

머신러닝 강의 02

머신러닝 개요

머신러닝이란

Contents

스칼라

벡터

행렬

머신러닝

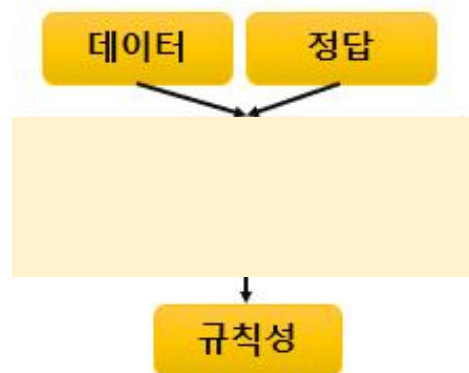
- 머신러닝은 컴퓨터 라는 도구로 경험을 활용해 시스템 자체를 개선해 나가는 방법을 연구하는 학문
- 컴퓨터 시스템에서 일반적으로 경험은 데이터라는 형식으로 존재하고, 따라서 머신러닝은 데이터로부터 규칙성을 발견해 나가는 알고리즘이라 할 수 있다.



전통적인 알고리즘
문제를 해결하는 Instruction의 sequence.



머신 러닝
Data-Driven의 알고리즘



인공지능, 머신러닝, 딥러닝

| 인공지능 (Artificial general intelligence)

- 약 인공지능 (Artificial general intelligence)
 - ✓ 지능을 모방해서 특정한 문제를 풀기위한 기술
- 강 인공지능 (Artificial general intelligence)
 - ✓ 인간처럼 생각하고 감정과 의식을 가지며 창의성을 발휘하는 기계



<https://sdc-james.gitbook.io/onebook/1.1.-artificial-intelligence/1.4.-greater-than-greater-than>

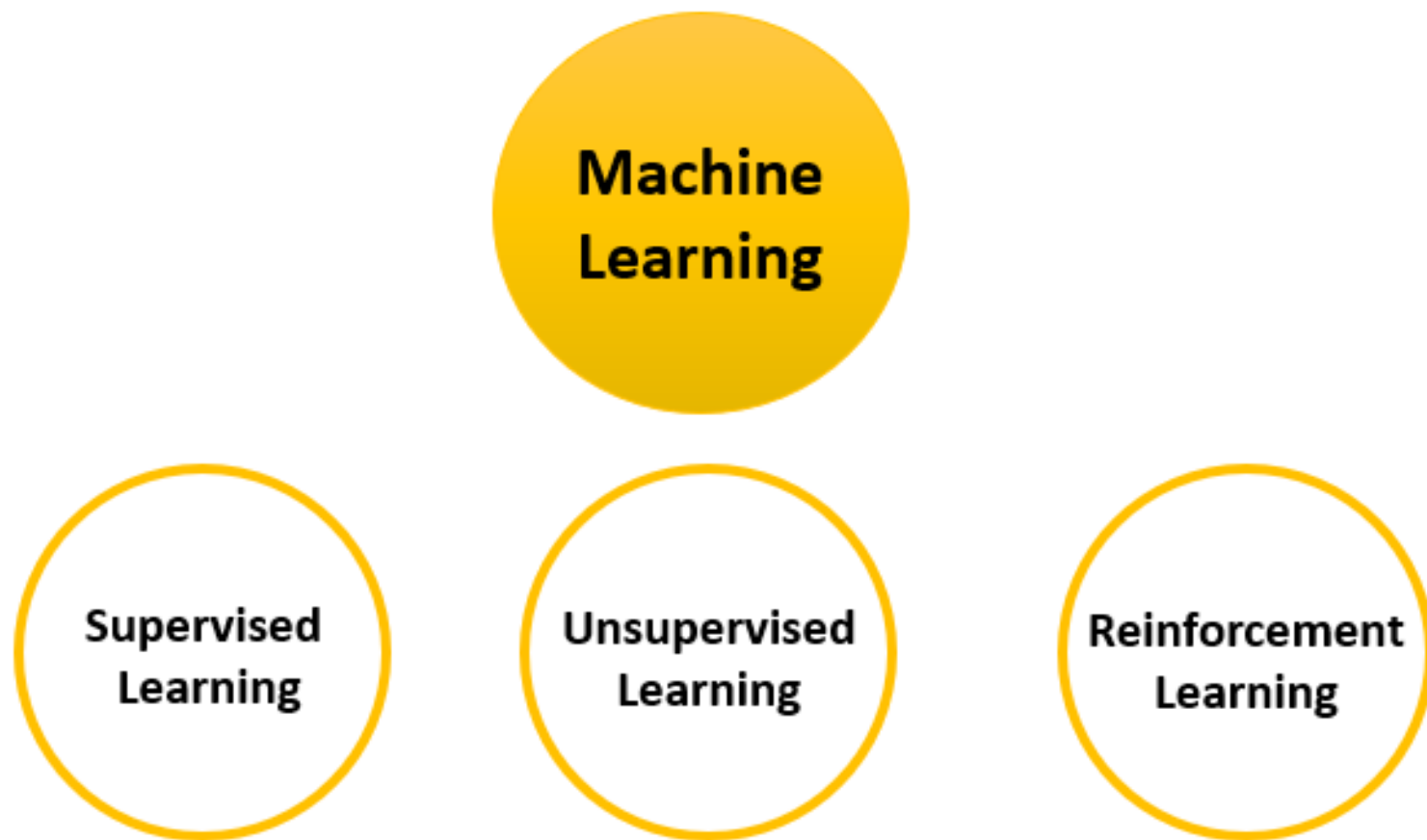
| 머신러닝 (Machine learning)

- 컴퓨터 스스로 데이터에 대한 규칙을 학습하는 알고리즘에 관한 학문

| 딥러닝 (Deep learning)

- 여러가지 머신러닝 알고리즘 중 인간의 신경망을 본 딴 인공신경망에서 발전한 개념
- 신경망은 큰 연산을 필요로 하기 때문에 초기에는 딥러닝의 상용화가 어려웠으나 2012년 이후 하드웨어 성능의 폭발적 향상과 알고리즘 발전 등에 따라 획기적으로 발전

머신러닝의 종류



문제의 종류

| 회귀 (Regression)

- 연속적인 값의 범위 내에서 예측해야 하는 문제

X

국어	수학	영어
80	90	95
45	50	60
30	70	90
85	80	85

Y

환산
86
55.5
70
83

| 이진분류 (Binary Classification)

- 주어진 입력에 대해 두개 선택지 중 하나를 선택해야 하는 문제

X

국어	수학	영어
80	90	95
45	50	60
30	70	90
85	80	85

Y

합불
합격
불합격
합격
합격

| 다중 클래스 분류 (Multi-class Classification)

- 주어진 입력에 대해 세 개 이상의 선택지 중 하나를 선택해야 하는 문제

X

국어	수학	영어
80	90	95
45	50	60
30	70	90
85	80	85

Y

학점
A
C
B
A

데이터 세트 (Dataset)

feature1	feature2	feature3
80	90	95
45	50	60
30	70	90
85	80	85

샘플 1	80	90	95
------	----	----	----

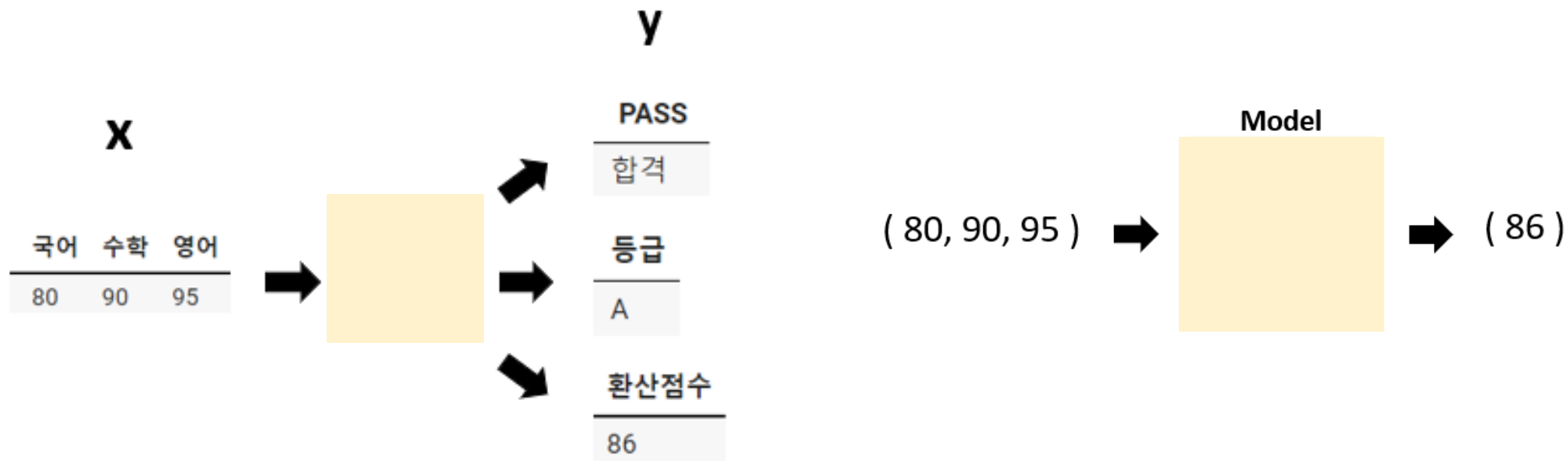
- D
- $x_i^{(j)}$ (x_{ij})
- m
- d

- 위 데이터에서 x_2 에 해당하는 것은?
- 위 데이터에서 x_{31} 에 해당하는 것은?
- 위 데이터에서 $x_{1,*}$ 에 해당하는 것을 쓰고, 크기를 나타내세요.
- 위 데이터에서 $x_{*,3}$ 에 해당하는 것을 쓰고, 크기를 나타내세요.
- 위 데이터에서 m 에 해당하는 것은?
- 위 데이터에서 d 에 해당하는 것은?

모델 (Model)

| 머신러닝 모델

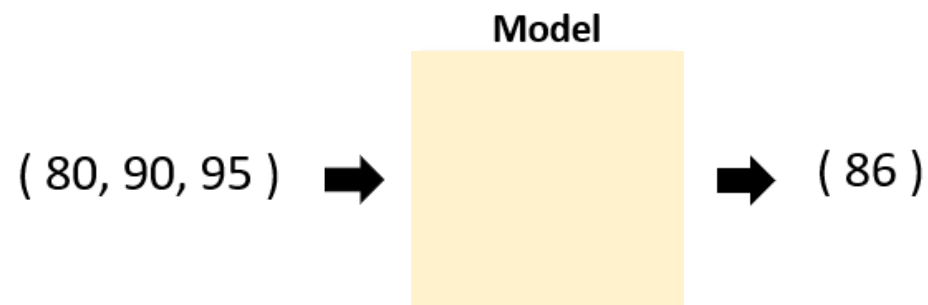
- 데이터 세트로부터 패턴을 찾거나 이를 근거로 결정을 내릴 수 있는 프로그램
- 이전에 접한 적 없는 새로운 데이터 세트에서도 적절한 결정을 내릴 수 있다.
- 주어진 데이터 X 를 가지고 예측값 Y 를 '잘' 예측하기 위해서는 X 와 Y 에 대한 이해가 필수적



학습 (learning)

| 학습(Learning)

- 적절한 함수의 파라미터를 찾아가는 과정



- w_1, w_2, w_3 이 0.5, 0.2, 0.3 일 때 계산
- w_1, w_2, w_3 이 1, 0.5, 0.5 일 때 계산

평가 (Evaluation)

| 평가 (Evaluation)

- 모델 성능을 평가하는 것

| 손실함수(Loss function)

- 실제값과 예측값의 차이를 측정하는 척도

| 데이터셋의 분류

- 훈련용 데이터 (Training set)
 - ✓ 모델 학습에 사용되는 데이터
- 검증용 데이터 (Validation set)
 - ✓ 학습 과정 중 학습이 잘 이루어 지는지 사용되는 데이터
- 테스트 데이터 (Test set)
 - ✓ 학습이 완료된 모델을 평가하는데 사용되는 데이터

훈련 (Training)

검증
(Validation)

테스트
(Testing)

차원 축소와 군집화

Contents

차원 축소와 군집화

샘플 공간

투영

매니폴드

차원 축소의 목적

차원 축소와 군집화

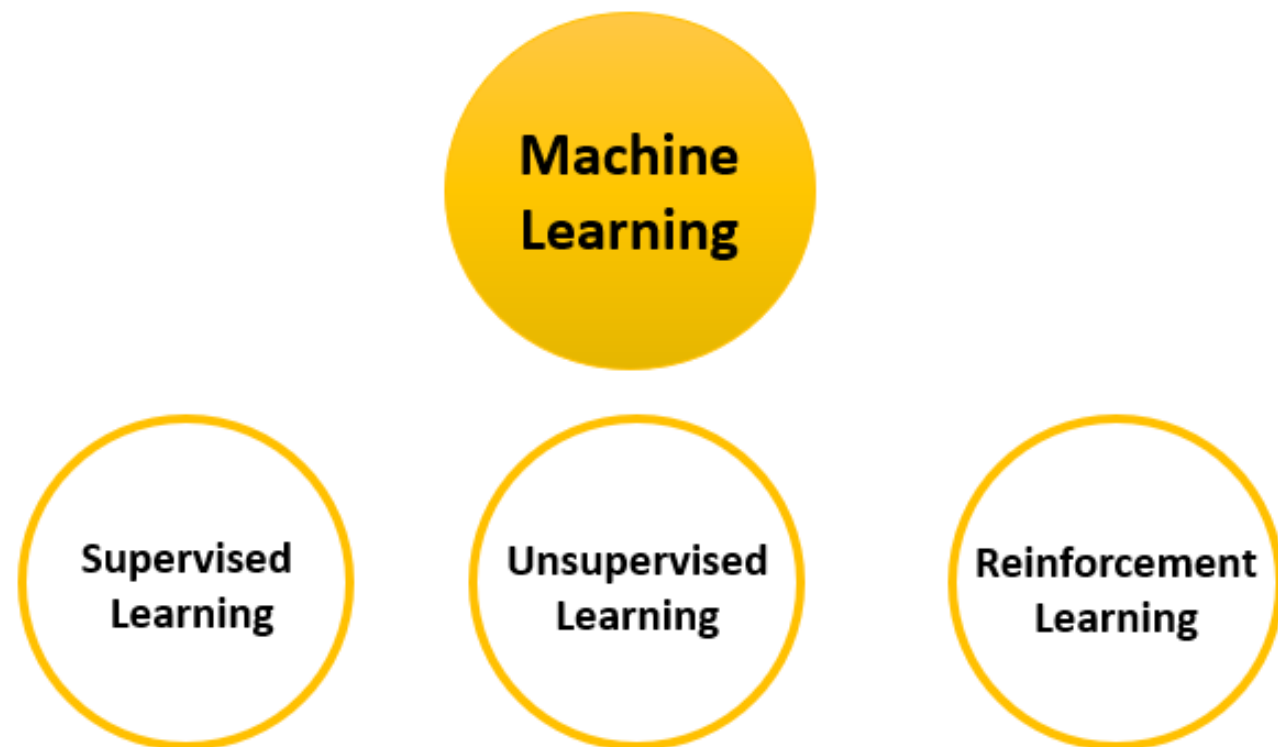
차원 축소와 군집화는 비지도학습 (Unsupervised learning)의 대표적인 예로 정답 데이터셋을 필요로 하지 않음

| 차원 축소 (Dimensionality Reduction)

- 수학적 변환을 통해 고차원의 속성 공간을 저차원의 부분공간으로 변환하는 기법

| 군집화 (Clustering)

- 주어진 데이터 집합을 유사한 데이터들의 그룹으로 나누는 것을 군집화(clustering) 이라 하며 군집화를 통해 나누어지는 유사한 데이터 그룹을 군집(cluster) 이라고 한다.



샘플 공간 (Sample space)

| 샘플 공간 (속성 공간, attribute space)

- 샘플이 존재할 수 있는 공간

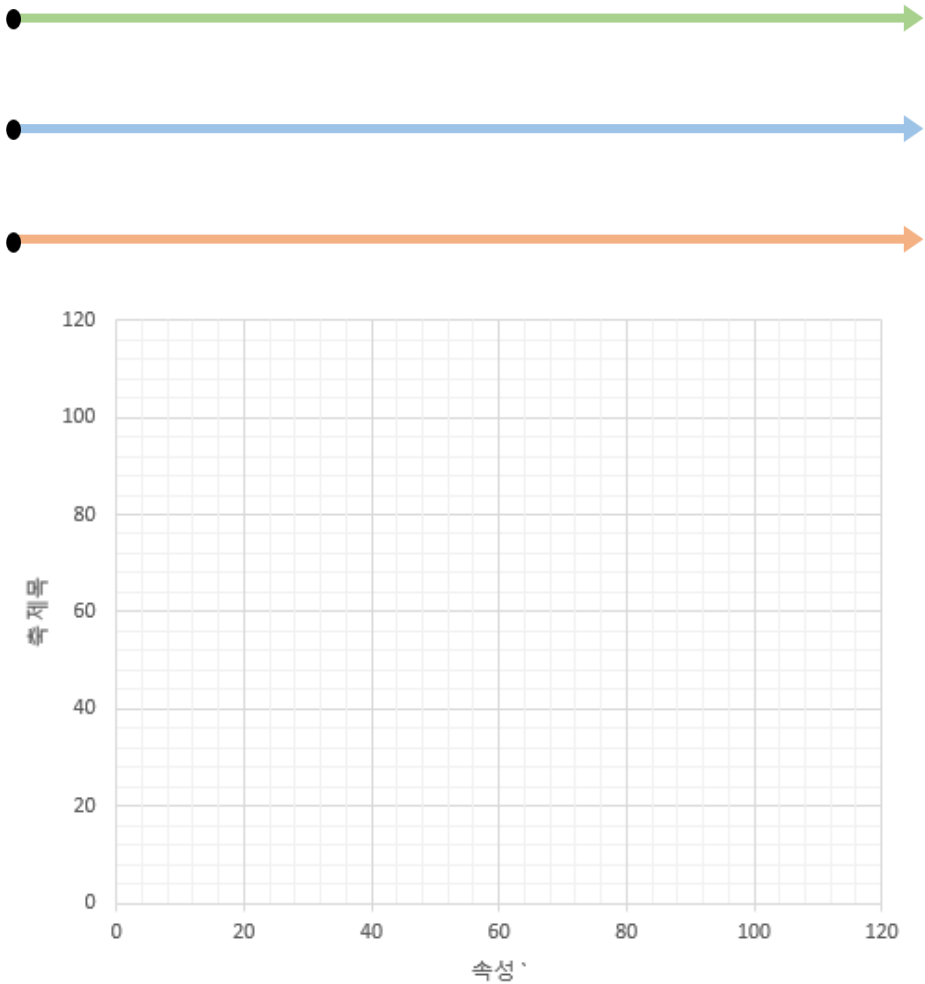
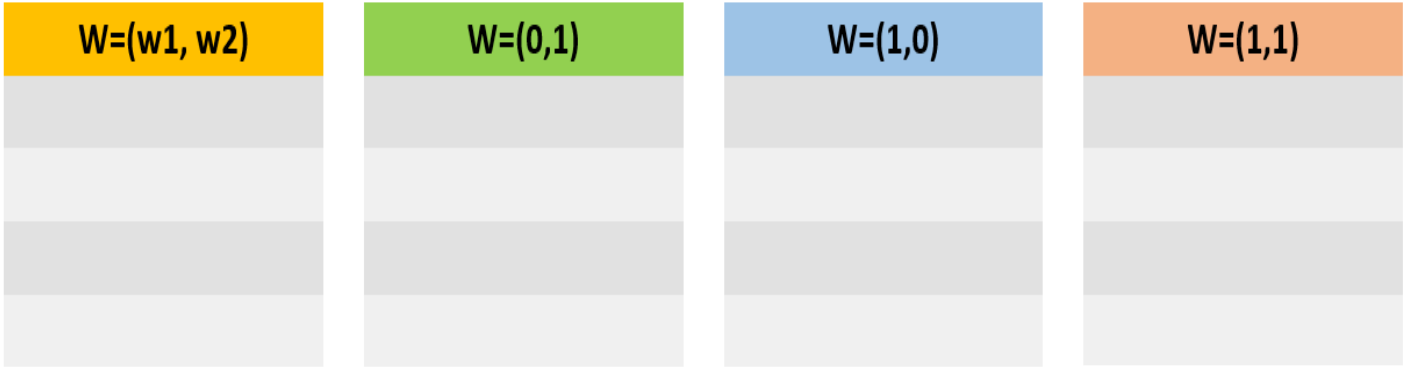
국어	수학	영어
100	90	90
45	50	60
60	70	65
85	80	90

- 데이터셋 D 의 shape?
- 샘플 한 개 x_i 의 shape?
- 샘플의 변수가 (국어) 일 때 샘플 공간 시각화
- 샘플의 변수가 (국어, 수학) 일 때 샘플 공간 시각화
- 샘플의 변수가 (국어, 수학, 영어) 일 때 샘플 공간 시각화

투영 (Projection)

| 투영

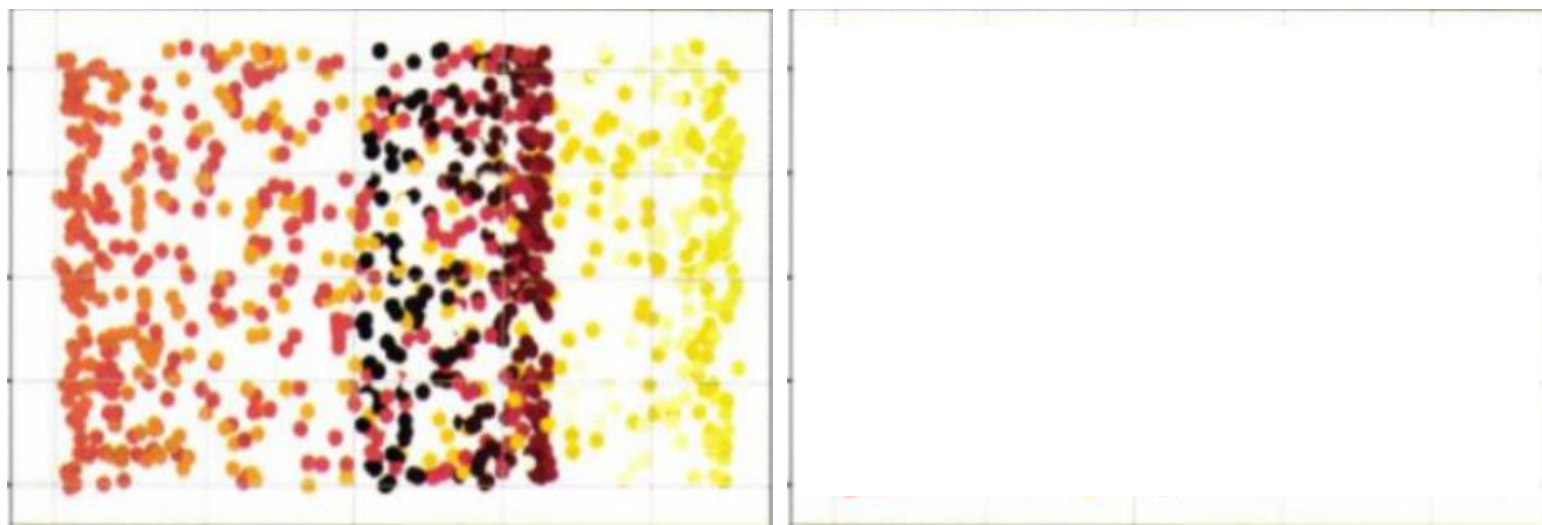
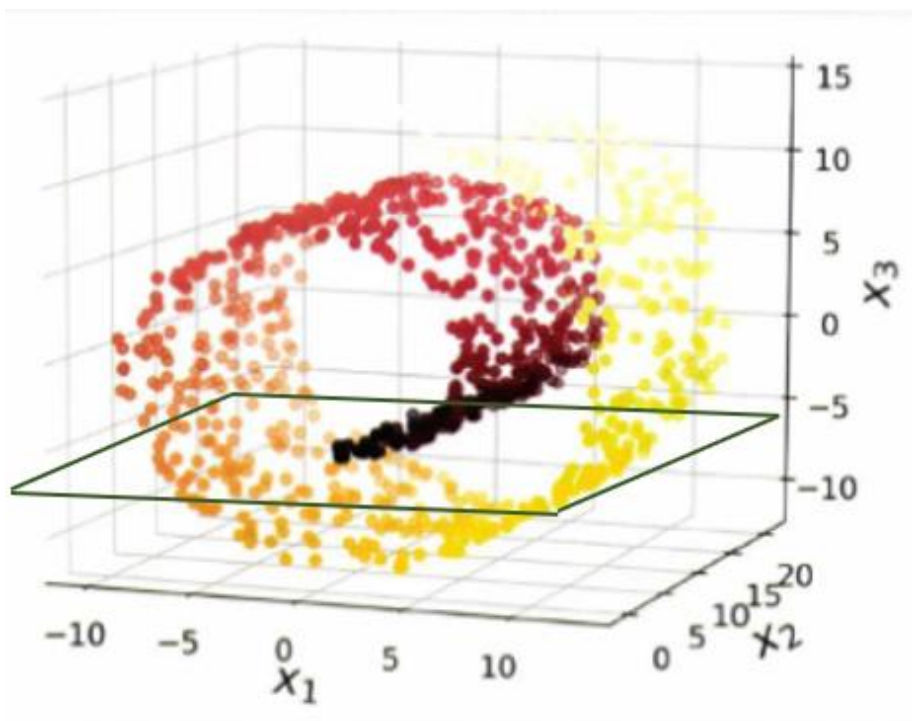
- 차원축소의 한가지 방법으로 고차원 공간 안에서 데이터의 양상을 최대한 보존하면서 차원을 축소시키는 기법



매니폴드 (Manifold)

| 매니폴드

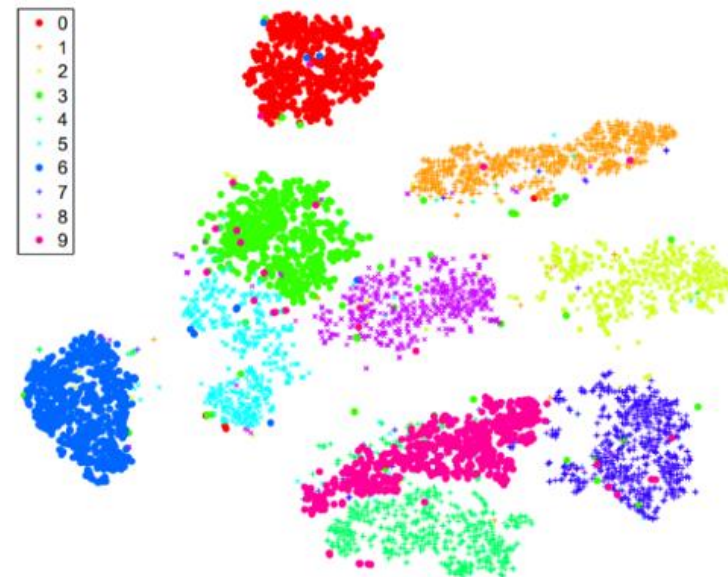
- 고차원의 데이터를 저차원으로 옮길 때 데이터를 잘 설명하는 집합의 모형
- 스위스 롤 데이터는 3차원 공간에서 휘어지거나 뒤틀린 모습으로 데이터가 존재하며 이런 경우 단순 투영을 이용하면 데이터가 더 복잡한 형태로 축소된다.



차원축소의 목적

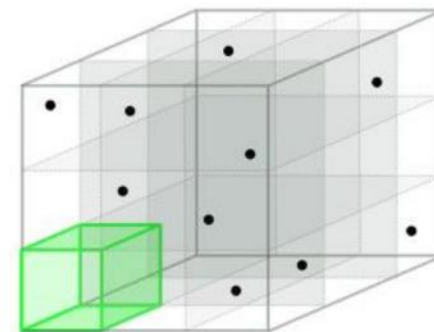
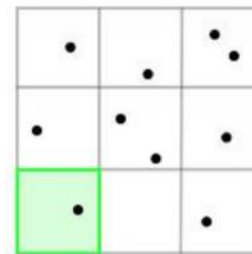
| 시각화 (Visualization)

- 문제 해결 목적이 아니더라도 데이터를 시각화하기 위해서도 차원축소 진행
- 4차원 이상의 feature를 가지는 데이터에 대해서는 시각화가 불가능하며, 이 경우 차원 축소를 통해 데이터를 2차원 혹은 3차원으로 나타내면 데이터 패턴을 쉽게 인지 가능



| 차원의 저주 (The curse of dimensionality)

- 데이터의 차원이 증가할 수록 데이터가 존재하는 공간의 크기가 기하급수적으로 증가하여 데이터의 밀도가 희박해지는 현상
- 모델이 탐색해야 하는 데이터 공간은 증가하는 반면 정보가 있는 공간 (데이터가 있는 공간)은 줄어들기 때문에 모델의 성능이 급격히 감소
- 이를 해결하기 위해서는 데이터 개수를 증가시키거나 feature 개수를 줄여야 함



회귀

Contents

단순 선형 회귀 분석

다중 선형 회귀 분석

손실함수

최적화

평가

회귀 (Regression)

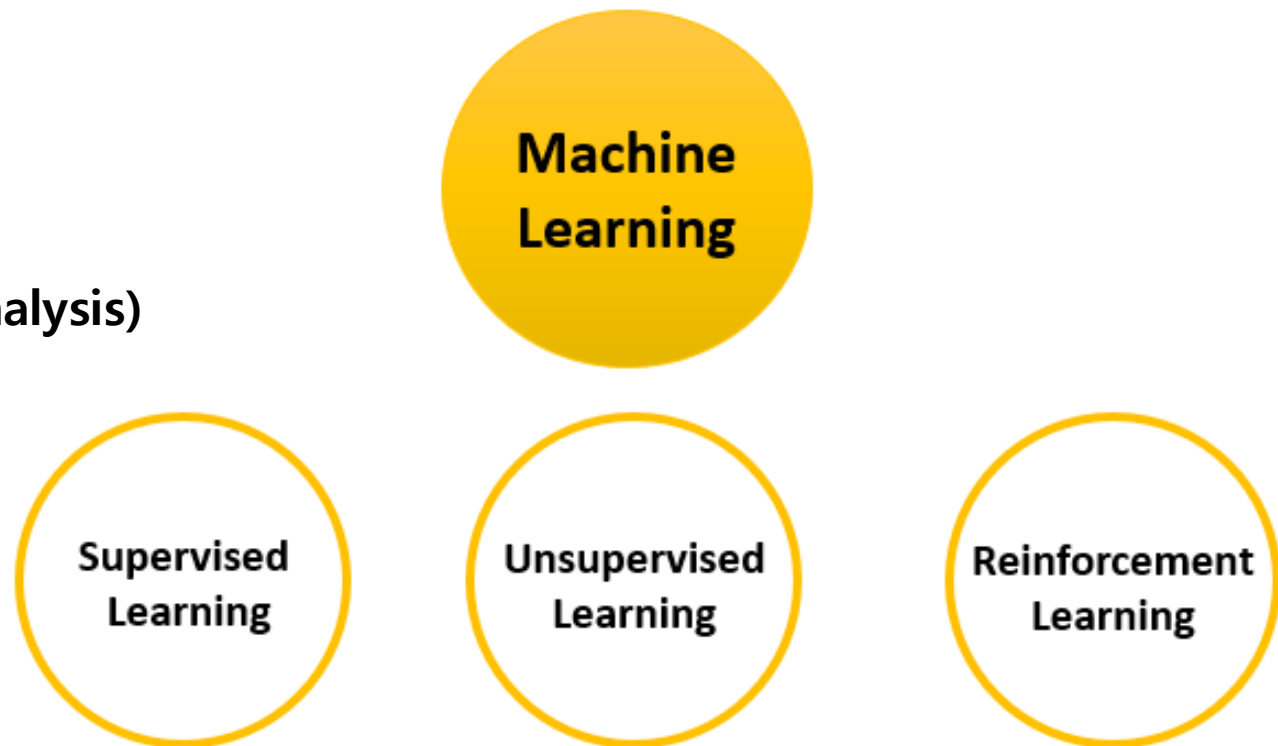
한 개 이상의 독립변수 x 에 대해 연속적인 값을 가지는 종속변수 y 를 예측하는 문제

| 단순 선형 회귀 분석 (Simple linear regression analysis)

- 독립변수 x 의 feature 가 1개일 때

| 다중 선형 회귀 분석 (Multiple linear regression analysis)

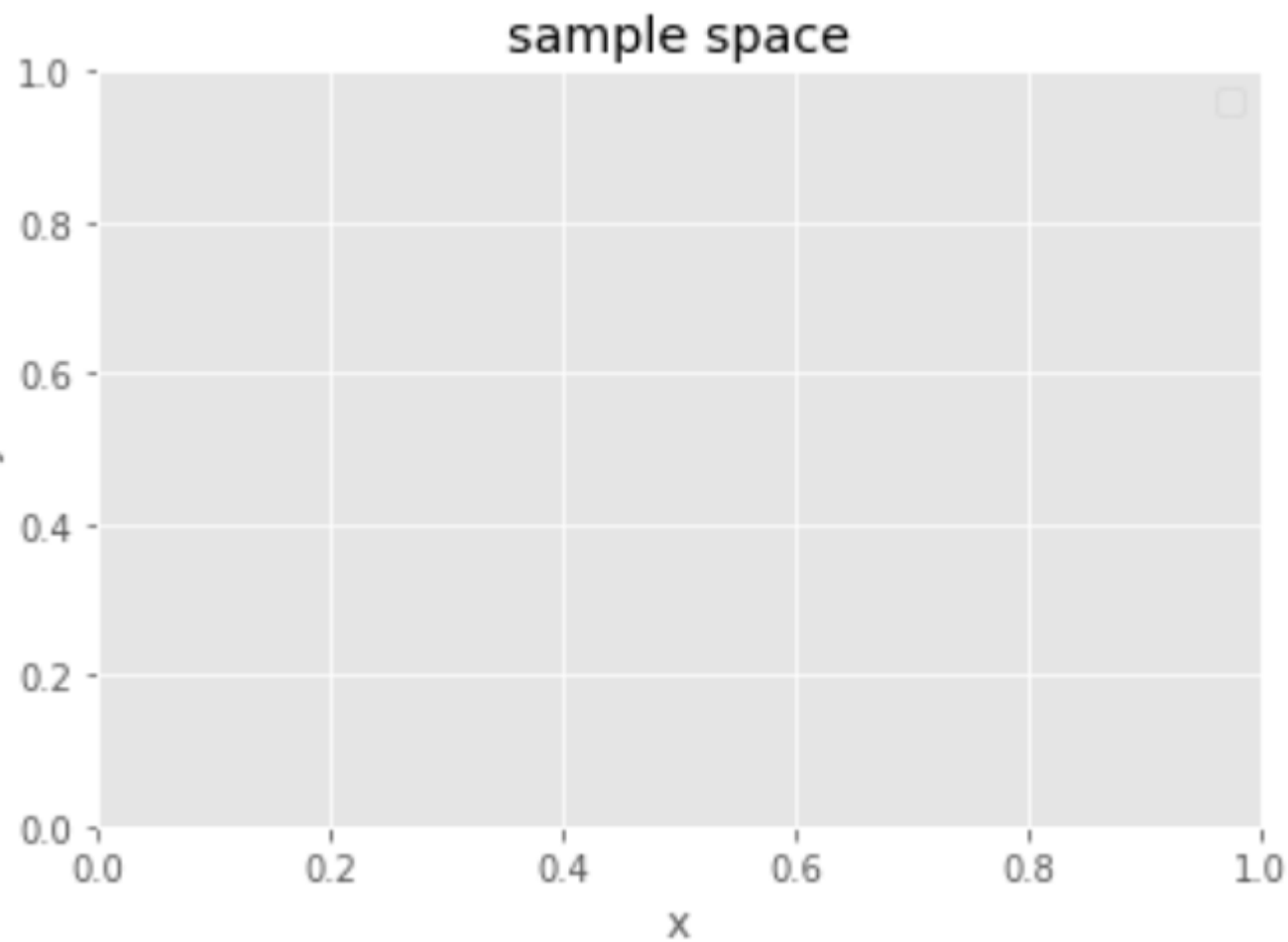
- 독립변수 x 의 feature 가 2개 이상일 때



단순 선형 회귀 분석 예제

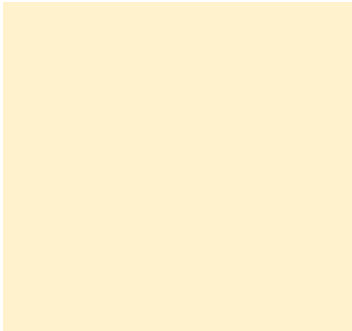
공부 시간에 대한 종합점수를 예측해보자

공부시간(x)	종합점수(y)
3	20
5	50
5	60
7	90
8	100



Step 1. 모델 구현

$$model(x; w) = wx$$

x	Model	y pred	y
3			20
5			50
5			60
7			90
8			100

Step 2. 손실 (Loss) 계산

| 손실 함수(Loss function) = 목적 함수(Objective function) = 비용 함수(Cost function)

정답 값과 예측 값에 대한 오차에 대한 식

| 손실 (Loss = Error)

손실 함수로부터 얻은 정답 값과의 차이

x	Model		y pred	y	Loss function		Loss
3				20			
5				50			
5				60			
7				90			
8				100			

Step 3. 최적화 (Optimizer)

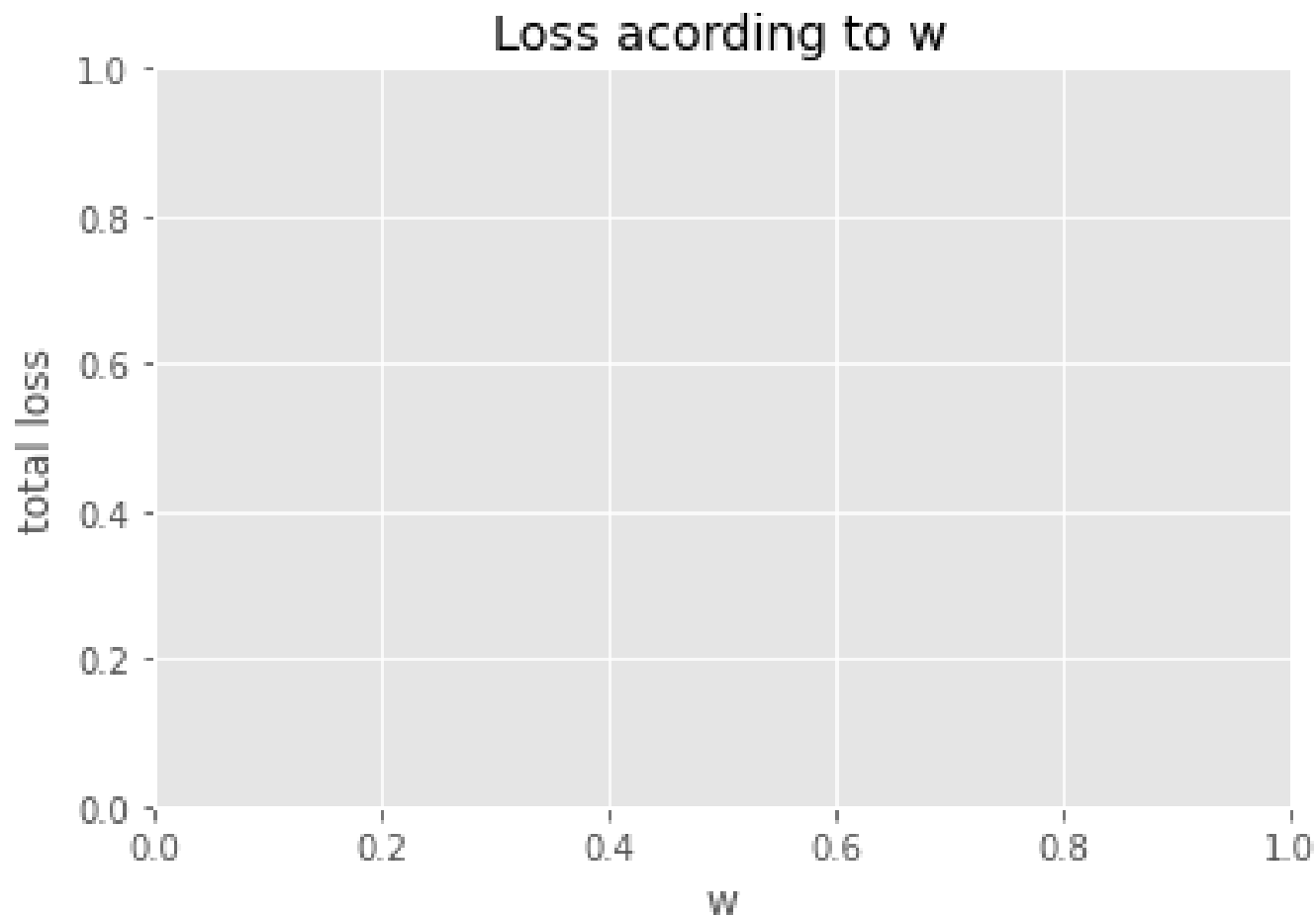
손실함수의 값을 최소화(경우에 따라서 최대화) 하도록 학습 파라미터를 조정해 가는 과정

x	Model	y pred	y	Loss function	Loss
3			20		
5			50		
5			60		
7			90		
8			100		

x	Model	y pred	y	Loss function	Loss
3			20		
5			50		
5			60		
7			90		
8			100		

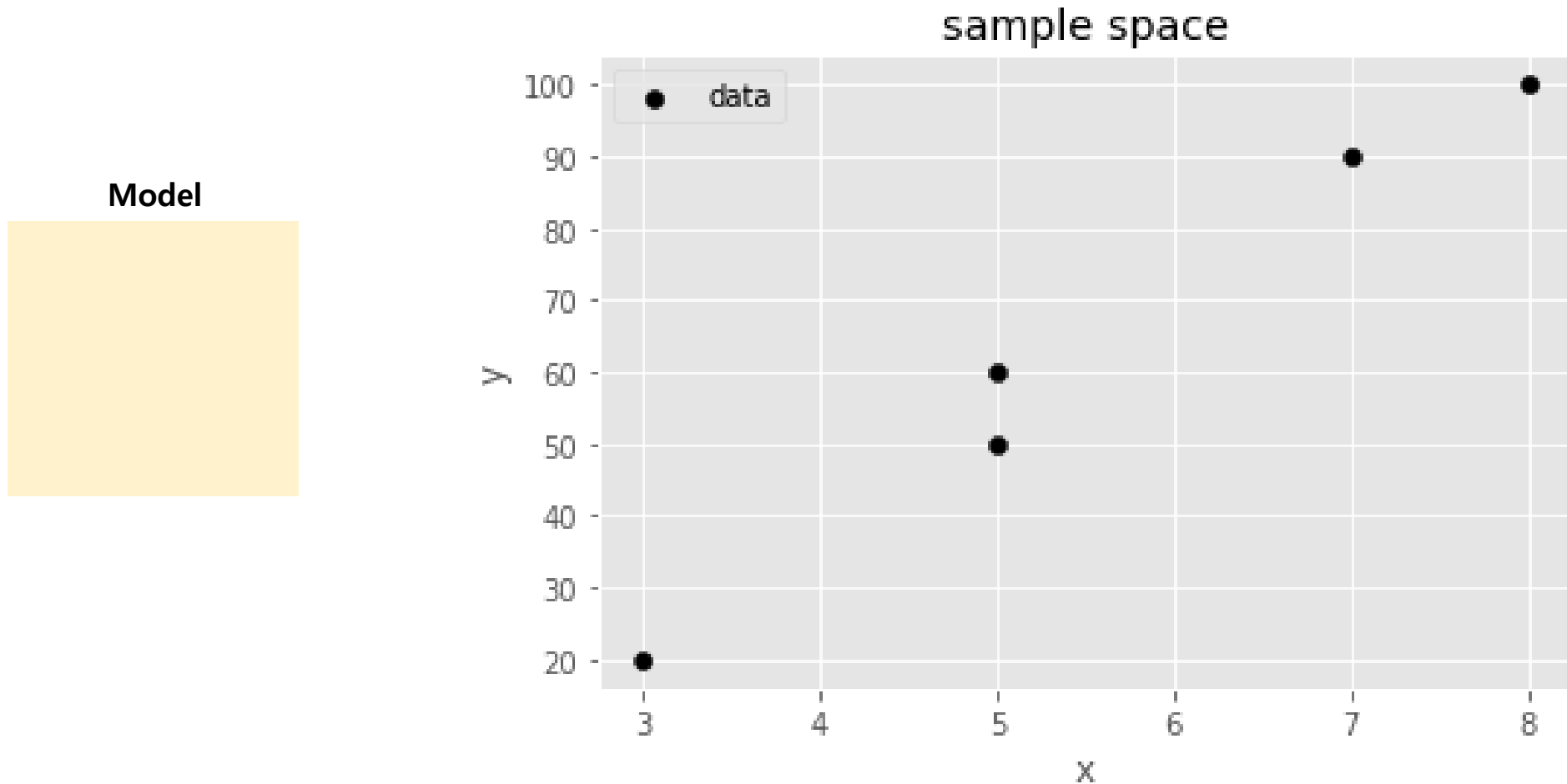
....

x	Model	y pred	y	Loss function	Loss
3			20		
5			50		
5			60		
7			90		
8			100		



Step 4. 예측 (Optimizer)

학습이 끝난 후 학습파라미터는 고정된 상태로 데이터셋에 대한 예측이 진행



Step 5. 평가 (Evaluation)

학습이 끝난 후 학습파라미터는 고정된 상태로 데이터셋에 대한 예측이 진행

테스트 데이터

공부시간(x)	종합점수(y)
1	5
2	15
5	55
6	80
7	85

Model

