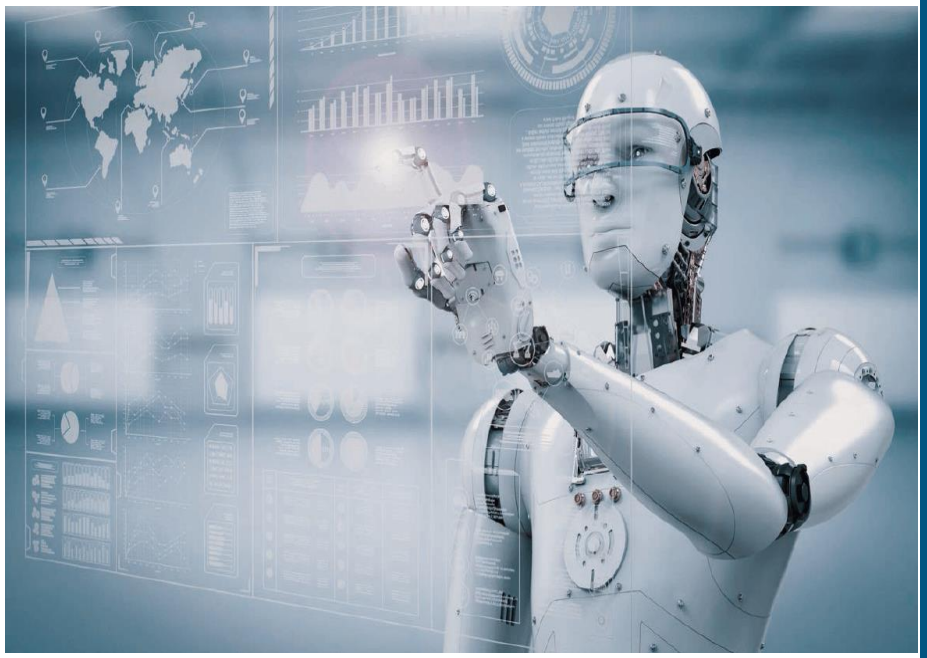


기계학습개론

- 데이터 다루기 -

교수 이홍로

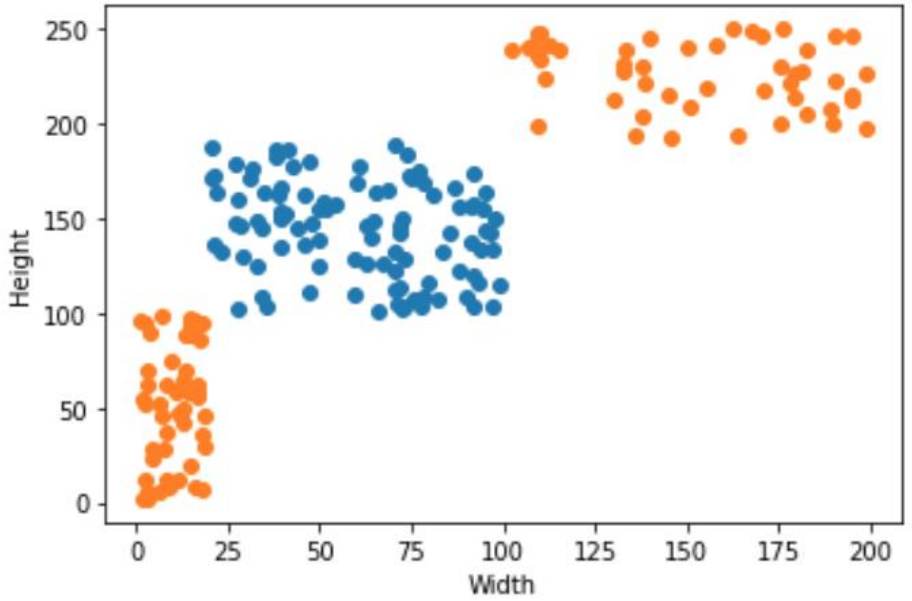
MP : 010-6611-3896
E-mail : hrlee@cnu.ac.kr
강의 홈페이지 : https://cyber.hanbat.ac.kr/

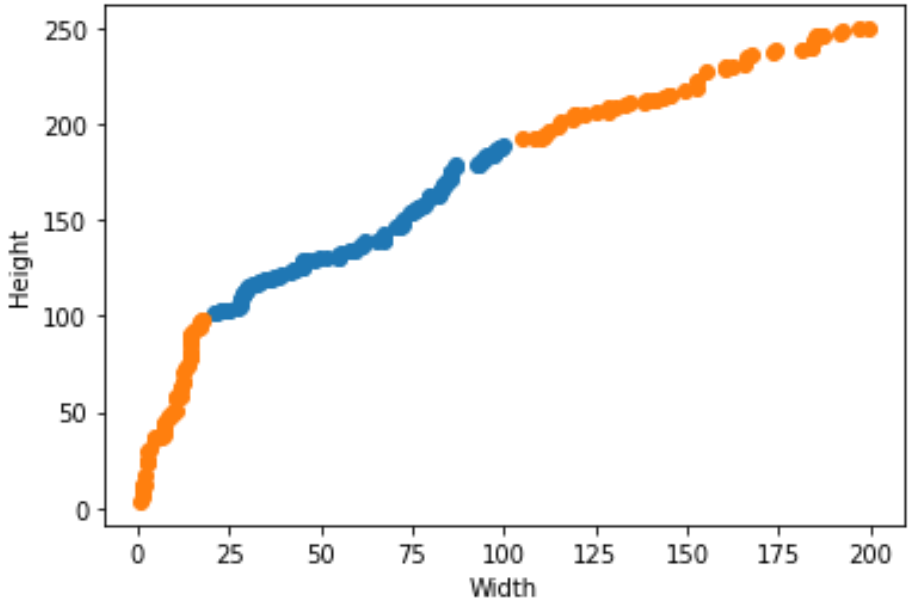
오늘의 강의 목표

- 데이터 정제
- Train Set 및 Test Set 생성

데이터 정제

- 좀 더 현실적인 데이터로 수정





데이터 정제

- Sort 적용

```
normal_person_width = np.sort(np.random.uniform(low=20, high=100, size=100))
normal_person_height = np.sort(np.random.uniform(low=100, high=190, size=100))

error_person_width = np.sort(np.append(np.random.uniform(low= 1, high=19, size=50),
                                       np.random.uniform(low=101, high=200, size=50)))
error_person_height= np.sort(np.append(np.random.uniform(low=  1, high=99, size=50),
                                       np.random.uniform(low=191, high=250, size=50)))
```

데이터 정제

- 넘파일 데이터로 수정

Numpy 2차원 array

person_data [[1,2],[3,4],[5,6],…,[49,50]] → [[1 2]
[3 4]
[5 6]
…
[49 50]]

```
width = np.append(normal_person_width, error_person_width)
height = np.append(normal_person_height, error_person_height)
answer = [1]*100 + [0]*100
person_data = [[l, w] for l, w in zip(width, height)]

person_data = np.array(person_data)
answer = np.array(answer)
```

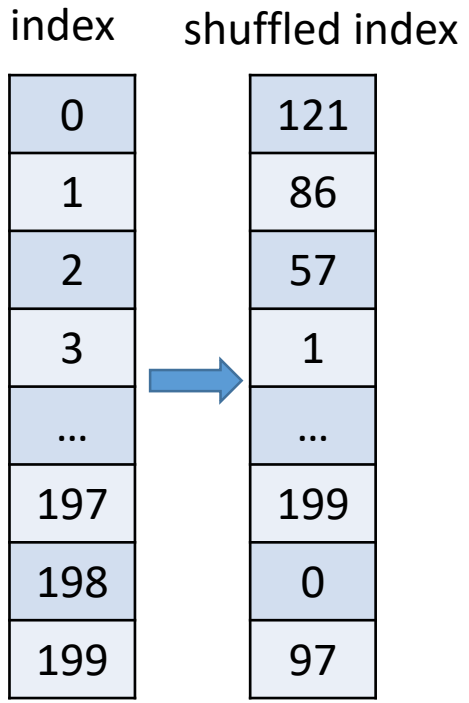
Train Set 및 Test Set 생성

- **Train Set(75%), Test Set(25%) 생성**

```
np.random.seed(42)
index = np.arange(200)
np.random.shuffle(index)

train_data = person_data[index[:150]]
train_answer = answer[index[:150]]

test_data = person_data[index[150:]]
test_answer = answer[index[150:]]
```



Height	Width	answer
220	50	0
200	55	0
190	54	0
175	52	1
160	100	1
⋮		⋮
155	45	1
150	50	1
50	10	0
45	44	0
32	31	0
20	53	0

```
from sklearn.model_selection import train_test_split

train_data, test_data, train_answer, test_answer
= train_test_split(person_data, answer, random_state=42)
```

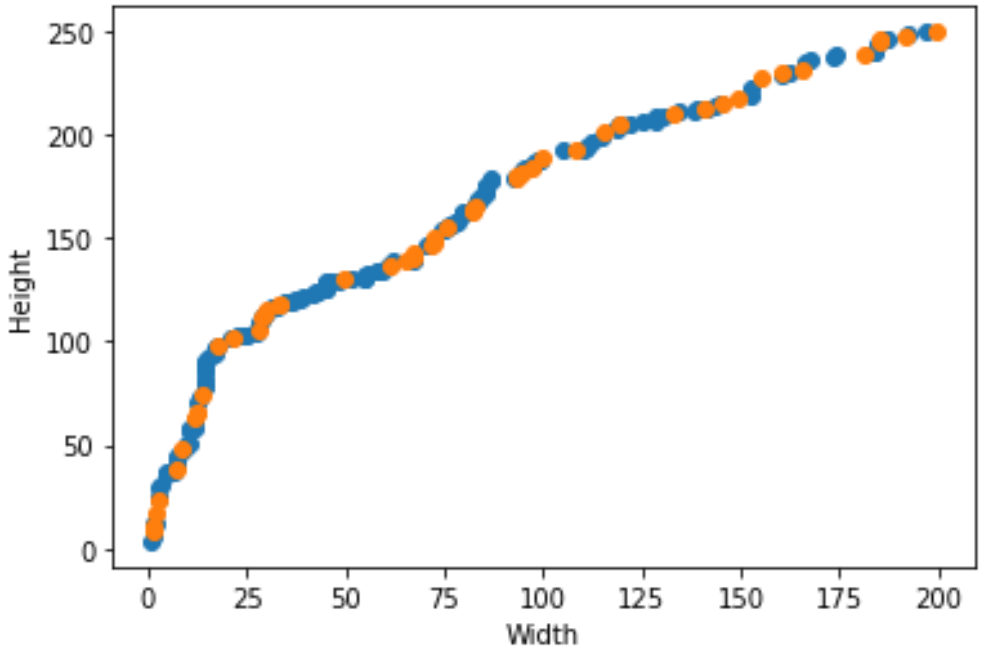
Train Set 75%

Test Set 25%

Train Set 및 Test Set 생성

- **Train Set 및 Test Set 출력**

```
plt.scatter(train_data[:,0], train_data[:,1])
plt.scatter(test_data[:,0], test_data[:,1])
plt.xlabel('Width')
plt.ylabel('Height')
plt.show()
```



Train Set 및 Test Set 생성

- 재 학습 후, 모델 평가하기

```
from sklearn.neighbors import KNeighborsClassifier


kn = KNeighborsClassifier()
kn.fit(train_data, train_answer)
kn.score(test_data, test_answer)
```

Train Set 및 Test Set 생성

- K 파라미터 찾기 알고리즘 수정

```
kn = KNeighborsClassifier()
kn.fit(train_data, train_answer)

for n in range(1, 100):
    kn.n_neighbors = n
    score = kn.score(test_data, test_answer)
    print(str(n)+":"+str(score))
```

“하루에 3시간을 걸으면 7년 후에
지구를 한바퀴 돌 수 있다”

-사무엘존슨-