딩

Feature Engineering

- gTLD: 일반 최상위 도메인
- ccTLD: 국가 코드 최상위 도메인

■ Feature Engineering

□ Clustering 적용을 위한 Feature Engineering

Domain	gTLD	ccTLD	ccTLD_Grouping
corp1ontheweb.com	com		
www.albemarlebulkheads.com	com		
www.kanggyeong.ms.kr		kr	EastAsia
apmab.ap.nic.in		in	SouthAsia
yll.loxa.edu.tw	edu	tw	EastAsia
m4.dpps.tcc.edu.tw	edu	tw	EastAsia
www.tcvhs.ylc.edu.tw	edu	tw	EastAsia
www.gcep.gov.jo	gov	jo	WestAsia
www.xyl.cc		cc	Australia
www.premyier.com	com		
www.duo-california.at		at	WesternEurope
bits.bpa.arizona.edu	edu		

OS	OS_Grouping
Win 2000	Window
Linux	Linux
Win NT9x	Window
Linux	Linux
Win 2003	Window
Linux	Linux
Linux	Linux
Win 2003	Window
Win 2000	Window
FreeBSD	Unix
FreeBSD	Unix
Win 2000	Window

Encoding	Encoding_Grouping
big5	Taiwanese
big5	Taiwanese
big5	Taiwanese
gb2312	Chinese
gb2312	Chinese
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope
iso-8859-1	WestEurope

- **Window**: Window 9x 기반, Window NT 기반
- **Linux**: RedHat, Fedora, CentOS, Tizen...
- **Unix**: BSD, Solaris, IRIX, HP-UX...

Character Encoding

- **문자 인코딩: 문자를 코드로 변환하는 것**
  - **ASCII, ANSI**
  - **Unicode (UTF-8, UTF-16)**
  - **Percent(%) Encoding**

예) cp949, iso-2022-kr, ks_c_5601-1987, euc-kr
-> 한글 표현을 위한 문자 인코딩

Feature Engineering

■ Underfitting & Overfitting

□ 과소적합 (Underfitting)

- 과소적합은 모델 학습 시 충분한 데이터의 특징을 활용하지 못할 경우 발생

사물	Class (분류)	생김새
야구공	공	원형
축구공	공	원형
테니스공	공	원형
딸기	과일	세모
포도알	공!?!	원형

Feature

학습데이터
(Training sets)

테스트데이터
(Testing sets)

오탐
(False Positive)

- 데이터 특징이 생김새 밖에 존재하지 않으므로, 생김새가 원형인 경우는 모두 공이라고 분류하게 됨
  → 오탐 발생

Feature Engineering

■ Underfitting & Overfitting

□ 과소적합 (Underfitting)

- 과소적합은 모델 학습 시 충분한 데이터의 특징을 활용하지 못할 경우 발생

- 데이터 특징이 생김새 밖에 존재하지 않으므로, 생김새가 원형인 경우는 모두 공이라고 분류하게 됨

→ 오탐 발생

Feature Engineering

■ Underfitting & Overfitting

□ 과대적합 (Overfitting)

- 필요 이상의 특징을 발견한 후 학습 데이터에서는 높은 정확도를 보이나, 테스트 데이터 (새로운 데이터)에서는 낮은 정확도를 보이는 경우

학습데이터
(Training sets)

사물	Class (분류)	생김새	크기	줄무늬
야구공	공	원형	중간	있음
축구공	공	원형	큼	있음
테니스공	공	원형	중간	있음
딸기	과일	세모	작음	없음
포도알	과일	원형	작음	없음

Features

- 생김새가 원형이고
- 크기가 작지 않으며
- 줄무늬가 있으면

공!! (학습)

테스트데이터
(Testing sets)

사물	Class (분류)	생김새	크기	줄무늬
골프공	공	원형	작음	없음
수박	공!?!?	원형	큼	있음
당구공	공	원형	중간	없음
럭비공	공	타원형	큼	있음
볼링공	공	원형	큼	없음

미탐 (False Negative)

오탐 (False Positive)

Thank you

KOREA
UNIVERSITY