2장

01.사이킷런 소개와 특징

사이킷런: 파이썬 머신러닝 랑브러리 중 가장 많이 사용됨

특징

- 쉽고 파이썬스러운 API 제공
- 다양한 알고리즘, 편리한 프레임워크 및 API 제공
- 오랜 기간 검증됨. 많은 환경에서 사용되는 성숙한 라이브러리

02. 첫 번째 머신러닝 만들어 보기 – 붓꽃 품종 예측하기

지도학습: 학습을 위한 다양한 피처와 분류 결정값이 레이블 데이터로 모델 학습 > 테스트 데이터 세트에서 미지의 레이블 예측. 즉, 명확한 정답이 주어진 데이터 세트 학습 후 미지의 정답 예측

사이킷런 패키지 내 모듈명: sklearn으로 시작

Sklearn.datasets 내의 모듈

> 사이킷런에서 자체적으로 제공하는 데이터 세트를 생성하는 모듈의 모임

load_00(): 데이터 세트 로딩

train_test_split(): 데이터 세트를 학습 데이터와 테스트 데이터로 분리

➔ 학습 데이터로 학습된 모델이 얼마나 뛰어난 성능을 가지는지 평가하려면 테스트 데이터 세트가 필요하기 때문에 반드시 분리 필요

➔ Test_size 파라미터의 입력 값의 비율로 분할
예) test_size =0.2 > 테스트 데이터 20%, 학습 데이터 80%

➔ Train_test_split(피처 데이터 세트,레이블 데이터 세트,데이터세트의비율,난수발생값)
예) train_test_split(iris_data, iris_label, test_size=0.2, random_state=11)

➔ Random_state는 호출할 때마다 같은 학습/테스트 용 데이터 세트를 생성하기 위함

데이터 세트로 분류를 예측한 프로세스

1. 데이터세트 분리
2. 모델 학습
3. 예측 수행
4. 평가

03. 사이킷런의 기반 프레임워크 익히기

메서드 이해하기

1. Fit( ): ML 모델 학습을 위해 사용
2. Predict( ): 학습된 모델의 예측을 위해 사용

사이킷런에서

분류 알고리즘을 구현한 클래스 > Classifier,

회귀 알고리즘을 구현한 클래스 > Regressor

Classifier + Regressor = Estimator

사이킷런 주요 모듈

- sklearn.datasets: 사이킷런에 내장되어 예제로 제공하는 데이터 세트
- sklearn.preprocessing: 데이터 전처리에 필요한 다양한 가공 긴으 제공
- sklearn.feature_selection: 알고리즘에 큰 영향을 미치는 피처를 우선순위대로 셀렉션 작업을 수행하는 다양한 기능 제공
- sklearn.feature_extraction: 텍스트 데이터나 이미지 데이터의 벡터화된 피처를 추출하는 데 사용
- sklearn.decomposition: 차원 축소와 관련한 알고리즘을 지우너하는 모듈
- sklearn.model_selection: 교차 검증을 위한 학습용/테스트용 분리 그리고 서치로 최적 파라미터 추출 등의 API 제공
- skleanr.metrics: 다양한 성능 측정 방법 제공
- sklearn.ensemble: 앙상블 알고리즘 제공
- sklearn.linear_model: 회귀 관련 알고리즘 지원, SGD 관련 알고리즘 제공

- sklearn.naive_bayes: 나이브 베이즈 알고리즘 제공
- sklearn.neighbors: 최근접 이웃 알고리즘 제공
- sklearn.svm: 서포트 벡터 머신 알고리즘 제공
- sklearn.tree: 의사 결정 트리 알고리즘 제공
- sklearn.cluster: 비지도 클러스터링 알고리즘 제공
- sklearn.pipeline: 피처 처리 등의 변환과 ML 알고리즘 학습, 예측 등을 함께 묶어서 실행할 수 있는 유틸리티 제공

분류와 클러스터링을 위한 표본 데이터 생성기

Datasets.make_classifications( ): 분류를 위한 데이터 세트를 만듦

Datasets.make_blobs( ): 클러스터링을 위한 데이터 세트를 무작위 생성

04. Model Selection 모듈 소개

Train_test_split()

- test_size: 전체 데이터에서 테스트 데이터 세트 크기를 얼마로 샘플링할지 결정
- train_size: 전체 데이터에서 학습용 데이터 세트 크기를 얼마로 샘플링할지 결정
- shuffle: 데이터 분리 전 데이터 섞을지 결정 (디폴트는 true)
- random_state: 호출할 때마다 동일한 학습/테스트용 데이터 세트 생성을 위한 난수 값

교차 검증: 데이터 편중을 막기 위해서 별도의 여러 세트로 구성된 학습 데이터 세트와 검증 데이터 세트에서 학습과 평가를 수행하는 것

과적합: 모델이 학습 데이터에만 과도하게 최적화되어 실제 예측을 다른 데이터로 수행할 경우에는 예측 성능이 과도하게 떨어지는 것

>이를 개선하기 위해 교차 검증이 필요

K 폴드 교차 검증: K개의 데이터 폴드 세트를 만들어서 K번만큼 각 폴드 세트에 학습과 검정 평가를 반복적으로 수행하는 방법

사이킷런에서는 KFOLD와 StratifiedKFOLD 클래스 제공

Stratified K 폴드: 불균형한 분포도를 가진 레이블 데이터 집합을 위한 K 폴드 방식

Cross_val_score(): 교차 검증을 편리하게 수행할 수 있게 해주는 API

폴드 세트 설정, for 루프로 인덱스 추출, 반복적으로 학습 및 예측 수행을 한번에 수행

gridsearchCV: 하이퍼 파라미터를 순차적으로 입력하면서 편리하게 최적의 파라미터를 도출할 수 있는 방안 제공

05. 데이터 전처리

사이킷런의 ML 알고리즘 적용 전 데이터에 대해 미리 처리해야할 기본 사항

결손값을 다른 값으로 변환하는 것

문자열을 숫자형으로 변환하는 것

데이터 인코딩

레이블 인코딩: 카테고리 피처를 코드형 숫자 값으로 변환하는 것

원-핫 인코딩: 피처 값의 유형에 따라 새로운 피처를 추가해 고유 값에 해당하는 칼럼에만 1을 표시하고 나머지 칼럼에는 0을 표시하는 방식. 즉, 행 형태로 되어 있는 피처의 고유 값을 열 형태로 차원을 변환한 뒤, 고유 값에 해당하는 칼럼에만 1을 표시하고 나머지 칼럼에는 0을 표시하는 것

피처 스케일링: 서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업

-표준화: 데이터의 피처 각각이 평균이 0이고 분산이 1인 가우시안 정규 분포를 가진 값으로 변환하는 것 > StandardScaler

-정규화: 서로 다른 피처의 크기를 통일하기 위해 크기를 변환해주는 것

>MinMaxScaler: 데이터값을 0과 1사이의 범위 값으로 변환

3장

분류의 성능 평가 지표에는 정확도, 오차행렬, 정밀도, 재현율, F1 스코어, ROC AUC가 있다

01.정확도

: 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 지표. 직관적으로 모델예측 성능을 나타냄.

이진 분류의 경우 데이터 구성에 따라 모델의 성능 왜곡할 수 있음

02.오차행렬

4분면 행렬에서 실제 레이블 클래스 값과 예측 레이블 클래스 값이 어떠한 유형을 가지고 매핑되는지 나타낸다

03. 정밀도와 재현율

정밀도: 예측을 positive으로 한 대상 중에 예측과 실제 값이 postivie으로 일치한 데이터 비율의 뜻

재현율: 실제 값이 positive인 대상 중에 예측과 실제 값이 positive으로 일치히나 데이터의 비율

04.F1 스코어: 정밀도와 재현율을 결합한 지표

정밀도와 재현율이 어느 한 쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가진다

05. ROC 곡선과 AUC

이진 분류의 예측 성능 측정에서 중요하게 사용되는 지표

ROC: 수신자 판단 곡선, 이진 분류 모델의 예측 성능 판단에 중요 지표

AUC: ROC 곡선 밑의 면적을 구한것으로 일반적으로 1에 가까울 수록 좋은 수치이다