R 교육 세미나 1주차

ToBig's 7기 이광록

Rgorithm

알아두면 1정도는 도움이 될 지식들

Contents

Unit 01. prologue

들어가기에 앞서

- 여러분은 무엇을 하러 이 곳에 오셨나요?

데이터 마이닝

- 데이터 광산에서 의미를 채굴하는 일
- '빅'데이터가 되어서 사람 눈으로 패턴을 찾기가 힘듬
- 기계가 정해진 알고리즘에 따라 패턴을 찾음 → 머신러닝

머신 러닝

- 머신 러닝 ⊂ 인공 지능(A.I.)
- 데이터 속에서(데이터 마이닝) 통계적 지식을(통계) 디지털 환경에서 구현(컴퓨터 과학)한 것

Unit 01 | prologue

들어가기에 앞서

- 여러분은 무엇을 하러 이 곳에 오셨나요?
- 우리가 할 것은 머신 러닝을 활용한 데이터 분석
- 그 전반적인 과정에 대해 알아보자

Unit 02. 데이터 분석하기

데이터 분석

데이터 관찰 → 모델 선정 → 데이터 전처리 → 모델 적합 & 테스트 → 모델 튜닝

	age	workclass	fnlwgt	education	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
1	49	Local-gov	223342	Some-college	Divorced	Adm-clerical	Not-in-family	White	Female	0	0	44	United-States	small
2	42	Federal-gov	108183	Masters	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	1902	40	South	large
3	63	Private	30813	Masters	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	50	United-States	large
4	43	Private	125461	Bachelors	Married-civ-spouse	Sales	Husband	White	Male	0	0	65	United-States	large
5	48	Private	143299	HS-grad	Never-married	Machine-op-inspct	Not-in-family	Black	Male	0	0	40	United-States	small
6	34	Self-emp-not-inc	203488	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	60	United-States	small
7	24	Private	196674	Bachelors	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	NA
8	34	Private	113198	Assoc-acdm	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	28	United-States	small
9	48	Private	249935	HS-grad	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	44	United-States	small
10	54	State-gov	123592	HS-grad	Separated	Adm-clerical	Unmarried	Black	Female	3887	0	35	United-States	small
11	34	Local-gov	93886	Bachelors	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	46	United-States	large
12	28	Private	285897	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
13	40	Private	207025	HS-grad	Never-married	Adm-clerical	Not-in-family	White	Female	6849	0	38	United-States	small
14	53	Private	217568	HS-grad	Widowed	Craft-repair	Unmarried	Black	Female	0	0	40	United-States	small
15	57	State-gov	222792	Some-college	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	40	United-States	NA
16	36	Private	75826	10th	Separated	Machine-op-inspct	Not-in-family	White	Female	0	0	40	United-States	NA
17	18	Private	192409	12th	Never-married	Other-service	Own-child	White	Female	0	0	25	United-States	small
18	33	Private	122116	HS-grad	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	small
19	39	Private	154641	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	large
20	42	Private	29702	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	42	United-States	NA
21	61	Private	105384	Bachelors	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	small
22	52	Self-emp-not-inc	95082	HS-grad	Married-civ-spouse	Sales	Husband	White	Male	0	0	60	United-States	NA
23	18	Private	106780	Some-college	Never-married	Other-service	Own-child	White	Female	0	0	12	United-States	small
24	20	NA	50163	Some-college	Never-married	NA	Not-in-family	White	Male	0	0	25	United-States	NA
25	45	Private	116163	HS-grad	Separated	Exec-managerial	Not-in-family	White	Female	0	0	40	United-States	small
26	31	Private	162572	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	small
27	45	Private	256866	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	45	United-States	NA
28	42	State-gov	147206	Assoc-voc	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
29	53	Local-gov	192982	Masters	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	38	United-States	large
30	32	NA	227160	Some-college	Divorced	NA	Not-in-family	White	Male	0	0	40	United-States	small
31	64	Self-emp-not-inc	388625	10th	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	10	United-States	large
32	48	Self-emp-not-inc	259412	Prof-school	Married-civ-spouse	Sales	Husband	White	Male	0	0	20	United-States	NA
33	24	Private	187717	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	0	0	40	United-States	small
34	62	Private	82906	Bachelors	Married-civ-spouse	Exec-managerial	Wife	White	Female	4064	0	35	England	NA
35	78	Private	135692	Some-college	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	NA
36	22	NA	125040	Some-college	Never-married	NA	Own-child	White	Male	0	0	40	United-States	NA

데이터 관찰

	age	workclass	fnlwgt	education	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
1	49	Local-gov	223342	Some-college	Divorced	Adm-clerical	Not-in-family	White	Female	0	0	44	United-States	small
2	42	Federal-gov	108183	Masters	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	1902	40	South	large
3	63	Private	30813	Masters	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	50	United-States	large
4	43	Private	125461	Bachelors	Married-civ-spouse	Sales	Husband	White	Male	0	0	65	United-States	large
5	48	Private	143299	HS-grad	Never-married	Machine-op-inspct	Not-in-family	Black	Male	0	0	40	United-States	small
6	34	Self-emp-not-inc	203488	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	60	United-States	small
7	24	Private	196674	Bachelors	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	NA
8	34	Private	113198	Assoc-acdm	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	28	United-States	small
9	48	Private	249935	HS-grad	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	44	United-States	small
10	54	State-gov	123592	HS-grad	Separated	Adm-clerical	Unmarried	Black	Female	3887	0	35	United-States	small
11	34	Local-gov	93886	Bachelors	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	46	United-States	large
12	28	Private	285897	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
13	40	Private	207025	HS-grad	Never-married	Adm-clerical	Not-in-family	White	Female	6849	0	38	United-States	small
14	53	Private	217568	HS-grad	Widowed	Craft-repair	Unmarried	Black	Female	0	0	40	United-States	small
15	57	State-gov	222792	Some-college	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	40	United-States	NA
16	36	Private	75826	10th	Separated	Machine-op-inspct	Not-in-family	White	Female	0	0	40	United-States	NA
17	18	Private	192409	12th	Never-married	Other-service	Own-child	White	Female	0	0	25	United-States	small
18	33	Private	122116	HS-grad	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	small
19	39	Private	154641	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	large
20	42	Private	29702	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	42	United-States	NA
21	61	Private	105384	Bachelors	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	small
22	52	Self-emp-not-inc	95082	HS-grad	Married-civ-spouse	Sales	Husband	White	Male	0	0	60	United-States	NA
23	18	Private	106780	Some-college	Never-married	Other-service	Own-child	White	Female	0	0	12	United-States	small
24	20	NA	50163	Some-college	Never-married	NA	Not-in-family	White	Male	0	0	25	United-States	NA
25	45	Private	116163	HS-grad	Separated	Exec-managerial	Not-in-family	White	Female	0	0	40	United-States	small
26	31	Private	162572	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	small
27	45	Private	256866	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	45	United-States	NA
28	42	State-gov	147206	Assoc-voc	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
29	53	Local-gov	192982	Masters	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	38	United-States	large
30	32	NA	227160	Some-college	Divorced	NA	Not-in-family	White	Male	0	0	40	United-States	small
31	64	Self-emp-not-inc	388625	10th	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	10	United-States	large
32	48	Self-emp-not-inc	259412	Prof-school	Married-civ-spouse	Sales	Husband	White	Male	0	0	20	United-States	NA
33	24	Private	187717	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	0	0	40	United-States	small
34	62	Private	82906	Bachelors	Married-civ-spouse	Exec-managerial	Wife	White	Female	4064	0	35	England	NA
35	78	Private	135692	Some-college	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	NA
36	22	NA	125040	Some-college	Never-married	NA	Own-child	White	Male	0	0	40	United-States	NA

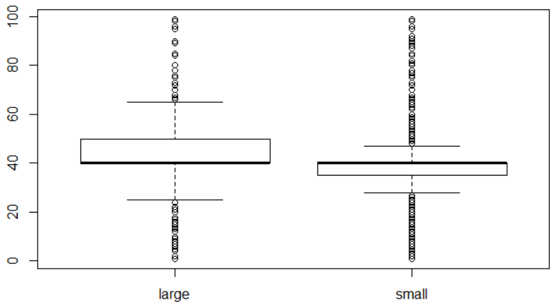
- 데이터를 살피면서 인사이트를 얻는 과정

데이터 관찰

```
'data.frame':	34189 obs. of  14 variables:
 $ age           : int  49 42 63 43 48 34 24 34 48 54 ...
 $ workclass     : Factor w/ 8 levels "Federal-gov",..: 2 1 4 4 4 6 4 4 4 7 ...
 $ fnlwgt        : int  223342 108183 30813 125461 143299 203488 196674 113198 249935 123592 ...
 $ education     : Factor w/ 16 levels "10th","11th",..: 16 13 13 10 12 12 10 8 12 12 ...
 $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 1 3 3 3 5 3 3 3 3 6 ...
 $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 1 10 10 12 7 3 10 1 14 1 ...
 $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 1 1 2 1 1 1 1 5 ...
 $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 2 5 5 3 5 5 5 5 3 ...
 $ sex           : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 2 2 1 ...
 $ capital.gain  : int  0 0 0 0 0 0 0 0 0 3887 ...
 $ capital.loss  : int  0 1902 0 0 0 0 0 0 0 0 ...
 $ hours.per.week: int  44 40 50 65 40 60 40 28 44 35 ...
 $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 39 35 39 39 39 39 39 39 39 39 ...
 $ income        : Factor w/ 2 levels "large","small": 2 1 1 1 2 2 NA 2 2 2 ...
```

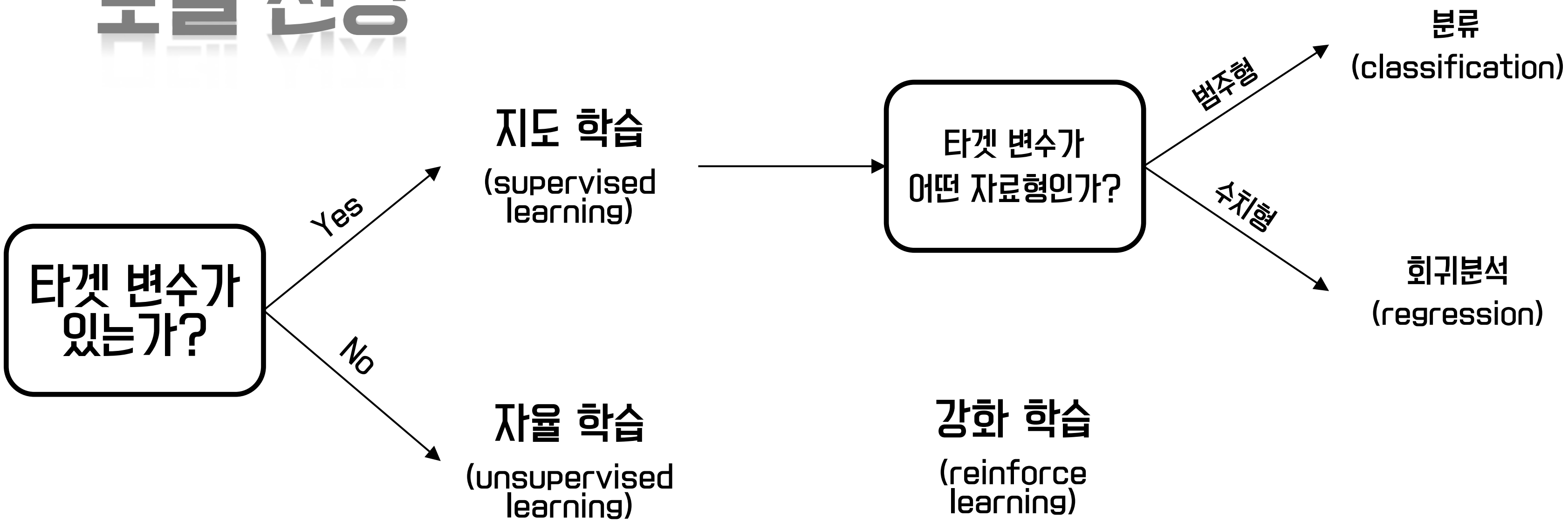
- 데이터를 살피면서 인사이트를 얻는 과정
- 대략적인 데이터 형태, 특성, 분포 등을 살핌

데이터 관찰

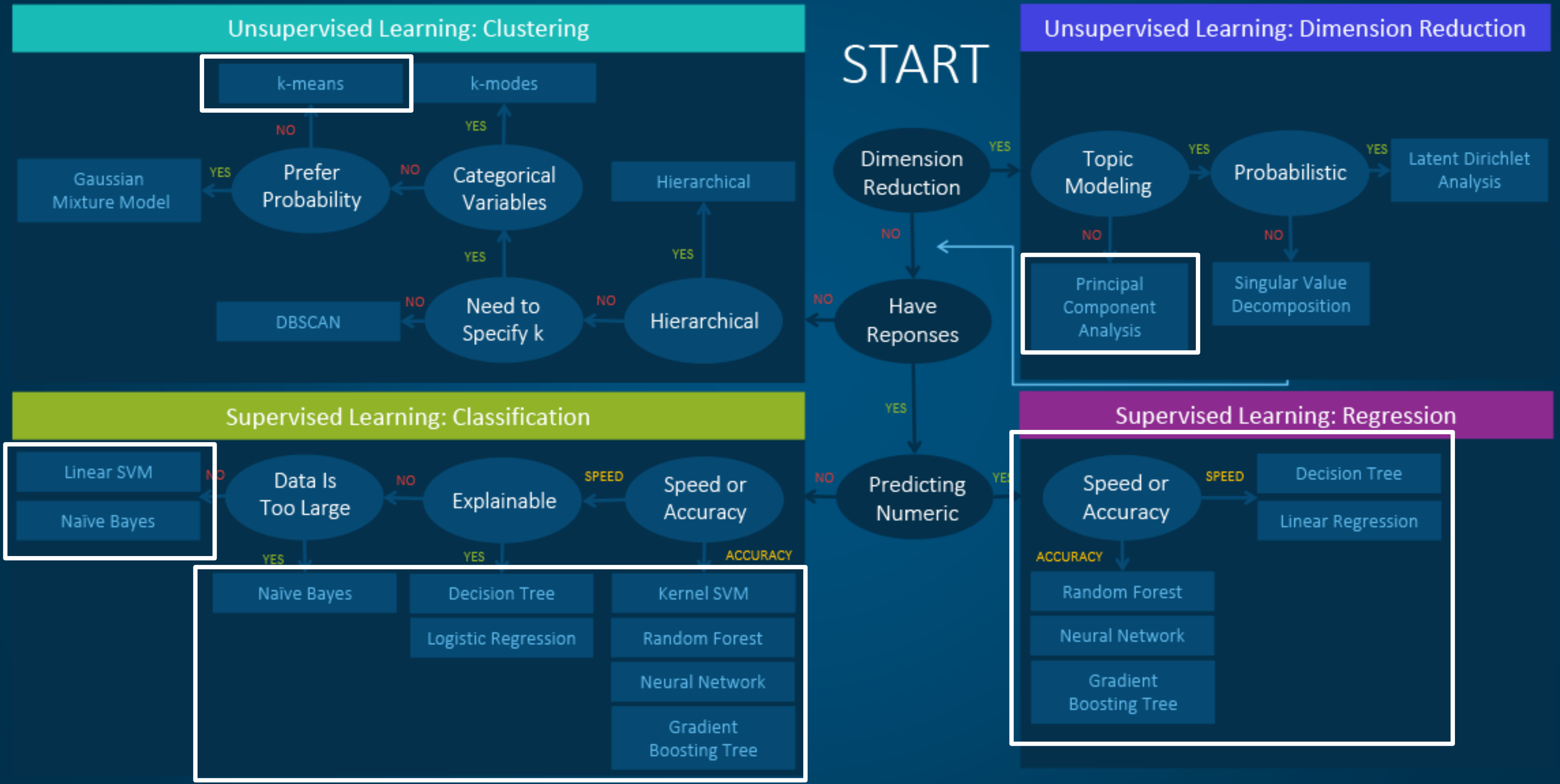


- 데이터를 살피면서 인사이트를 얻는 과정
- 대략적인 데이터 형태, 특성, 분포 등을 살핌
- 그래프, 히스토그램, 테이블 등을 그려보면서 데이터에 대한 파악을 함

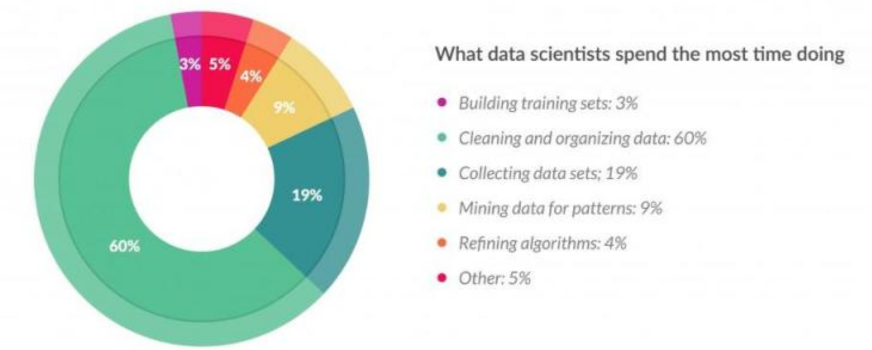
모델 선정

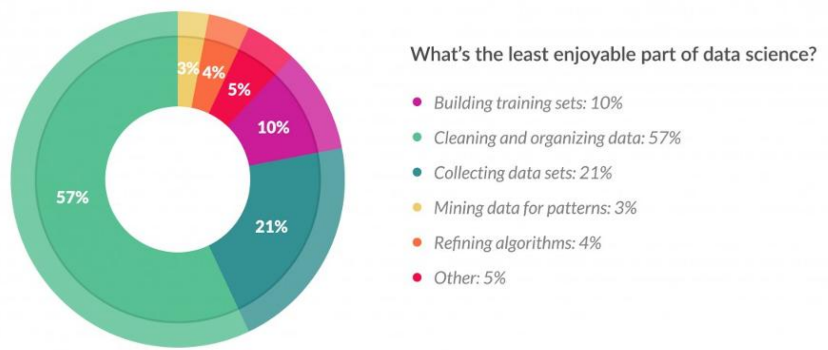

타겟 변수가 있는가?
Yes
No
지도 학습
(supervised learning)
자율 학습
(unsupervised learning)
타겟 변수가 어떤 자료형인가?
범주형
수치형
분류
(classification)
회귀분석
(regression)
강화 학습
(reinforce learning)

Machine Learning Algorithms Cheat Sheet



데이터 전처리


3%
5%
4%
9%
19%
60%
What data scientists spend the most time doing
Building training sets: 3%
Cleaning and organizing data: 60%
Collecting data sets; 19%
Mining data for patterns: 9%
Refining algorithms: 4%
Other: 5%


3%
4%
5%
10%
57%
21%
What's the least enjoyable part of data science?
Building training sets: 10%
Cleaning and organizing data: 57%
Collecting data sets: 21%
Mining data for patterns: 3%
Refining algorithms: 4%
Other: 5%

```
$ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 1 3 3 3 5 3 3 3 3 6 ...
$ occupation    : Factor w/ 14 levels "Adm-clerical",..: 1 10 10 12 7 3 10 1 14 1 ...
$ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 1 1 2 1 1 1 1 5 ...
$ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 2 5 5 3 5 5 5 5 3 ...
$ sex           : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 2 2 1 ...
```

	age	workclass	fnlwgt	education	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
1	49	Local-gov	223342	Some-college	Divorced	Adm-clerical	Not-in-family	White	Female	0	0	44	United-States	small
2	42	Federal-gov	108183	Masters	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	1902	40	South	large
3	63	Private	30813	Masters	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	50	United-States	large
4	43	Private	125461	Bachelors	Married-civ-spouse	Sales	Husband	White	Male	0	0	65	United-States	large
5	48	Private	143299	HS-grad	Never-married	Machine-op-inspct	Not-in-family	Black	Male	0	0	40	United-States	small
6	34	Self-emp-not-inc	203488	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	60	United-States	small
7	24	Private	196674	Bachelors	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	NA
8	34	Private	113198	Assoc-acdm	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	28	United-States	small
9	48	Private	249935	HS-grad	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	44	United-States	small
10	54	State-gov	123592	HS-grad	Separated	Adm-clerical	Unmarried	Black	Female	3887	0	35	United-States	small
11	34	Local-gov	93886	Bachelors	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	46	United-States	large
12	28	Private	285897	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
13	40	Private	207025	HS-grad	Never-married	Adm-clerical	Not-in-family	White	Female	6849	0	38	United-States	small
14	53	Private	217568	HS-grad	Widowed	Craft-repair	Unmarried	Black	Female	0	0	40	United-States	small
15	57	State-gov	222792	Some-college	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	40	United-States	NA
16	36	Private	75826	10th	Separated	Machine-op-inspct	Not-in-family	White	Female	0	0	40	United-States	NA
17	18	Private	192409	12th	Never-married	Other-service	Own-child	White	Female	0	0	25	United-States	small
18	33	Private	122116	HS-grad	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	small
19	39	Private	154641	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	large
20	42	Private	29702	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	42	United-States	NA
21	61	Private	105384	Bachelors	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	small
22	52	Self-emp-not-inc	95082	HS-grad	Married-civ-spouse	Sales	Husband	White	Male	0	0	60	United-States	NA
23	18	Private	106780	Some-college	Never-married	Other-service	Own-child	White	Female	0	0	12	United-States	small
24	20	NA	50163	Some-college	Never-married	NA	Not-in-family	White	Male	0	0	25	United-States	NA
25	45	Private	116163	HS-grad	Separated	Exec-managerial	Not-in-family	White	Female	0	0	40	United-States	small
26	31	Private	162572	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	small
27	45	Private	256866	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	45	United-States	NA
28	42	State-gov	147206	Assoc-voc	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
29	53	Local-gov	192982	Masters	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	38	United-States	large
30	32	NA	227160	Some-college	Divorced	NA	Not-in-family	White	Male	0	0	40	United-States	small
31	64	Self-emp-not-inc	388625	10th	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	10	United-States	large
32	48	Self-emp-not-inc	259412	Prof-school	Married-civ-spouse	Sales	Husband	White	Male	0	0	20	United-States	NA
33	24	Private	187717	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	0	0	40	United-States	small
34	62	Private	82906	Bachelors	Married-civ-spouse	Exec-managerial	Wife	White	Female	4064	0	35	England	NA
35	78	Private	135692	Some-college	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	NA
36	22	NA	125040	Some-college	Never-married	NA	Own-child	White	Male	0	0	40	United-States	NA

데이터 전처리

1. 데이터 정제 (Data Cleaning) : 결측값을 채우거나, 이상점를 발견하여 이를 제거하고 불일치를 해결하는 과정
2. 데이터 통합 (Data Integration) : 데이터들을 하나의 형태로 합쳐서 표현
3. 데이터 정리 (Data Reduction) : 분석 결과에 영향을 끼치지는 않지만, 데이터 크기를 줄임
4. 데이터 변환 (Data Transformation) : 모델에 적합한 데이터로의 변환

데이터 전처리

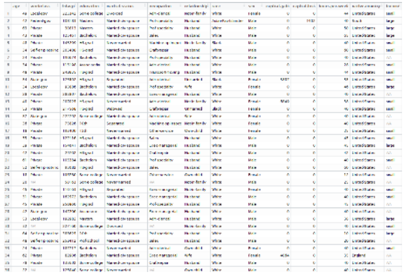
1. 데이터 정제 (Data Cleaning)
   a. 결측값 처리
      - 삭제
      - 대체
      - 예측값 삽입

데이터 전처리

1. 데이터 정제 (Data Cleaning)
   b. 이상치 처리
   - 삭제
   - 대체
   - 분리

데이터 전처리

2. 데이터 통합 (Data Integration)



+

	education	education.num
1	Some-college	10
2	Masters	14
3	Bachelors	13
4	HS-grad	9
5	Assoc-acdm	12
6	10th	6
7	12th	8
8	Assoc-voc	11
9	Prof-school	15
10	5th-6th	3
11	11th	7
12	7th-8th	4
13	Preschool	1
14	Doctorate	16
15	1st-4th	2
16	9th	5

=

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
1	49	Local-gov	223342	Some-college	10	Divorced	Adm-clerical	Not-in-family	White	Female	0	0	44	United-States	small
2	42	Federal-gov	108183	Masters	14	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	1902	40	South	large
3	63	Private	30813	Masters	14	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	50	United-States	large
4	43	Private	125461	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	65	United-States	large
5	48	Private	143299	HS-grad	9	Never-married	Machine-op-inspct	Not-in-family	Black	Male	0	0	40	United-States	small
6	34	Self-emp-not-inc	203488	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	60	United-States	small
7	24	Private	196674	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	NA
8	34	Private	113198	Assoc-acdm	12	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	28	United-States	small
9	48	Private	249935	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	44	United-States	small
10	54	State-gov	123592	HS-grad	9	Separated	Adm-clerical	Unmarried	Black	Female	3887	0	35	United-States	small
11	34	Local-gov	93886	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	46	United-States	large
12	28	Private	285897	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	NA
13	40	Private	207025	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	6849	0	38	United-States	small
14	53	Private	217568	HS-grad	9	Widowed	Craft-repair	Unmarried	Black	Female	0	0	40	United-States	small
15	57	State-gov	222792	Some-college	10	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	40	United-States	NA
16	36	Private	75826	10th	6	Separated	Machine-op-inspct	Not-in-family	White	Female	0	0	40	United-States	NA
17	18	Private	192409	12th	8	Never-married	Other-service	Own-child	White	Female	0	0	25	United-States	small
18	33	Private	122116	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	small
19	39	Private	154641	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	large
20	42	Private	29702	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	42	United-States	NA
21	61	Private	105384	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	small
22	52	Self-emp-not-inc	95082	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	60	United-States	NA

데이터 전처리

3. 데이터 정리 (Data Reduction)

데이터 전처리

4. 데이터 변환 (Data Transformation)

$$Z = \frac{X - \mu}{\sigma}$$

$$Z \sim N(0,1)$$

모델 적합 & 평가

이제 진짜 데이터를 모델에 넣고 돌려볼 시간! (supervised learning)

1. 전처리된 데이터를 훈련용 데이터(train data)와 평가용 데이터(test data)로 나누고,
2. 훈련용 데이터로 모델을 학습 시키고
3. 평가용 데이터로 모델이 잘 작동하는 지 확인

모델 튜닝

- 데이터를 더욱 맛깔나게 수정
- 하이퍼 파라미터(hyper parameter) 수정 (k-fold Cross-Validation)
- 부트스트랩, 배깅, 앙상블 등등

최적의 모델을 완성?

그럼 이제 튜닝도 다 끝나고 여기저기 써먹을 수 있는 최적의 모델을 완성한건가?

– 공짜 점심은 없다.

IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 1, NO. 1, APRIL 1997 67

No Free Lunch Theorems for Optimization

David H. Wolpert and William G. Macready

***Abstract*—A framework is developed to explore the connection between effective optimization algorithms and the problems they are solving. A number of "no free lunch" (NFL) theorems are presented which establish that for any algorithm, any elevated performance over one class of problems is offset by performance over another class. These theorems result in a geometric interpretation of what it means for an algorithm to be well suited to an optimization problem. Applications of the NFL theorems to information-theoretic aspects of optimization and benchmark measures of performance are also presented. Other issues addressed include time-varying optimization problems and *a priori* "head-to-head" minimax distinctions between optimization algorithms, distinctions that result despite the NFL theorems' enforcing of a type of uniformity over all algorithms.**

***Index Terms*— Evolutionary algorithms, information theory, optimization.**

I. Introduction

THE past few decades have seen an increased interest in general-purpose "black-box" optimization algorithms that exploit limited knowledge concerning the optimization problem on which they are run. In large part these algorithms have drawn inspiration from optimization processes that occur in nature. In particular, the two most popular black-box optimization strategies, evolutionary algorithms [1]–[3] and simulated annealing [4], mimic processes in natural selection and statistical mechanics, respectively.

In light of this interest in general-purpose optimization algorithms, it has become important to understand the relationship between how well an algorithm a performs and the optimization problem f on which it is run. In this paper we present a formal analysis that contributes toward such an understanding by addressing questions like the following: given the abundance of black-box optimization algorithms and of optimization problems, how can we best match algorithms to problems (i.e., how best can we relax the black-box nature of the algorithms and have them exploit some knowledge concerning the optimization problem)? In particular, while serious optimization practitioners almost always perform such matching, it is usually on a heuristic basis; can such matching be formally analyzed? More generally, what is the underlying mathematical "skeleton" of optimization theory before the "flesh" of the probability distributions of a particular context and set of optimization problems are imposed? What can information theory and Bayesian analysis contribute to an understanding of these issues? How *a priori* generalizable are the performance results of a certain algorithm on a certain class of problems to its performance on other classes of problems? How should we even measure such generalization? How should we assess the performance of algorithms on problems so that we may programmatically compare those algorithms?

Broadly speaking, we take two approaches to these questions. First, we investigate what *a priori* restrictions there are on the performance of one or more algorithms as one runs over the set of all optimization problems. Our second approach is to instead focus on a particular problem and consider the effects of running over all algorithms. In the current paper we present results from both types of analyses but concentrate largely on the first approach. The reader is referred to the companion paper [5] for more types of analysis involving the second approach.

We begin in Section II by introducing the necessary notation. Also discussed in this section is the model of computation we adopt, its limitations, and the reasons we chose it.

One might expect that there are pairs of search algorithms A and B such that A performs better than B on average, even if B sometimes outperforms A. As an example, one might expect that hill climbing usually outperforms hill descending if one's goal is to find a maximum of the cost function. One might also expect it would outperform a random search in such a context.

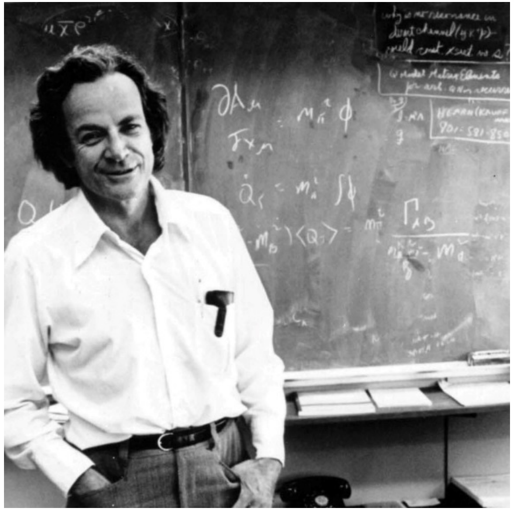
One of the main results of this paper is that such expectations are incorrect. We prove two "no free lunch" (NFL) theorems in Section III that demonstrate this and more generally illuminate the connection between algorithms and problems. Roughly speaking, we show that for both static and time-dependent optimization problems, the average performance of any pair of algorithms across all possible problems is identical. This means in particular that if some algorithm a_1's performance is superior to that of another algorithm a_2 over some set of optimization problems, then the reverse must be true over the set of all other optimization problems. (The reader is urged to read this section carefully for a precise statement of these theorems.) This is true even if one of the algorithms is random; any algorithm a_1 performs worse than randomly just as readily (over the set of all optimization problems) as it performs better than randomly. Possible objections to these

Unit 03. 코드 짜기

9가지 이야기

- 파인만 알고리즘
- 컴퓨터는 계산기다
- 잘게 쪼개기
- 디버깅 열심히
- 보기 좋은 코드가 돌리기도 좋다
- 효율적인 코드
- 너의 고민, 이미 누군가는 했다1
- 너의 고민, 이미 누군가는 했다2
- 일반적으로 짜자

파인만 알고리즘



- 천재 물리학자라 불리었던 리처드 파인만(Richard Phillips Feynman, 1918~1988)이 문제를 풀 때 사용하였다는 알고리즘
  1. Write down the problem. (문제를 쓴다.)
  2. Think real hard. (열심히 생각한다.)
  3. Write down the Solution. (답을 쓴다.)

컴퓨터는 계산기다

- 컴퓨터는 대신 생각해주지 않는다.

잘게 쪼개기

- 최대한 잘게 기능을 나누자

디버깅 열심히

- 알던 함수라도 다시 한번

보기 좋은 코드가 돌리기도 좋다

- 컴퓨터가 이해하는 코드는 어느 바보나 짤 수 있다. 좋은 프로그래머는 사람이 이해하는 코드를 짠다.

Martin Fowler (리팩토링 저자)

- 보기에 깔끔한 코드여야 나중에 고치기도 쉽고, 재탕하기에도 좋다.
- 컴퓨터과학에서 중요한 것은 단 두 가지 뿐이다. 캐시 무효화와 이름 작명이다.

Phil Karlton

- 변수명, 함수명도 이해하기 쉽게
- 구조도 알아보기 쉽도록

효율적인 코드짜기

- 앞에서 했던 내용 중에, 같은 기능이지만 함수마다 속도가 달라지던 함수들이 존재
- 데이터 셋이 어느정도 커지면, 속도 이슈를 고려 안 할 수가 없어진다.
- 간단한 연산을 1,000,000 ~ 10,000,000번 정도 하게되면, 느려지는 것에 체감이 오기 시작한다.(R)
- 함수에 따라 C로 짜여진 함수가 있고, R로 짜여진 함수가 있다. 태생적 차이가 존재

너의 고민, 이미 누군가는 했다 1

- 코딩을 하다보면, 이런 함수 있지 않을까 생각이 들때가 있다.
- 있다. 찾아보자.
- 공식 라이브러리로 등재된 함수들은 충분히 효율적이게 짜진 함수들이다. 믿자.

너의 고민, 이미 누군가는 했다 2

- 당신이 띄운 에러, 이미 누군가가 stackoverflow에 질문을 올렸고, 그에 대한 답이 달렸다.
- 뜬 에러를 그대로 복사 + 웹 주소창 + 붙여넣기 + 엔터를 누르면 글이 곧바로 보일 것이다.
- 에러에 대해 잠깐 고민해보는거는 좋지만, 너무 오래 고민하지 말자, 힘들다.

일반적으로 짜자

- 코드가 익숙해진 당신, 약간의 여유를 부려보자
- 함수를 짤 때, 조금 더 범용적인 함수를 짜보자
  - 입력되는 자료형 확장, 수치에서 벡터, 매트릭스로 확장
  - 옵션에 따라 변하는 알고리즘을 짜보자

일반적으로 짜자

- 코드가 익숙해진 당신, 약간의 여유를 부려보자
- 함수를 짤 때, 조금 더 범용적인 함수를 짜보자
  - 입력되는 자료형 확장, 수치에서 벡터, 매트릭스로 확장
  - 옵션에 따라 변하는 알고리즘을 짜보자

Q &A

끝!

들어주셔서 감사합니다.