CNU 데이터 분석 교육

7th lecture

"Linear Analysis-1"

2022 - 11 - 07

수학과 최명수

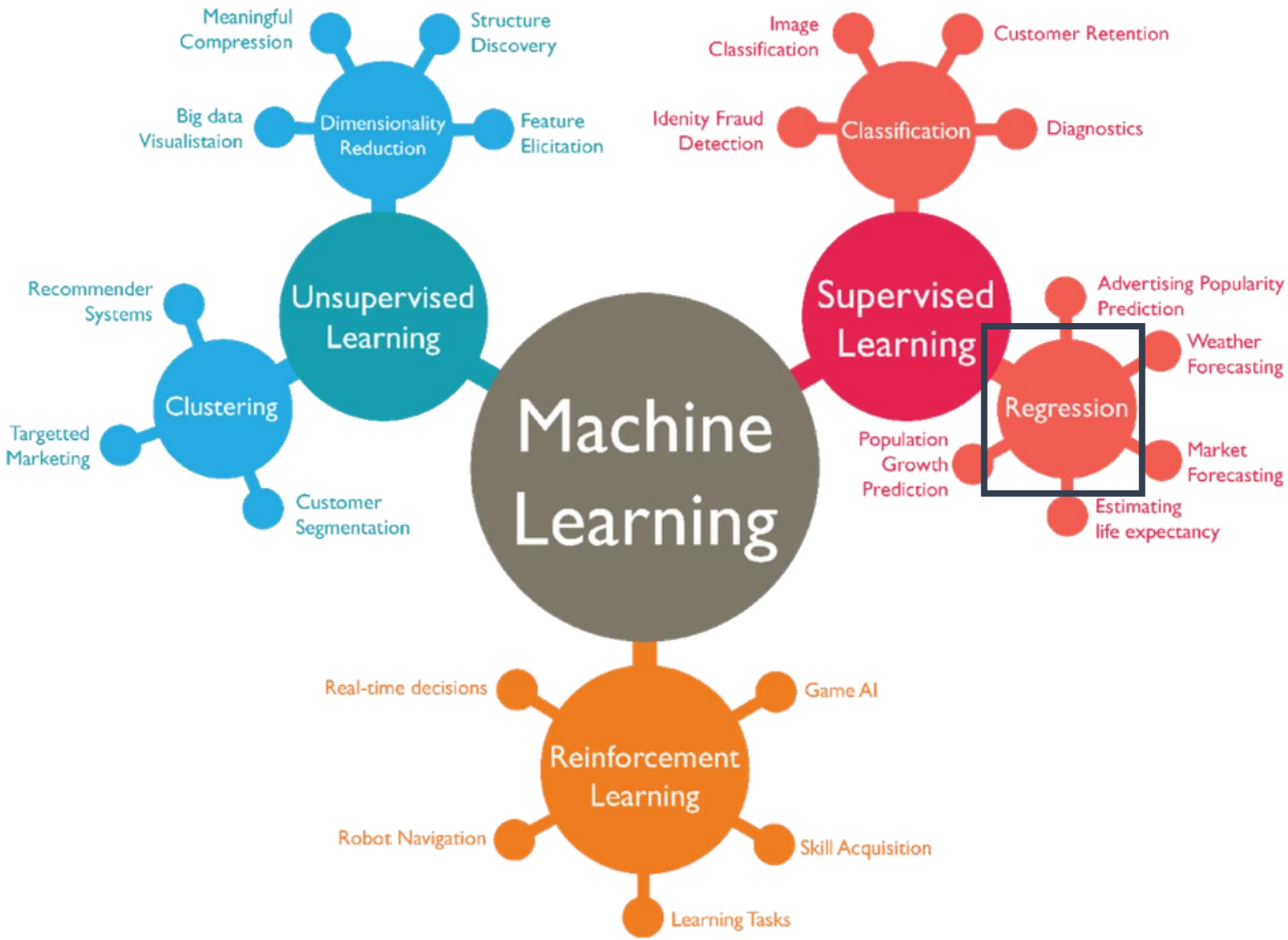
지난시간?

1. Matplotlib, Seaborn – Visualization

오늘은 무엇을?

1. 선형회귀(Linear regression) 모델 개발

   - 두 변수의 관계를 가장 잘 설명할 수 있는 선형함수가 무엇일까?
   - 그 선형함수를 찾는데 어떤 수학적인 방법이 이용될까?
   - 수학적인 아이디어를 코딩으로 나타내는 방법?


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating life expectancy
Population Growth Prediction
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

지도학습(Supervised learning)

지도학습(Supervised learning)

data	Label
	사과
	바나나

학습!

: 사과

: 바나나

회귀분석(Regression)이란?

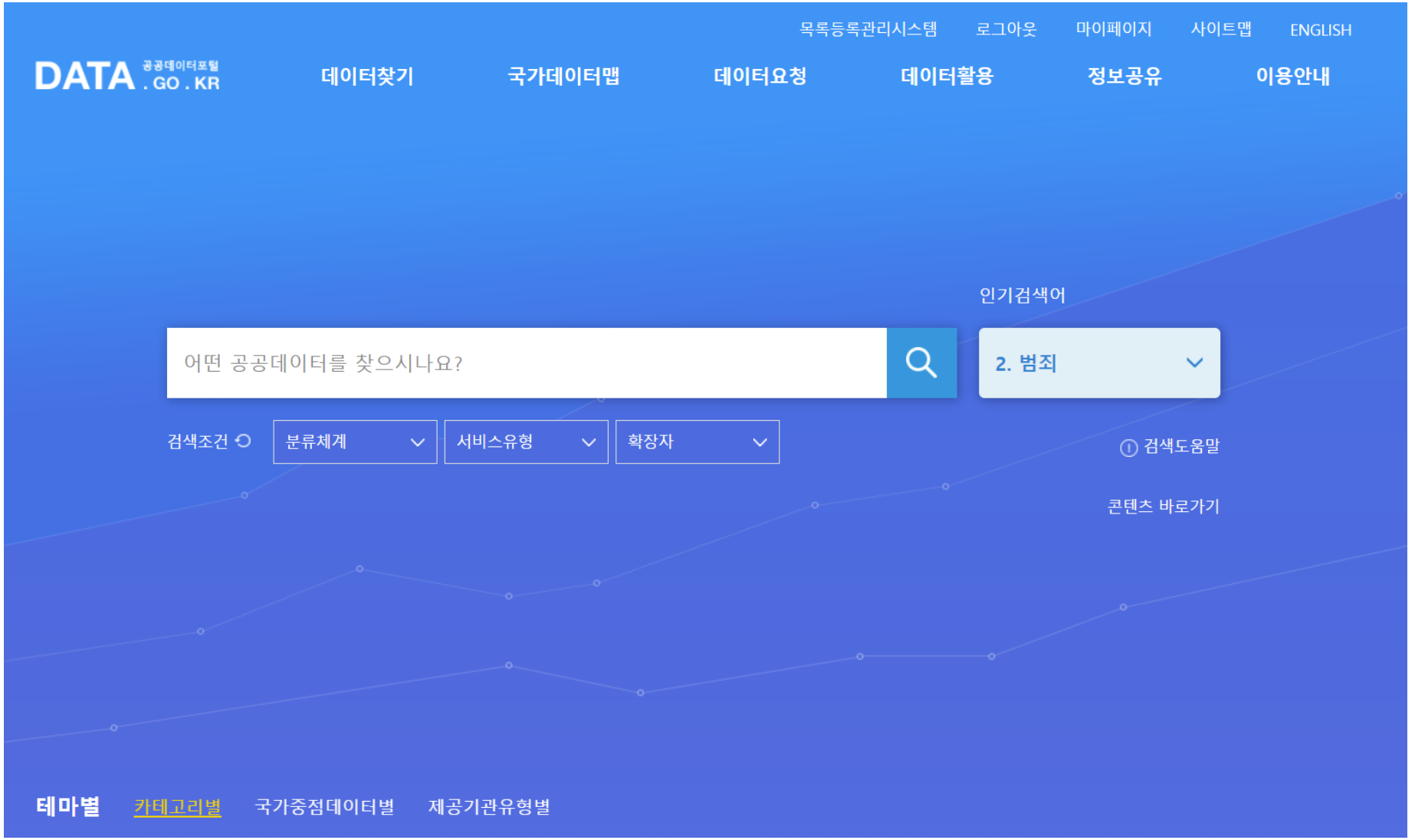
1. 독립변수 X와 종속변수 Y의 관계를 모델링하는 기법
   -니네 둘 무슨 사이니…?

2. X는 일반적으로 n차원, Y는 1차원

3. 둘 사이의 관계를 수치로 평가

   data로 부터 원하는 값을 예측하고 싶을때 주로 사용.

공공데이터포털


목록등록관리시스템
로그아웃
마이페이지
사이트맵
ENGLISH
DATA 공공데이터포털 .GO.KR
데이터찾기
국가데이터맵
데이터요청
데이터활용
정보공유
이용안내
인기검색어
어떤 공공데이터를 찾으시나요?
2. 범죄
검색조건
분류체계
서비스유형
확장자
검색도움말
콘텐츠 바로가기
테마별
카테고리별
국가중점데이터별
제공기관유형별

공공데이터포털

파일데이터 (1,187건)

더보기 >

교육 국가행정기관 미리보기

CSV JSON + XML **교육부_학교건강검사 표본조사결과(건강조사)_2017**

교육부_학교건강검사 표본조사결과(건강조사)_2017

제공기관 교육부 **수정일** 2020-09-15 **조회수** 3029 **다운로드** 704 **주기성 데이터** 2 **키워드** 학교건강,건강검사,표본조사

다운로드

교육 국가행정기관 미리보기

CSV JSON + XML **교육부_학교건강검사 조사 결과 2016**

전국 초·중·고등학생의 건강조사 표본조사 결과

제공기관 교육부 **수정일** 2020-09-03 **조회수** 2557 **다운로드** 426 **키워드** 학교건강,건강검사,표본조사

다운로드

교육 국가행정기관 미리보기

CSV JSON + XML **교육부_학생건강검사 키rawdata 2016**

학생표본 신체(키) 검사 rawdata

제공기관 교육부 **수정일** 2020-09-22 **조회수** 3470 **다운로드** 846 **키워드** 학교건강,건강검사,표본조사

다운로드

교육 국가행정기관 미리보기

CSV JSON + XML **교육부_학생건강검사 결과분석 rawdata 서울 2015**

2015년도 학생건강검사 결과(표본) 서울지역 학생 키, 몸무게, 혈당, 고지혈증, B형간염, 혈색소에 대한 raw data

제공기관 교육부 **수정일** 2020-09-03 **조회수** 4109 **다운로드** 1214 **키워드** 학교건강,건강검사,표본조사

의견수렴 게시판

공공데이터포털

CSV 교육부_학생건강검사 결과분석 rawdata 서울 2015

다운로드 | 오류신고 및 담당자 문의

파일데이터 정보 | 메타데이터 다운로드

- 파일데이터명
- 분류체계
- 관리부서명
- 보유근거
- 업데이트 주기
- 매체유형
- 확장자
- 데이터 한계
- 등록
- 제공형태
- 설명
- 기타 유의사항
- 비용부과유무
- 이용허락범위

미리보기

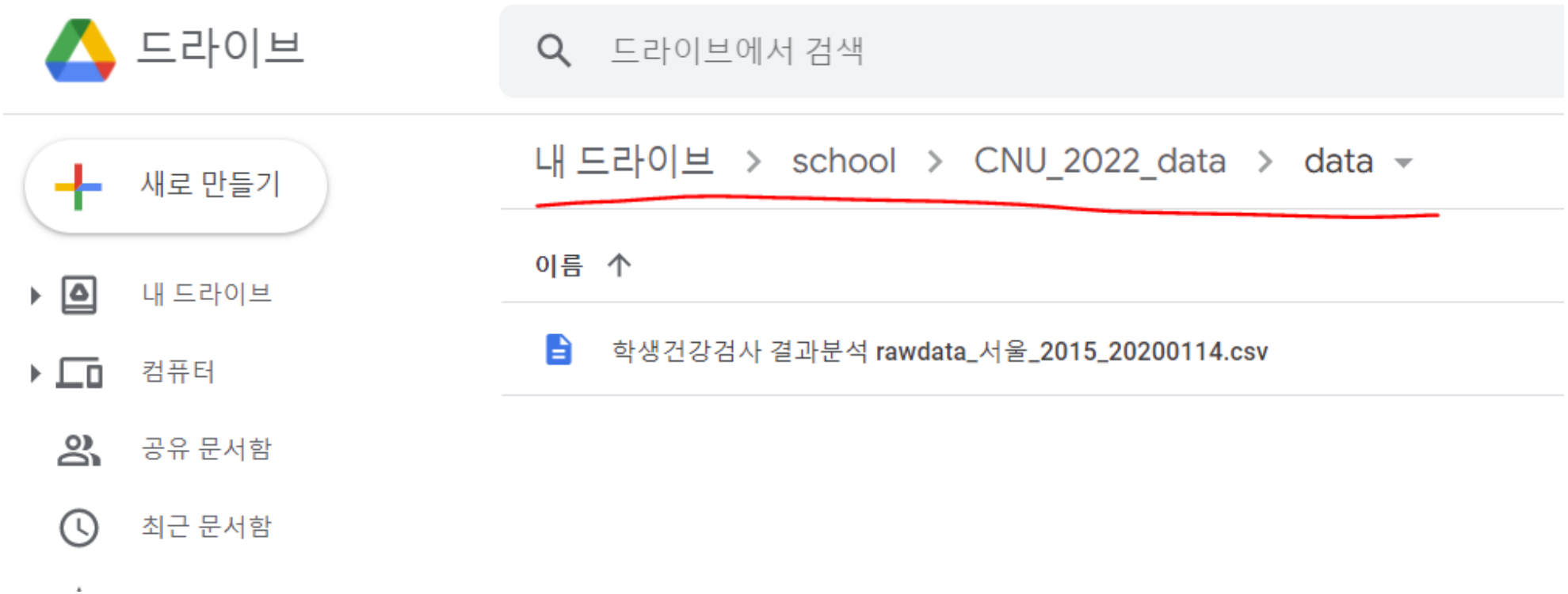
※ 파일 데이터의 일부 내용을 제공하고 있으며, 전체 내용이 필요한 경우 해당 파일을 다운로드 받으시기 바랍니다.

ID	최종가중치	학교ID	도시규모	도시규모별분석용	학년도	광역시도	시도별	학
Aa011남10101	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10102	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10103	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10104	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10105	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10106	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10107	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10108	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1
Aa011남10109	169.550665	Aa01	대도시/중소도시	특별/광역시	2015	서울	서울특별시교육청	1

Github 다운로드

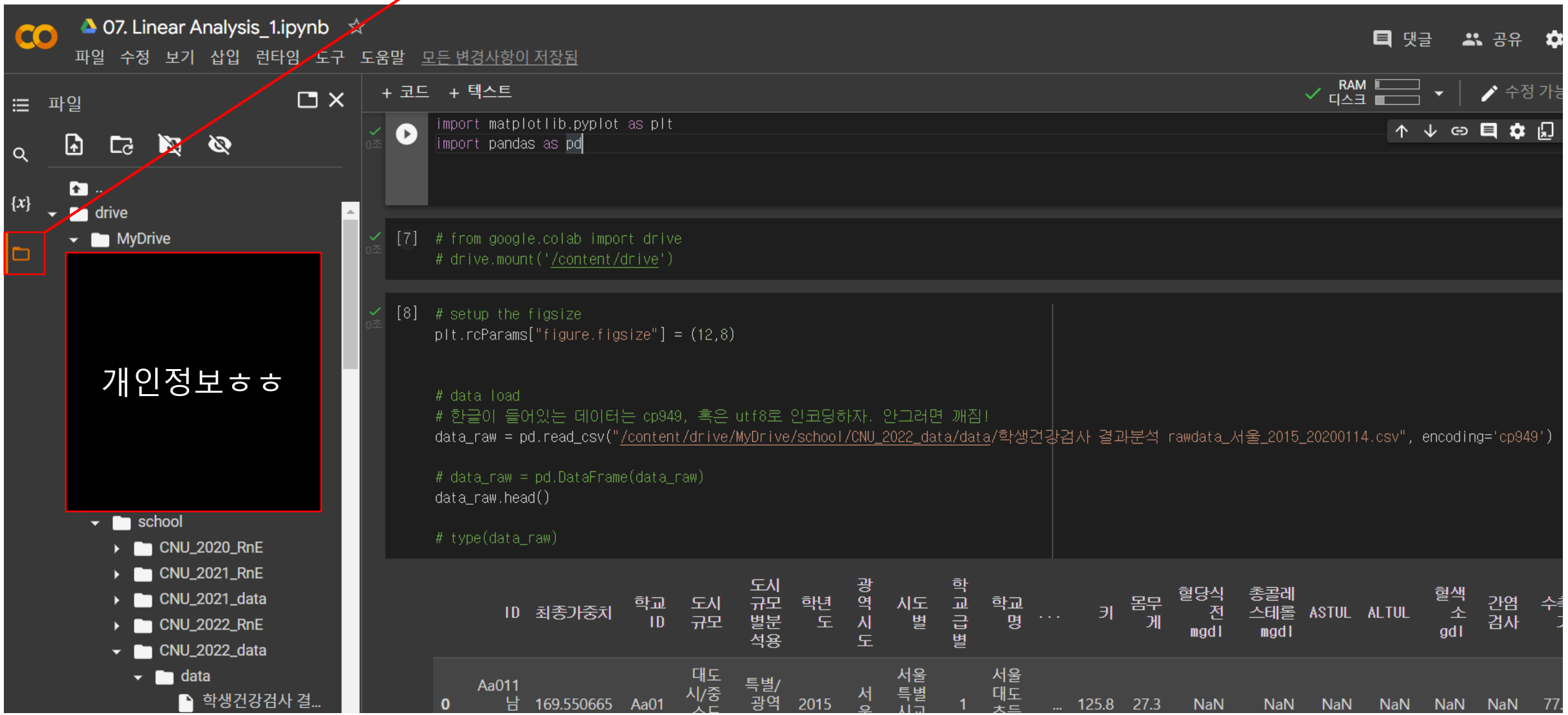
https://github.com/NumberL/NumberT

데이터 업로드


드라이브
드라이브에서 검색
새로 만들기
내 드라이브
컴퓨터
공유 문서함
최근 문서함
내 드라이브 > school > CNU_2022_data > data
이름
학생건강검사 결과분석 rawdata_서울_2015_20200114.csv

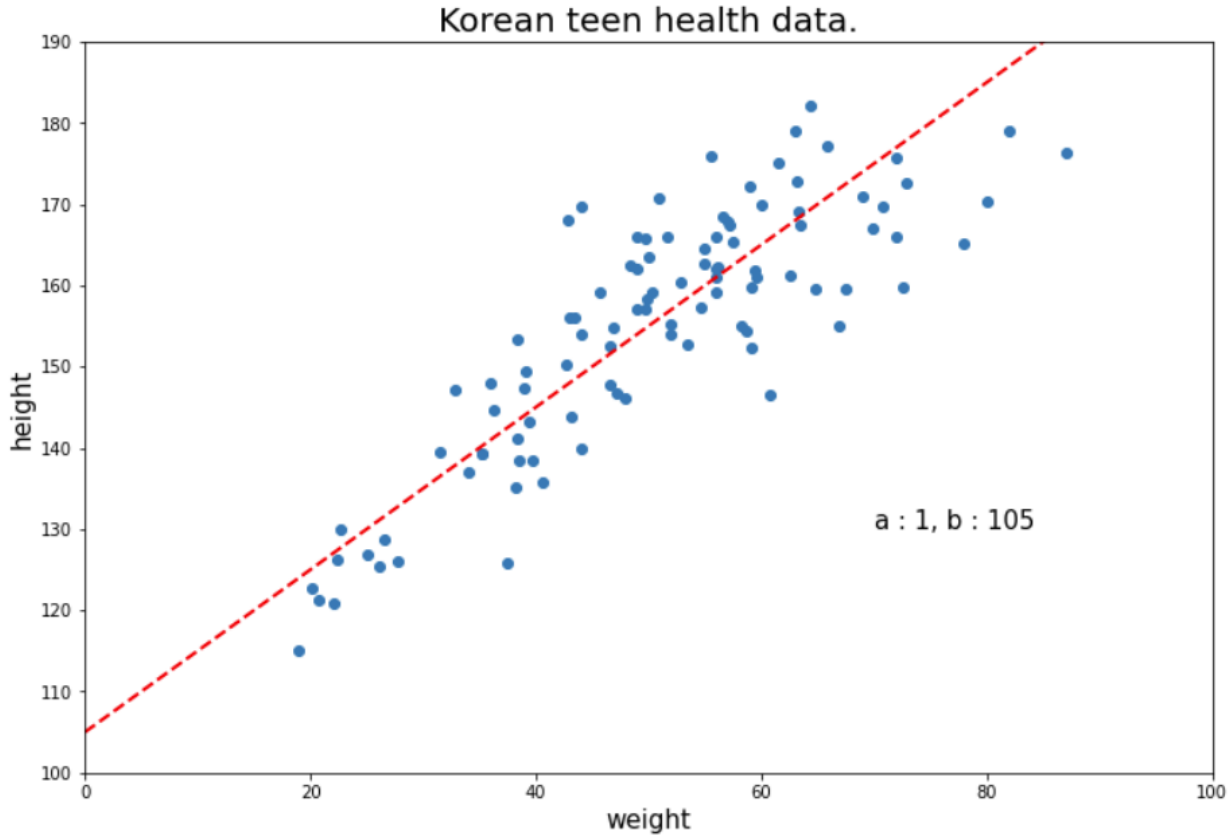
데이터 업로드

1. 드라이브 마운트 후 데이터를 올린 경로를 복사!
2. Encoding = 'cp949'로 설정 후 데이터가 잘 로딩이 되면 끝!



문제 1

- 주어진 scatter plot 위에 직선을 그려주는 함수를 def로 선언한 후 plt.plot 기능을 이용하여 다음과 같은 그래프를 그려라.

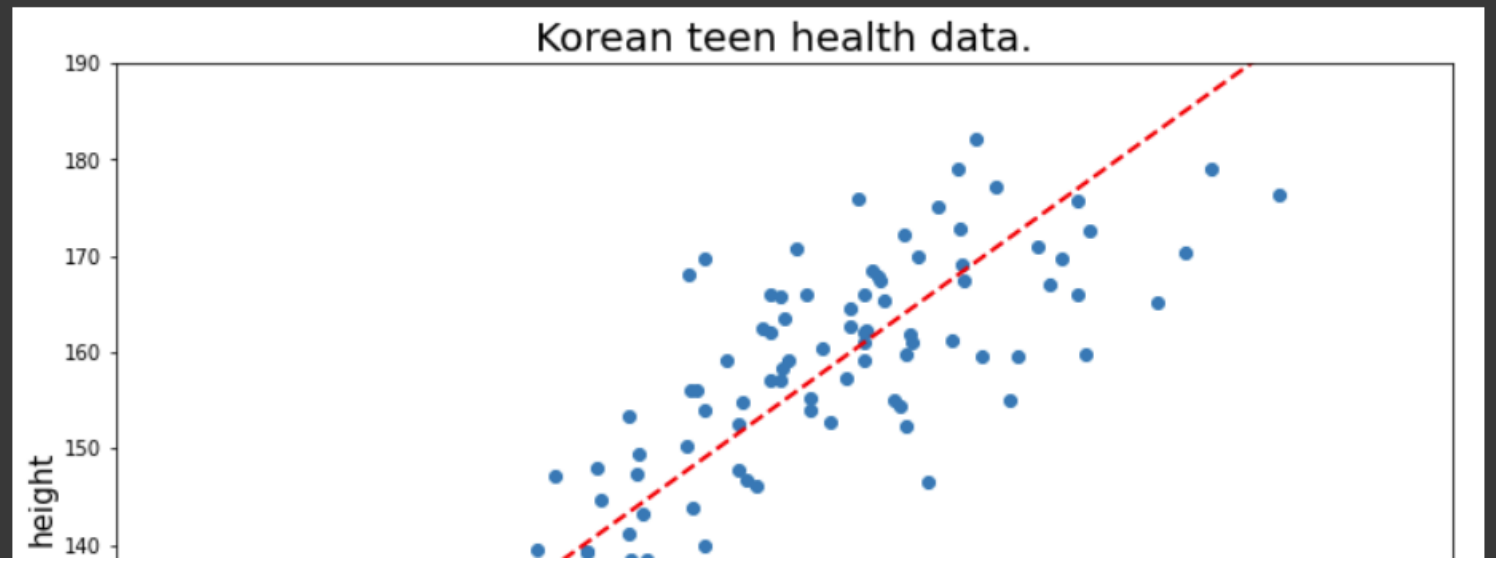


문제 1

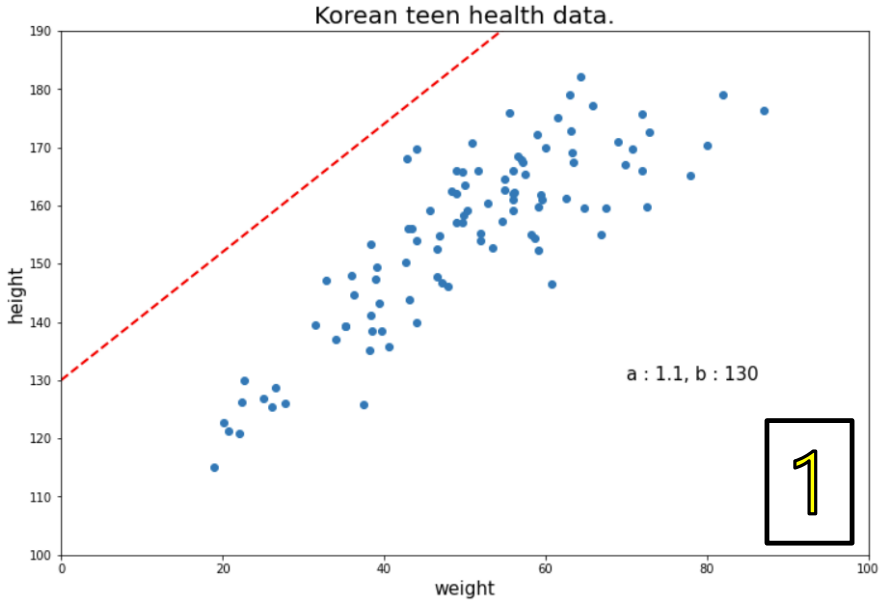
- 정답

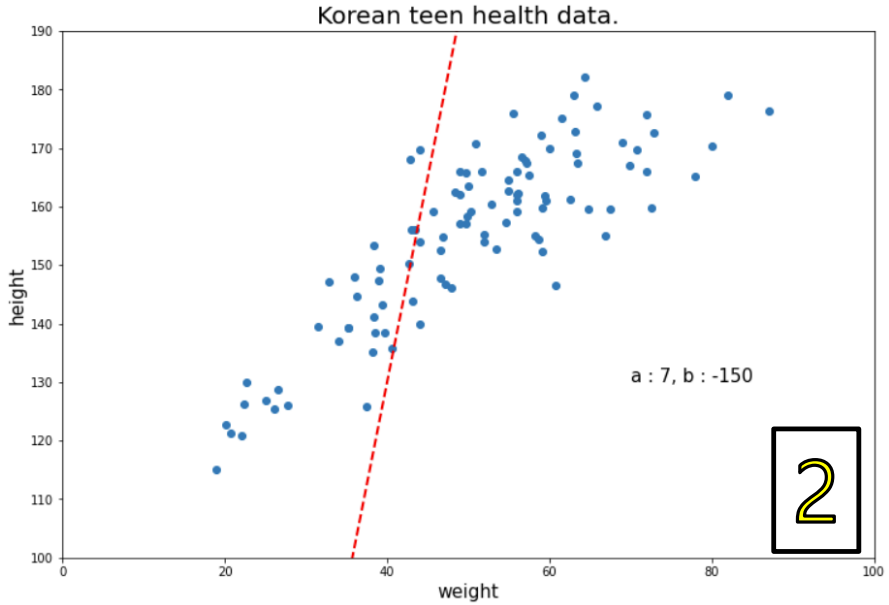
```
# linear function 생성
# y = ax + b


def height_pred(x, a, b):
    return a*x + b
```

```
# Line 그리기

a = 1
b = 105

x = np.linspace(0,100,10001)
y = height_pred(x, a, b)

plt.scatter(weight_100, height_100)
plt.plot(x,y, linestyle = '--', linewidth = 2, color = 'red')
plt.title("Korean teen health data.", fontsize = 20)
plt.xlabel("weight", fontsize = 15)
plt.ylabel("height", fontsize = 15)
plt.text(70, 130, "a : {}, b : {}".format(a,b), fontsize = 15)

plt.xlim([0,100])
plt.ylim([100,190])
plt.show()
```

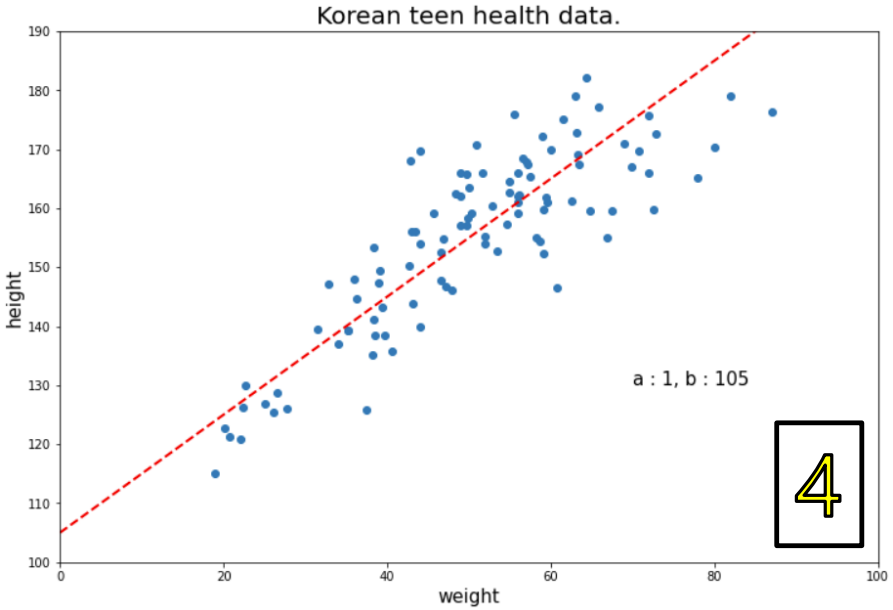


어느 것이 적당해 보임?


Korean teen health data.
height
weight
a : 1.1, b : 130
1


Korean teen health data.
height
weight
a : 7, b : -150
2


Korean teen health data.
height
weight
a : 0.1, b : 150
3

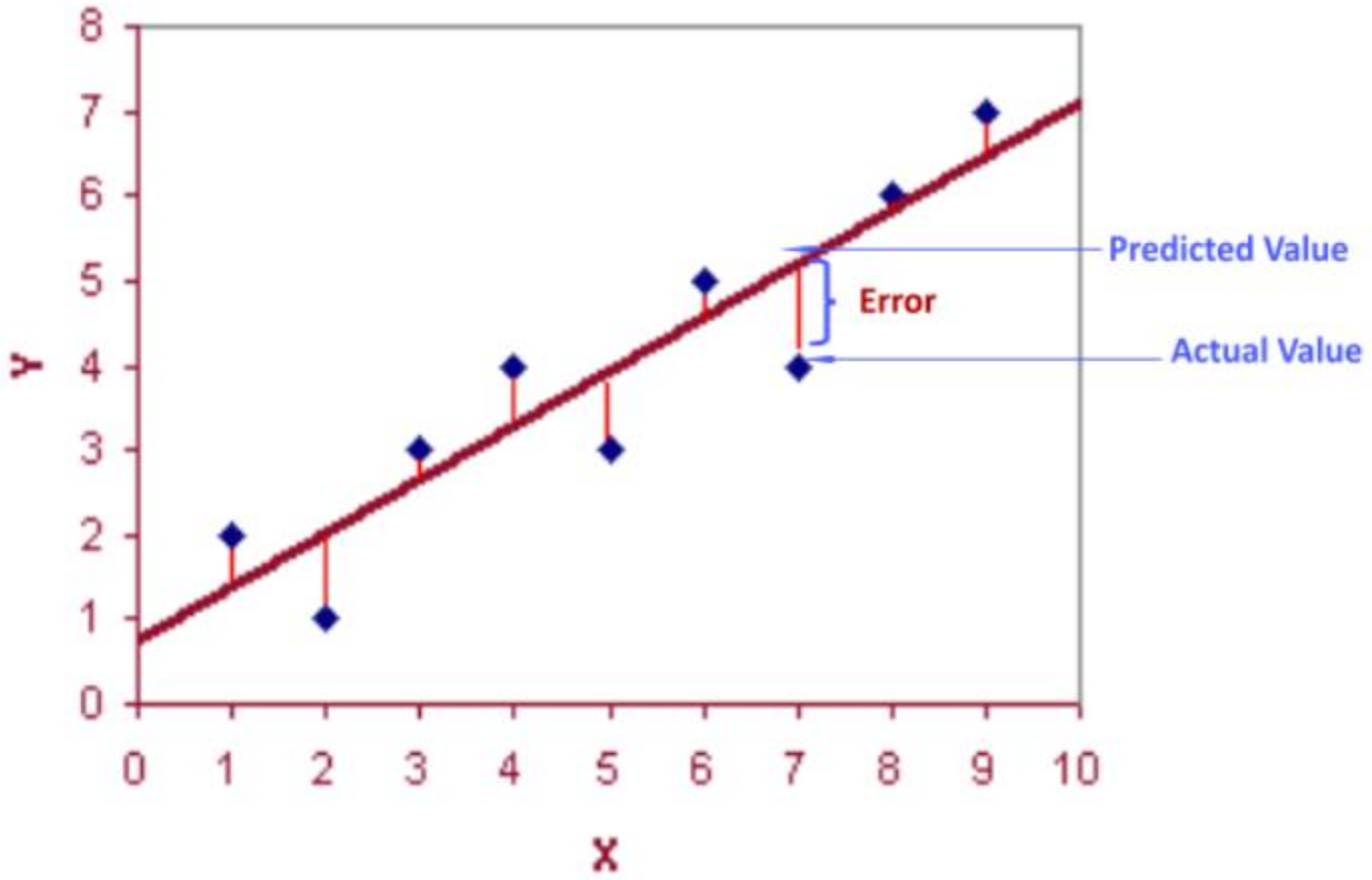

Korean teen health data.
height
weight
a : 1, b : 105
4

어느 것이 적당해 보임?

당신들이 선택한것의 공통점이 무엇인가?

1. 데이터와 직선간의 거리가 가깝다
->점과 직선사이의거리...?

어느 것이 적당해 보임?



손실함수(loss function, cost function)

Means Square Error

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

오차의 정도를 '양수'로 표현해준다.

Root Mean Square Error

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Object!

이것을 최소화 시키자!

*손실함수를 목적함수(Object function) 이라고도 한다.

손실함수(loss function)

HOW??????

가능한 모든 a, b의 쌍을 입력해서 그중 최소의 RMSE를 선택!

문제2

- a(기울기)와 b(y절편)의 범위를 설정 한 후 각각의 대한 RMSE 를 DataFrame으로 작성하라.

	a	b	RMSE
0	0.9	90.0	21.278450
1	0.9	100.0	12.617387
2	0.9	110.0	8.100892
3	1.0	90.0	16.810529
4	1.0	100.0	9.479130
5	1.0	110.0	9.854639
6	1.1	90.0	12.943688
7	1.1	100.0	8.696726
8	1.1	110.0	13.554595

문제2

- 정답

```
# 문제 2

a_range = np.linspace(0.9,1.1,3)
b_range = np.linspace(90,110,3)

a_lst = []
b_lst = []
rmse_list = []

x = np.linspace(0,100,10001)
y = height_pred(x, a, b)

df = pd.concat([weight_100, height_100], axis = 1)
```

```
for a in a_range:
    for b in b_range:

        df["예상 키"] = height_pred(df["몸무게"], a, b)
        diff = df["키"] - df["예상 키"]
        RMSE = np.sqrt(sum(diff**2)/len(diff))

        a_lst.append(a)
        b_lst.append(b)
        rmse_list.append(RMSE)

data_rmse = pd.DataFrame([a_lst, b_lst,rmse_lst]).T
data_rmse.columns = ["a", "b", "RMSE"]
data_rmse
```

손실함수(loss function)

HOW??????

가능한 모든 a, b의 쌍을 입력해서 그중 최소의 RMSE를 선택!

그러면 우리가 찾은 **parameter** a와 b는
한국청소년들의 키와 몸무게를 설명하는 선형함수 모델에서
가장 좋은 성능을 나타내는 것이다.

이러한 방법을 "

그거슨 다음시간에 확인

정리 요약

1. 선형회귀(Linear regression)란 데이터를 가장 잘 설명해줄 수 있는 선형함수를 찾는 것이다

2. 데이터를 가장 잘 설명해준다는 것은 정의된 "loss function" 을 최소화 하는 것이다.

3. 우리는 데이터의 수(n)과 parameter의 범위(a_range, b_range)를 입력(input)하여 loss function을 최소화하는 a, b를 출력(output)는 함수(SLR)를 만들었다.

4. 데이터로 부터 가장 좋은 값을 얻을 수 있는 a,b를 학습(learn)하였다.

회귀분석(Regression test)이란?

1. 독립변수 X와 종속변수 Y의 관계를 모델링하는 기법
   -니네 둘 무슨 사이니…?

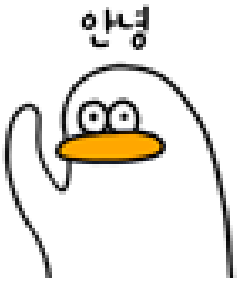
2. X는 일반적으로 n차원, Y는 1차원

3. Error의 척도를 가장 낮추는 parameter를 구하는 방법

공지

1. 어려웠?

끝



담에봅시당