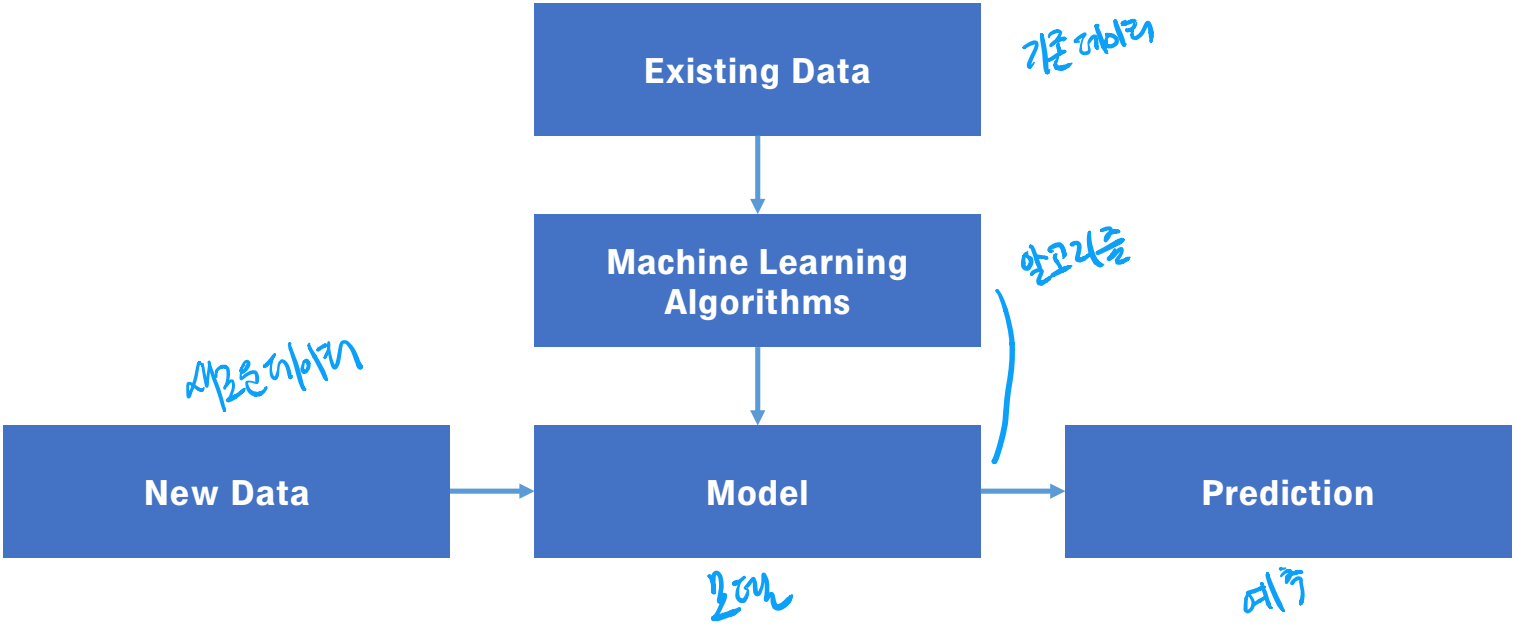
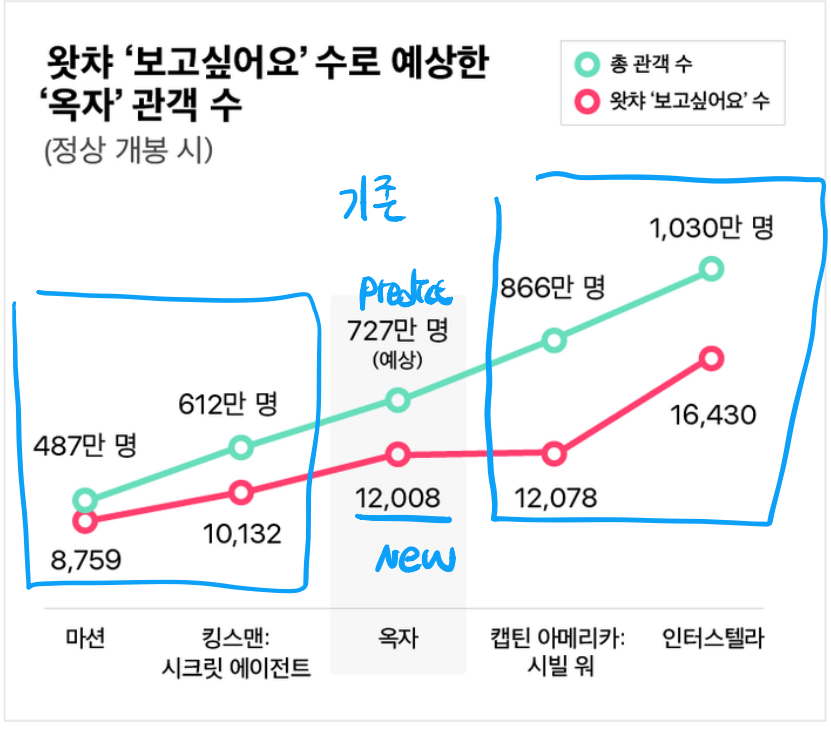
How to learn machine learning

Machine Learning Overview

Director of TEAMLAB
Sungchul Choi

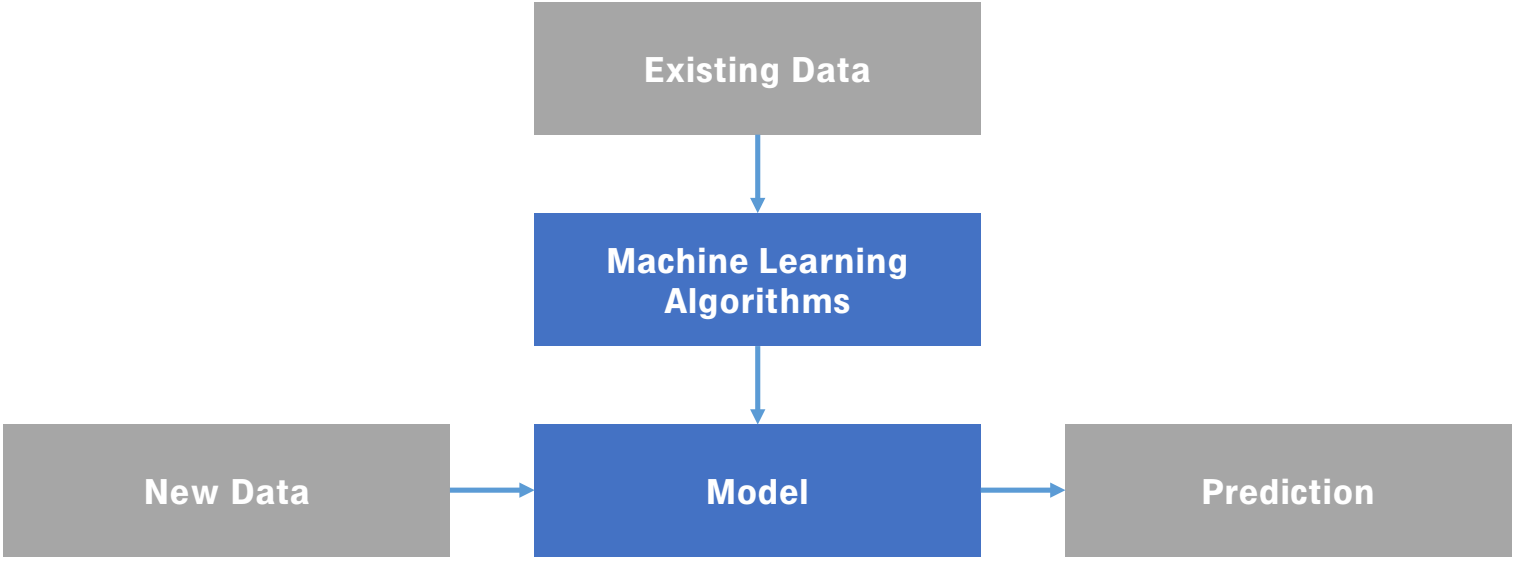
Machine Learning Process

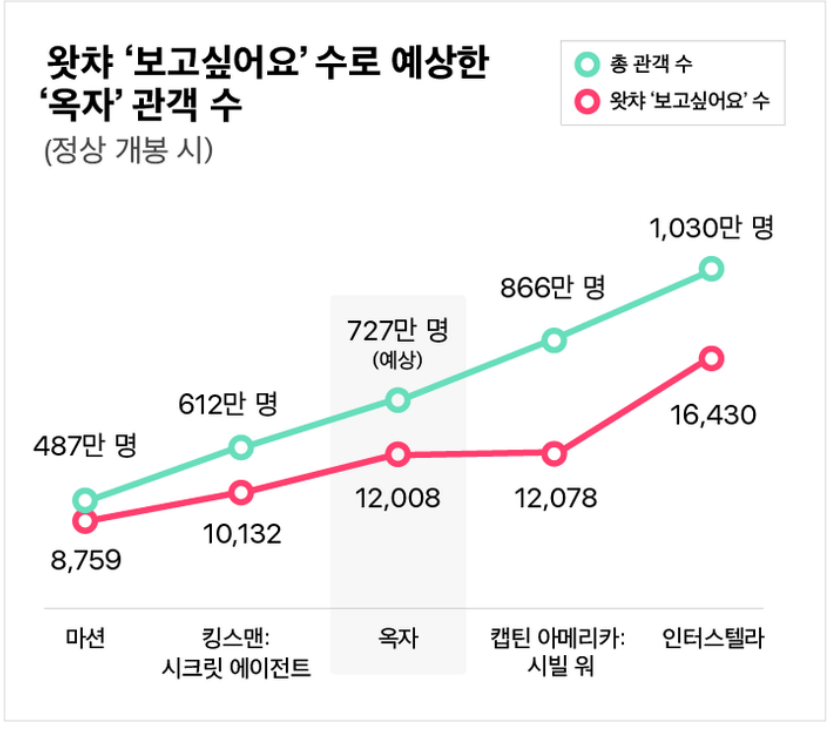




Source: http://platum.kr/archives/83757

Machine Learning Process





Source: http://platum.kr/archives/83757

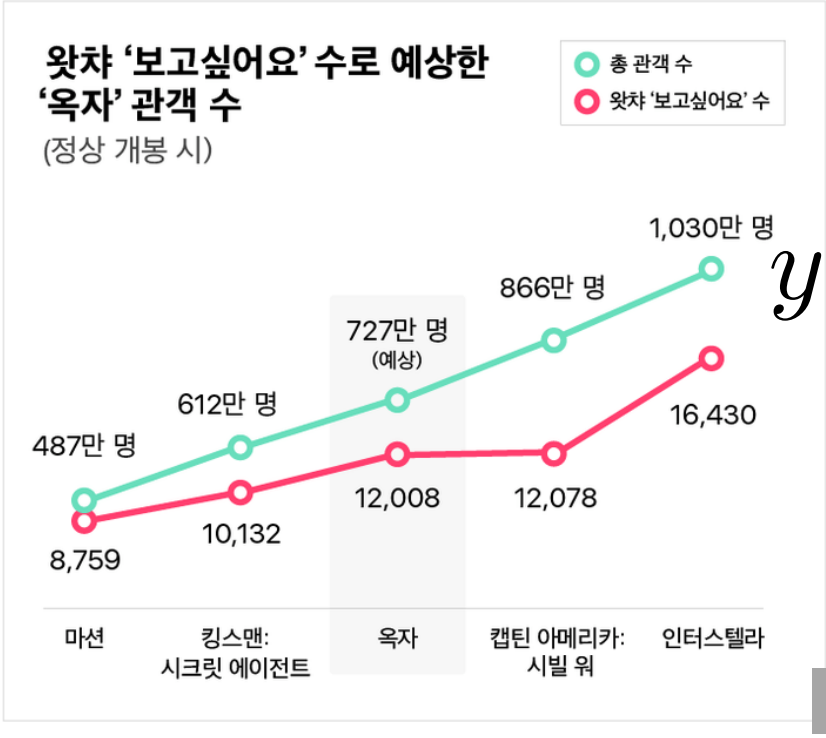
Key concepts

Model - 예측을 위한 수학 공식, 함수

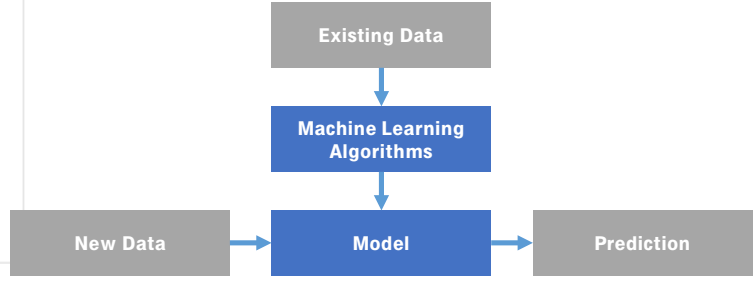
1차 방정식, 확률분포, condition rule

Algorithms - 어떠한 문제를 풀기 위한 과정

Model을 생성하기 위한 (훈련) 과정


왓챠 '보고싶어요' 수로 예상한
'옥자' 관객 수
(정상 개봉 시)
총 관객 수
왓챠 '보고싶어요' 수
1,030만 명
866만 명
727만 명
(예상)
612만 명
487만 명
16,430
12,008
12,078
10,132
8,759
마션
킹스맨:
시크릿 에이전트
옥자
캡틴 아메리카:
시빌 워
인터스텔라

$$y = ax + b$$


Existing Data
Machine Learning
Algorithms
New Data
Model
Prediction

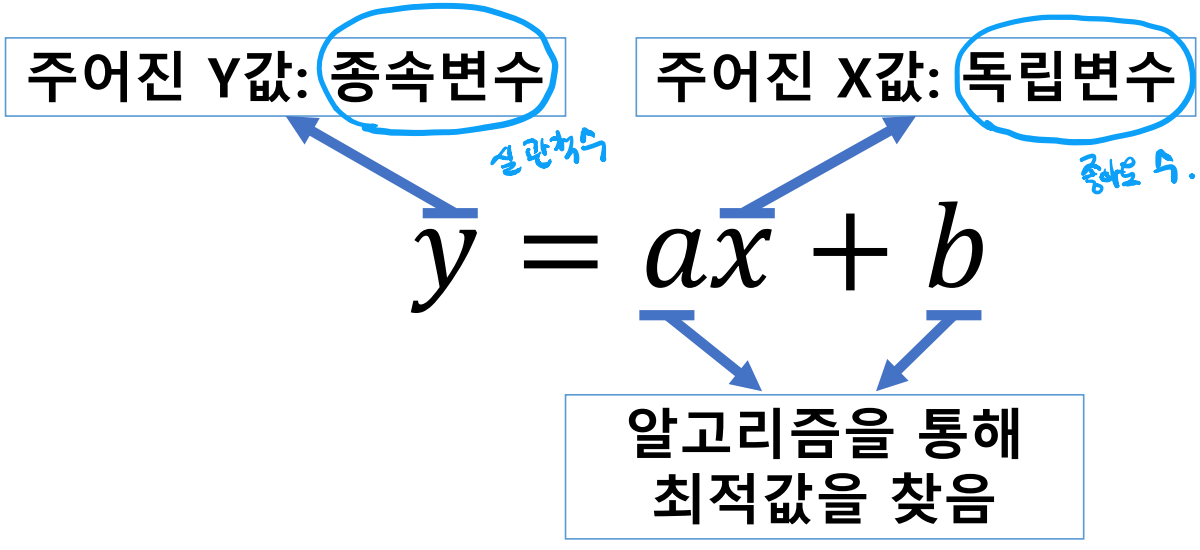
Human knowledge belongs to the world.

The concept of a feature

An understanding of data

Director of TEAMLAB
Sungchul Choi

모델을 학습할 때 영향을 주는 것들


주어진 Y값: 종속변수
주어진 X값: 독립변수
실관찰수
종속도 수.
$y = ax + b$
알고리즘을 통해
최적값을 찾음

Y값에 영향을 주는 X값은 하나인가?


Y값에
주는
X값은
인가?

Boston House Price Dataset

- 머신 러닝 등 데이터 분석을 처음 배울 때,
  가장 대표적으로 사용하는 Example Dataset

- 1978년에 발표된 데이터로, 미국 인구통계 조사 결과
  미국 보스턴 지역의 주택 가격에 영향 요소들을 정리함

http://lib.stat.cmu.edu/datasets/boston

Boston House Price Dataset

$y = ax + b$.

X 변수 13개

Y 변수

[01] CRIM	자치시(town) 별 1인당 범죄율
[02] ZN	25,000 평방피트를 초과하는 거주지역의 비율
[03] INDUS	비소매상업지역이 점유하고 있는 토지의 비율
[04] CHAS	찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0)
[05] NOX	10ppm 당 농축 일산화질소
[06] RM	주택 1가구당 평균 방의 개수
[07] AGE	1940년 이전에 건축된 소유주택의 비율
[08] DIS	5개의 보스턴 직업센터까지의 접근성 지수
[09] RAD	방사형 도로까지의 접근성 지수
[10] TAX	10,000 달러 당 재산세율
[11] PTRATIO	자치시(town)별 학생/교사 비율
[12] B	1000(Bk-0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함.
[13] LSTAT	모집단의 하위계층의 비율(%)
[14] MEDV	본인 소유의 주택가격(중앙값) (단위: $1,000)

http://www.dator.co.kr/ctg258/textyle/1721307

http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \cdots \beta_{13} x_{13} + \beta_0 \cdot 1$$

13개의 x변수, 1개의 y변수

X변수의 실제 데이터는 특징(feature)을 나타냄

독립변수

Input 변수.

Feature

- 머신러닝에서 데이터의 특징을 나타내는 변수
- feature, 독립변수, input 변수 등은 동일의미로 사용
- 일반적으로 Table 상에 Data를 표현할 때, Column을 의미
- 하나의 data instance (실제 데이터)는 feature vector로 표현

Feature

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	0
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	1
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	1
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2	1
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7	0
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9	0
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1	0
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5	0
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9	0
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15	0
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9	0
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7	0
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4	0
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2	0

Feature

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0

$$y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + w_7x_7 + \cdots w_{13}x_{13} + w_0 \cdot 1$$

Feature vector

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} 1 \\ 0.00632 \\ 18 \\ 2.31 \\ 0.538 \\ \vdots \\ 24 \end{bmatrix} \qquad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{13} \end{bmatrix}$$

※ Scalar는 이탤릭체, vector는 소문자 볼드, matrix는 대문자 볼드

Feature

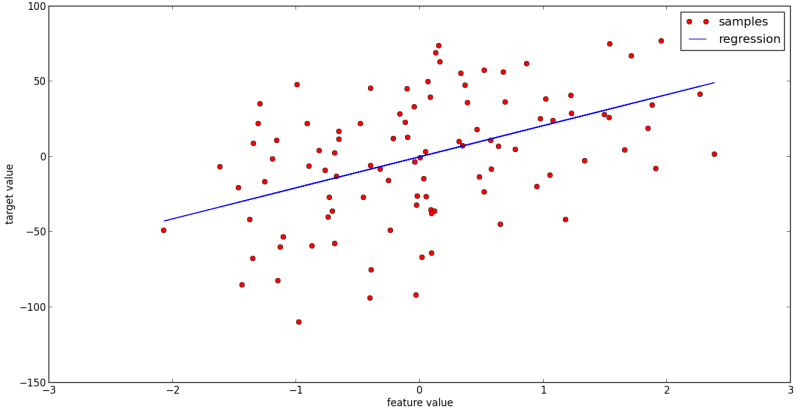
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0

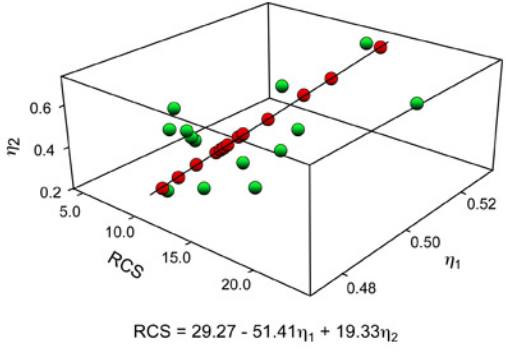
$$y = w_1 x_1 + w_2 x_2 \ldots w_{13} x_{13} + w_0 x_0$$

$$= \sum_{i=0}^{13} w_i x_i = \mathbf{w}^T \cdot \mathbf{x}$$

Feature의 개수?

Feature가 1개 일 때, Feature가 2개 일 때

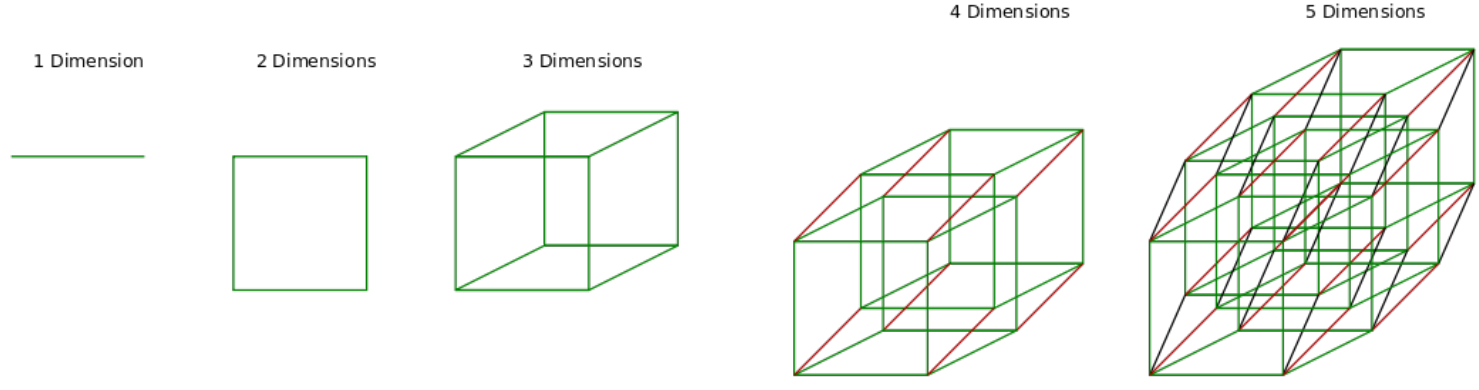




$$RCS = 29.27 - 51.41\eta_1 + 19.33\eta_2$$

https://goo.gl/d1zRGq

Feature가 n개 일 때?



Curse of dimensionality

https://goo.gl/mCg5nu

차원의 저주(curse of dimensionality)

- 데이터의 차원이 증가할 수록(= feature가 증가할 수록)
  데이터를 표현하는 공간이 증가하기 때문에
  1) 희박한 벡터가 증가 (값이 없는 feature가 늘어남)
  2) 샘플데이터가 급속도록 늘어남

- 데이터 분포나 모델 추정의 어려움이 생김

Human knowledge belongs to the world.

Data attributes

An understanding of data

Director of TEAMLAB
Sungchul Choi

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0

$$y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + w_7x_7 + \cdots w_{13}x_{13} + w_0 \cdot 1$$

x_i에는 어떤 종류의 값들이 들어갈까?

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0

$$y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + w_7x_7 + \cdots w_{13}x_{13} + w_0 \cdot 1$$

Feature별로 Data의 유형이 다름

DB를 알면 굳이 몰라도 되는 내용...

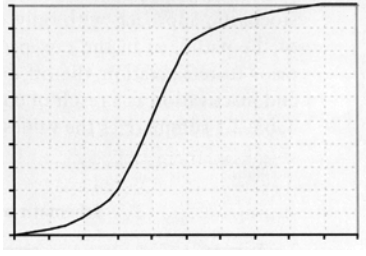
연속형 값 vs 이산형 값

continuous

값이 끊어지지 않고 연결됨

온도, 시험평균 점수, 속도

일반적으로 실수 값들

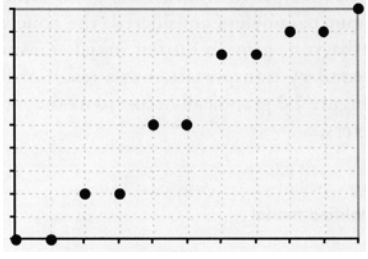


discrete

값이 연속적이지 않음

성별, 우편주소, 등수

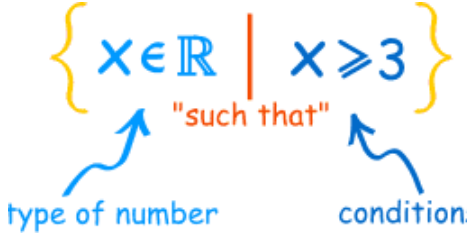
Label로 구분되는 값들



https://goo.gl/1sSRSV

Numeric Types

- 정량적으로 측정 가능한 data type
- 일반적으로 정수(integer) 또는 실수(real-number)로 표현
- 온도, 자동차 속도, 날짜의 차이(year or day)
- 단위(scale)이 있는 Interval-scaled type
- 비율이 있는 Ratio-scaled type



https://goo.gl/S7W69t

Nominal Types

- 범주(category)로 분류가 가능한 data type
- 명목 척도라는 표현으로 사용되기도 함
- 색깔, 학교명, ID, 전공명 등
- 두 개의 Category만 분류할 때는 Binary Type으로 구별

Ordinal Types

- 범주(category)로 분류가 가능하나 범주간의 순서가 있음
- 명목 척도라는 표현으로 사용되기도 함
- 음료수 병의 크기, 학점, 5점 척도 설문조사
- 측정되는 Scale 또는 Unit이 사람마다 다를 수 있음
- 순서가 있는 것 ≠ 배수로 증가하는 개념은 다름

실제 값을 넣어보면...

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 + w_7 x_7 + \cdots w_{13} x_{13} + w_0 \cdot 1$$

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	0
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	1
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	1
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2	1
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7	0
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9	0
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1	0
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5	0
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9	0
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15	0
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9	0
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7	0
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4	0
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2	0

생길 수 있는 문제점들

- 데이터의 최대/최수가 다름 → Scale에 따른 y값에 영향
- Ordinary 또는 Nominal 한 값 들의 표현은 어떻게?
- 잘 못 기입된 값들에 대한 처리
- 값이 없을 경우는 어떻게?
- 극단적으로 큰 값 또는 작은 값들은 그대로 놔둬야 하는가?

Human knowledge belongs to the world.

Loading data with pandas

An understanding of data

Director of TEAMLAB
Sungchul Choi

우리의 데이터는 누가 처리한다?

컴퓨터...

그러려면 먼저 불러오기 부터

전에 먼저...

Data table, Sample

attribute, field, feature, column

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	0
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	1
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	1
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2	1
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7	0
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9	0
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1	0
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5	0
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9	0
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15	0
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9	0
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7	0
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4	0
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2	0

instance, tuple, row

Feature vector

data

데이터의 형식

- 일반적으로 데이터분석시 사용하는 Raw data는 Binary가 아닌 text 형태의 데이터

- 주로 사용되는 데이터 포멧은 csv, json, xml 등

- pandas를 사용하여 데이터를 호출함

Padas
엑셀처럼 데이터 사용

Pandas

- 구조화된 데이터의 처리를 지원하는 Python 라이브러리
- 고성능 Array 계산 라이브러리인 Numpy와 통합하여, 강력한 “스프레드시트” 처리 기능을 제공
- 인덱싱, 연산용 함수, 전처리 함수 등을 제공함

Pandas 설치

```
conda create -n ml_scratch python=3.6 # 가상환경생성
activate ml_scratch # 가상환경실행
conda install pandas#  pandas 설치
```

```
jupyter notebook # 주피터 실행하기
```

데이터 로딩

In [1]:
```
import pandas as pd #라이브러리 호출
```

In [2]:
```
data_url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data' #Data URL
df_data = pd.read_csv(data_url, sep='\s+', header = None) #csv 타입 데이터 로드, separate는 빈공간으로 지정하고, Column은 없음
```

In [3]:
```
df_data.head() #처음 다섯줄 출력
```

Out[3]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

Column 지정

In [4]:
```
df_data.columns = ['CRIM','ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO' ,'B', 'LSTAT', 'MEDV']
# Column Header 이름 지정
df_data.head()
```

Out[4]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	0
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	1
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	1
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2	1
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7	0
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9	0
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1	0
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5	0
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9	0
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15	0
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9	0
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7	0
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4	0
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2	0

Human knowledge belongs to the world.