6/8 강의노트

KNN(K-Nearest Neighbor) 이진분류

텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

패턴을 파악하는 것이 아니라 모든 데이터를 다 기억해서 계산하므로 메모리 사용량이 높음

컬럼이 많을수록 해당 컬럼 개수만큼 n차원 거리 계산을 하고 그것을 모든 데이터 개수만큼 실행하므로 계산이 복잡하여 오래 걸릴 수 있음

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

.fit 명령어로 모델 작성 괄호 내 첫번째는 독립변수(원본데이터), 두번째는 종속변수(결과값)

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

.score : 모델의 정확도 평가 (0~1), 매개변수는 독립변수, 종속변수 순

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

길이 30, 무게 600의 물고기의 종을 예측

위 훈련 (score 함수 사용)은 잘못된 훈련 : 학습 데이터를 그대로 테스트에 적용하여 정확도가 100%일수밖에 없음

그러므로 원본 데이터를 분할하여 일부만 떼어서 학습모델을 만듦

텍스트, 영수증, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

일반적으로 전체 데이터의 70~80% 만큼의 분량으로 데이터를 분할하여 학습 모델을 만듦

텍스트, 폰트, 화이트, 영수증이(가) 표시된 사진

자동 생성된 설명

샘플링 편향 : 데이터를 완전 무작위로 뽑으면 확률에 따라 올바른 모델을 만들 수 없는 데이터가 뽑힐 수 있음

이를 막기 위하여 train\_test\_split 함수를 이용하여 데이터를 나눌 규칙을 이상적으로 조절할 수 있음

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

앞의 2개 변수는 독립변수(함수의 첫번째 매개변수), 뒤의 2개는 종속변수(함수의 두번째 매개변수)로 지정됨, train = 훈련 데이터, test = 테스트 데이터

test\_size 매개변수 : 테스트할 데이터의 비율 (0.25는 테스트 데이터를 25%로 지정, 훈련 데이터는 나머지 75%가 됨)

stratify 매개변수 : 테스트할 데이터의 비율 지정, 위에서 label로 지정했으므로 테스트 데이터의 비율은 label의 비율을 따라감

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

훈련 데이터와 테스트 데이터의 분할하여 모델 훈련과 테스트 진행

텍스트, 스크린샷, 디스플레이, 도표이(가) 표시된 사진

자동 생성된 설명

도미같아 보이는 데이터를 빙어로 예측

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

Kneighbors 함수로 이웃 샘플의 인덱스를 반환, KNN의 기본 K값은 5임

텍스트, 스크린샷, 번호, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

단위 차로 인해 X축 (길이)의 숫자값의 영향은 크지 않고, y축 (무게)의 숫자값의 영향이 큼

텍스트, 스크린샷, 디스플레이, 라인이(가) 표시된 사진

자동 생성된 설명

길이는 모델이 값을 예측하는데 영향을 거의 주지 못했음

텍스트, 영수증, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

스케일링으로 해당 각 단위의 기준을 맞출 수 있음

텍스트, 영수증, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

스케일링으로 평균과 표준편차를 계산식에 대입하여 모든 데이터 분포를 표준 분포로 변환

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

StandardScaler() 클래스의 .fit\_transform 메소드로 데이터를 스케일링

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

이 경우 테스트 데이터도 훈련 데이터와 같은 기준으로 스케일링해줘야함

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

.transform으로 데이터를 스케일링

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

스케일된 데이터로 훈련 및 테스트 재진행시 올바른 결과 확인

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

* Flatten : 다차원 데이터를 1차원 데이터로 변환