**<Chapter2. 데이터 다루기> : 특성의 스케일이 다른경우**

**-> Practice03, Practice04**

**배열은 (길이, 무게)**

지도학습 : 입력과 타깃(정답)으로 이뤄진 훈련 데이터로 학습.

**-분류** : 샘플을 몇개의 클래스 중 하나로 분류할것인가?

-**회귀** : 임의의 어떤 숫자를 예측. (두 변수 사이의 상관관계 예측)

비지도 학습 : 입력으로 이뤄진 훈련 데이터로 학습

훈련 데이터 : 지도학습에서의 입력과 타깃을 합친 것

**Practice03 (indexes 사용)**

**=> fish\_data, fish\_target으로 훈련하고 score하면 당연히 1.0나옴!**

샘플링 편향 : 특정 종류의 샘플이 골고루 섞여 있지 않고 한쪽으로 치우쳤음을 의미 -> Practice03

넘파이(numpy) : 파이썬의 대표적인 배열 라이브러리

배열인덱싱(array indexing)(배열 인덱싱) : 1개의 인덱스가 아닌 여러개의 인덱스로 한 번에 여러개의 원소를 선택할 수 있음. -> Practice03

**Practice04 (indexes 배열 사용해 훈련데이터중 이웃샘플 따로 구분)**

**=>길이가 25cm, 무게 150g인 도미를 빙어로 예측하는 문제 발생 (전처리데이터를 만들어야 제대로 된 이웃을 찾을 수 있음)**

**=> distances, indexes = kn.kneighbors([[25,150]]) 으로 해당 데이터의 훈련데이터중 이웃까지의 거리와 이웃 샘플의 인덱스 반환**

**=>x축은 범위가 좁고(10~40), y 축은 범위가 넓어서(0~1000) 도미 샘플이 이웃으로 선택되지 못함 => xlim((0,1000))으로 수정 =>모든 데이터가 수직으로 늘어선 형태가 됨 => y축만 고려 대상이 됨 (두 특성의 스케일이 다르기 때문에 이런 문제 발생 => 표준점수를 이용해 데이터 전처리**