The Apprentice Project

Lec05: Linear Regression (Part I)

충북대학교

문성태 (지능로봇공학과)

stmoon@cbnu.ac.kr

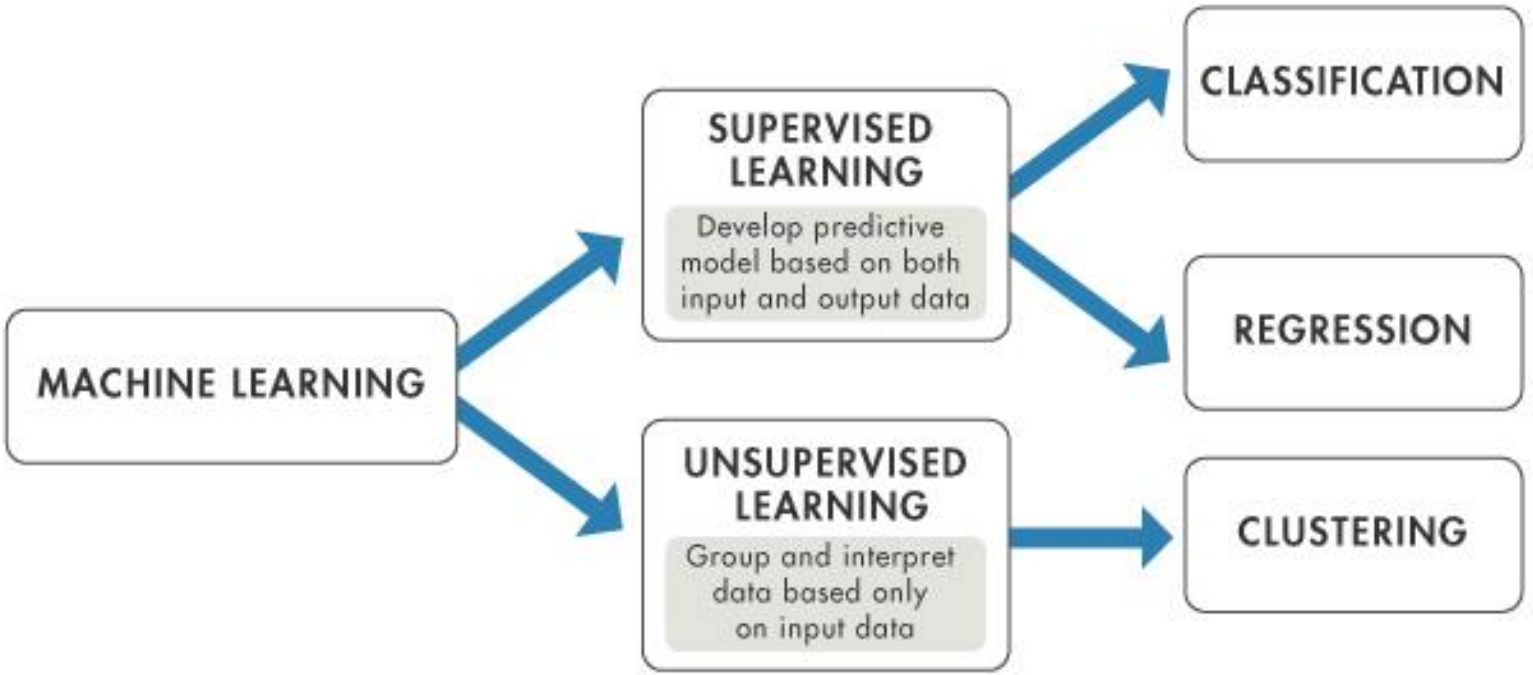
01

Introduction

Recap: Dataset

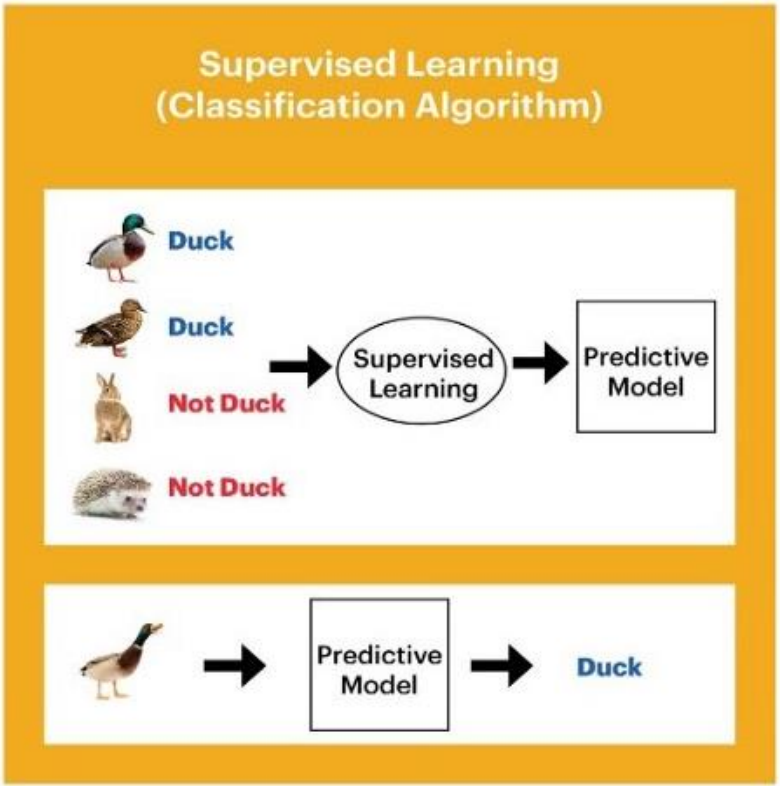
- Dataset
  - Training set
  - Test set
- Pre-processing
  - Normalization & Standardization
  - PCA (Principal Component Analysis

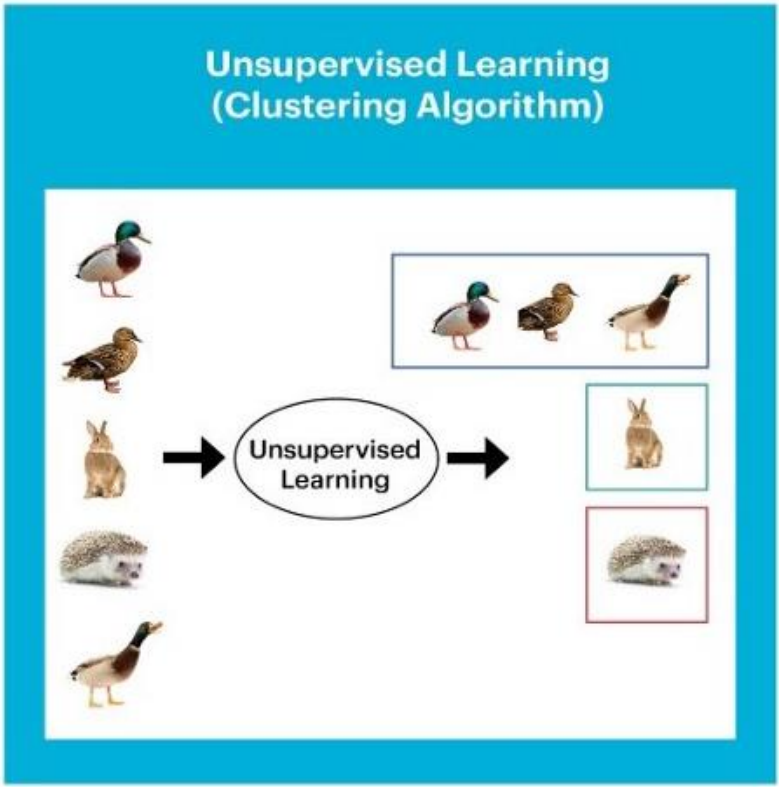
Recap: Machine Learning



https://images.deepai.org

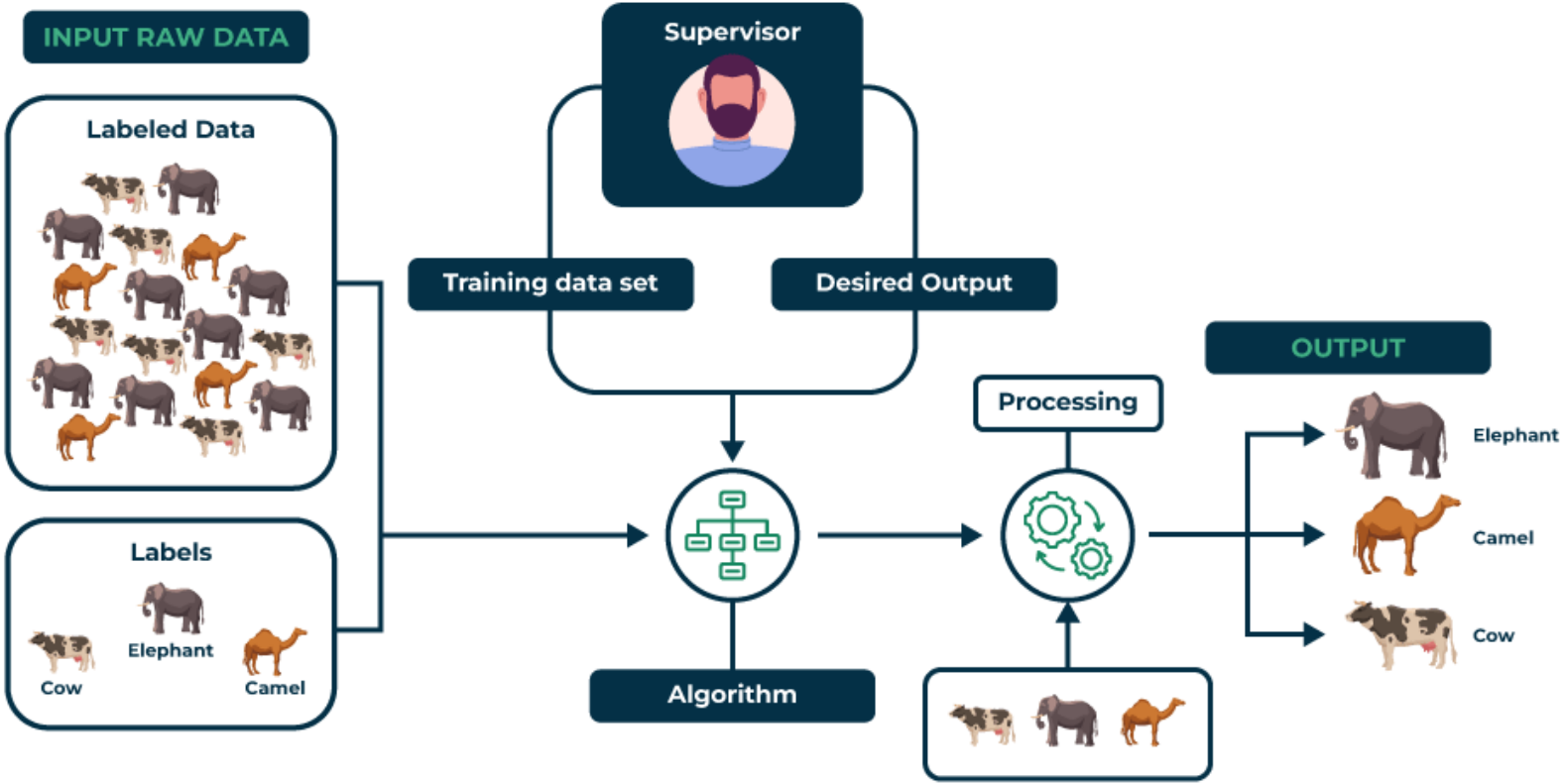
Recap: Machine Learning





https://images.deepai.org

What is Supervised Learning

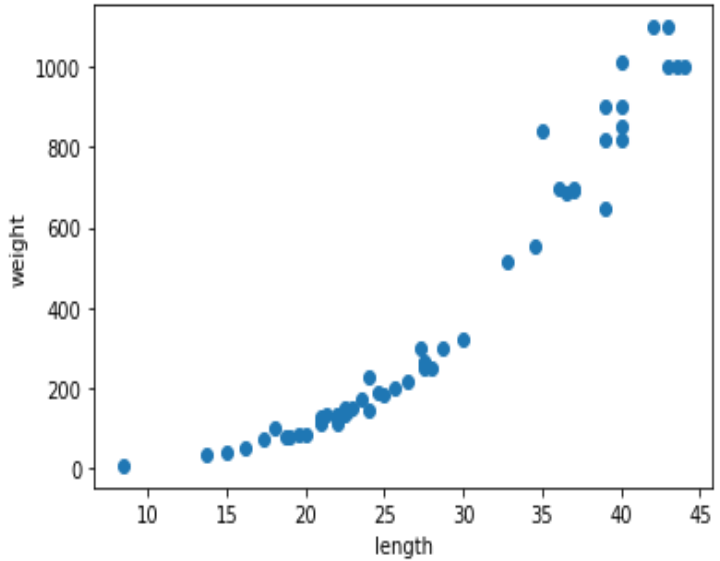


Supervised Learning

Input	Output	Application
Email	Spam? (0/1)	Spam filtering
Audio	Text transcripts	Speech recognition
English	Spanish	Machine translation
User info	Click? (0/1)	Online advertising
Image, lidar	Position of others cars	Self-driving car
Image of phone	Defect? (0/1)	Visual inspection

Supervised Learning

- Classification
  - Sample을 몇 개의 Class로 분류하는 방식
- Regression
  - 임의의 어떤 숫자를 예측하는 방식

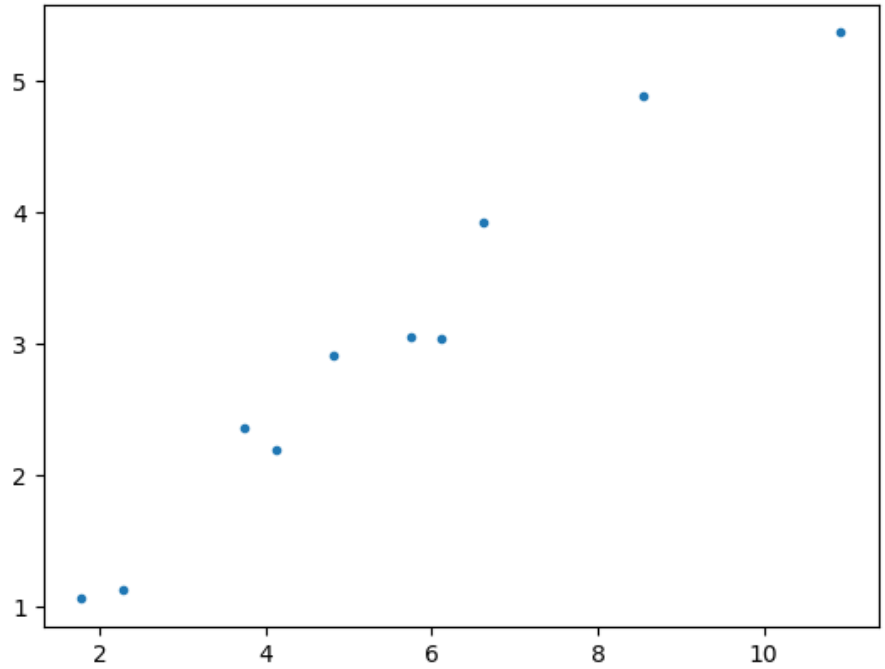


02

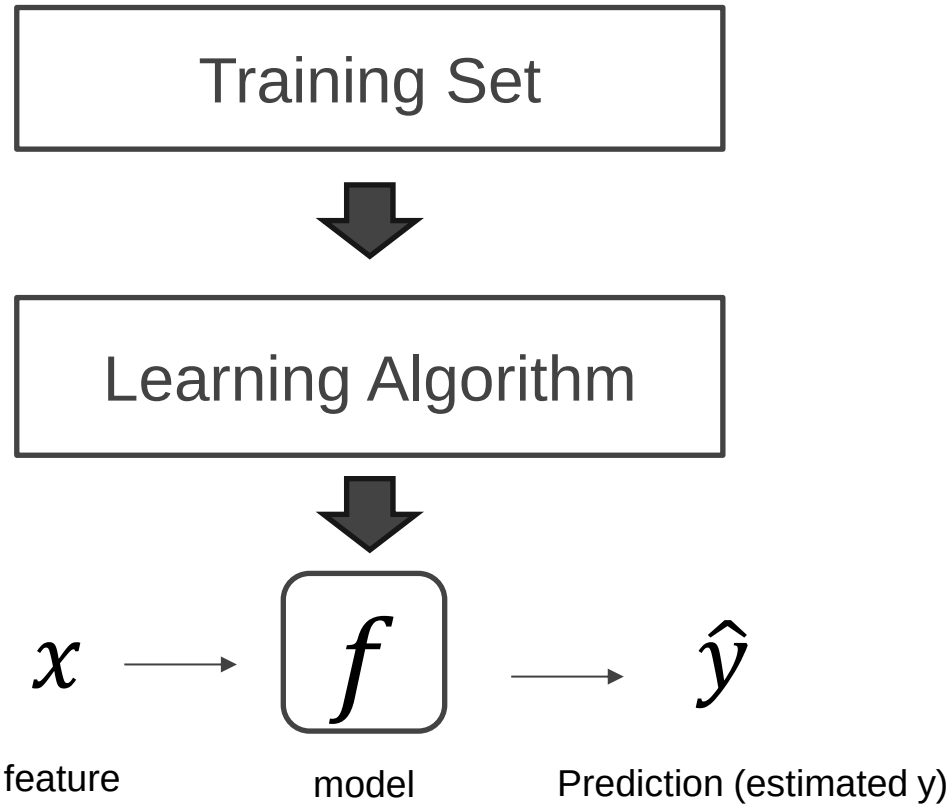
Linear Regression with univariate

Target

- 아래 그래프에서 학습 데이터 (x,y)를 잘 설명하는 (회귀) 직선을 구하세요



How to work


Training Set
Learning Algorithm
x
f
$\hat{y}$
feature
model
Prediction (estimated y)

Linear Regression

How to find regression

- 목적함수 (Objective Function)
  - 목적함수는 최소값/최대값을 찾는 **최적화 문제** (Optimization Problem)에서 사용하는 함수

Objective Function vs Cost Function vs Loss Function

- Objective Function: 어떠한 목적을 가지고 모델을 학습해 최적화하고자 하는 함수
- Cost Function: 입력으로 받은 데이터를 모아서 오차를 계산하는 함수
- Loss Function: 입력으로 받은 데이터를 하나하나 받아 실제값과 예측값 간의 오차를 계산하는 방식

Objective Function >= Cost Function >= Loss Function

How to calculate the objective function

- Gradient Descent
  - 목적함수의 값을 최소화하기 위해 경사를 내려가듯 최소값을 찾는 방법
  - Iterative optimization algorithm for finding the minimum of a function

https://towardsdatascience.com/understanding-gradient-descent-and-adam-optimization-472ae8a78c10

03

Linear Regression with Multivariate

Target: Boston House Prices

- The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970
- Number of Instances: 506
- Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.
  - CRIM: per capita crime rate by town
  - ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
  - INDUS proportion of non-retail business acres per town
  - CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  - NOX nitric oxides concentration (parts per 10 million)
  - RM average number of rooms per dwelling
  - AGE proportion of owner-occupied units built prior to 1940
  - DIS weighted distances to five Boston employment centres
  - RAD index of accessibility to radial highways
  - TAX full-value property-tax rate per $10,000
  - PTRATIO pupil-teacher ratio by town
  - B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  - LSTAT % lower status of the population
  - MEDV Median value of owner-occupied homes in $1000's (TARGET)

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

Boston House Prices

index	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	TARGET
0	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
1	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
2	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
5	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
6	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
7	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
8	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
9	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
10	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
11	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
12	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
13	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
14	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
15	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
16	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
17	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
18	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2
19	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2

Linear Regression

Linear Regression