1주차

인프런 강의

1강

지도학습 - 입력, 타겟 (ex.k-최근접 이웃)

비지도학습 - 타겟 x

강화학습 - 환경의 보상값 최대화

슬라이싱을 이용해 훈련과 테스트용으로 데이터셋 나눔

→ 훈련 데이터와 테스트 데이터 구별

샘플링 편향이 일어나지 않도록 잘 나누어야 함

numpy - 입력 데이터 전달, 여러 라이브러리들이 의존

다차원 배열 제작 가능

다차원 벡터 - 원소 개수 (벡터와 배열 차이)

하나의 리스트에 다른 타입의 원소 공존 불가능

배열 인덱스 - 데이터 섞기

1. 따로 섞지 않고 입력 데이터 특성값과 타겟값이 쌍으로 잘 섞이도록 해야 함 (정답을 제대로 매칭)
2. 인덱스 분류 - 랜덤 섞기 - 훈련, 테스트 데이터 나누기 - 인덱스에서 원본 데이터 찾도록 설계

np.arrange - 인덱스 배열 만들기

np.random.shuffle - 인덱스 배열 넣어서 섞기

배열 슬라이싱

a = np.array([5,6,7,8])

a[[1,3]]

train_input, train_target → 훈련 데이터 한 쌍

test_input, test_target → 테스트 데이터 한 쌍

plt.scatter(train_input[:,0], train_input[:,1])

plt.scatter(test_input[:,0], test_input[:,1])

[:,0] : 첫 번째 열 (length)

[:,1] : 두 번째 열 (weight)

kn.fit(train_input, train_target)

kn.score(test_input, test_target) → 데이터가 잘 섞였는지 검증

이번 시간에 배운 것

- 훈련 데이터와 테스트 데이터는 나눠야 함
- 편중되어 있지 않도록 나누는 방법
- numpy 라이브러리
- 데이터를 섞을 때 특성과 타겟의 쌍을 이루기 위해 배열의 인덱스를 만들어서 섞인 인덱스를 배열 슬라이싱 기능으로 훈련, 테스트 세트를 나누고 최근접 모델 훈련, 평가함

QnA

인덱스 지정 : 배열[행,열] (ex. a[0,1])

검증 데이터를 사용해 모델 평가

일반적으로 행 - 샘플, 열 - 특성으로 나열

2강

양성 클래스 : 1 -> 찾고자 하는 타겟

음성 클래스 : 0

인덱스 배열 - 메모리에 부담 x

넘파이로 데이터 준비

샘플 - 행

특성 - 열

column_stack : 주어진 배열을 나란히 세운 뒤 열로 붙임

column_stack(fish_length, fish_weight)

np.concatenate : 입력되는 배열을 가로로 붙여줌

np.ones, np.zeros : 1, 0으로 구성된 배열 만들기

np.full((n,m)x) : x으로 구성된 n*m 배열

사이킷런으로 데이터 나누기

train_test_split 함수 : data와 target을 train.input,
test.input, train.target, test.target으로 나눔 (ex.2개의 배열 -> 나눠진 4개의 배열)

straitify 매개변수에 타깃을 전달 -> 입력값이 불완전해도 잘 섞이도록 함

k 최근접 이웃의 오류

이상값 분석 과정 : 그래프에서 이상값 발견 -> 배열 인덱싱 이용하여 특성값이 가까운 다섯 값 표시 -> 비교

분석 결과 : y값과 x값의 스케일이 맞지 않아 더 멀리 있어도 가깝게 보임 -> 스케일 조정 필요 (기준 맞추기)

plt.xlim : 축의 스케일을 수동 지정

출력 결과 : length보다 weight이 생선 구별에 있어 영향을 더 많이 미침, 이상값은 빙어

기준 변환 방법

- 표준 점수 : (특성 - 평균) / 표준편차
  - mean 함수 - 평균
  - std 함수 - 표준편차
  - axis - 0으로 설정 (특성마다 행 계산, 1은 행마다)

넘파이 브로드캐스팅 : 한 행에서의 계산을 전체 행에 적용

훈련 데이터의 평균과 표준편차로 테스트 데이터를 맞춤

(훈련 세트가 기준)

변환 뒤 모델 훈련, 그래프 출력 -> k 최근접 이웃 특성에 맞는 올바른 값으로 예측

모델에 맞는 데이터 전처리 과정이 필요

QnA

k 최근접 이웃 메소드 : new와 가까운 5개의 샘플의 거리와 인덱스를 추출 - train_scaled 행 차원의 인덱스 배열을 넣어 가까운 샘플 출력

train_input, mean, std 모두 numpy 배열

axis - 차원에 맞는 숫자까지 지정 가능

0번째 축 - 행, 1번째 축 - 열, 그 다음 - 깊이..