2주차 예습과제 개념정리

2. 사이킷런 sklearn

- `sklearn.datasets`
- `sklearn.model_selection`
- `sklearn.tree`

▶ 분류 예측 수행 프로세스

1. 데이터 세트 분리
2. 모델 학습
3. 예측 수행
4. 평가

▶ Estimator : Classifier + Regressor

지도학습의 모든 알고리즘을 구현한 클래스

- `fit()`
- `predict()`

▶ 사이킷런의 주요 모듈

- 예제 데이터
- 피처 처리
- 피처 처리 & 차원 축소
- 데이터 분리, 검증 & 파라미터 튜닝
- 평가

- ML 알고리즘
- 유틸리티

▶ 내장된 예제 데이터 세트

- 사이킷런에는 연습용 데이터세트가 내제되어있음
- `fetch` 계열 명령 : 최초 사용 시에 인터넷 연결 필수

▶ `train_test_split()`

: 전체 데이터세트를 학습데이터와 테스트 데이터로 분리할 때 사용하는 함수

> 💡 `train_test_split()`의 반환값은 튜플형태이다

- 학습용 데이터의 피처 데이터 세트
- 테스트용 데이터의 피처 데이터 세트
- 학습용 데이터의 레이블 데이터 세트
- 테스트용 데이터의 레이블 데이터 세트

train_test_split() 파라미터

1. 피처 데이터 세트
2. 레이블 데이터 세트
3. 선택적으로 받는 파라미터들
    a. test_size
    b. train_size
    c. shuffle

d. random_state

▶ 교차검증

- K폴드 교차검증 : K개의 데이터 폴드세트 만들기 → K번만큼 각 폴드 세트에 학습과 검증 평가
- Stratified K 폴드 : for 불균형한 분포도를 가진 레이블 데이터 집합
- `cross_val_score()` : 사이킷런 제공 API, 교차 검증을 편리하게 할 수 있음

▶ GridSearchCV

- 교차 검증과 최적 하이퍼 파라미터 튜닝을 한 번에

- 편리 but, 수행시간이 상대적으로 오래 걸림

GridSearcdCV 주요 파라미터

- estimator
- param_grid
- scoring
- cv
- refit

▶ 데이터 전처리

- NULL 값 처리하기
- 문자열 값 인코딩 → 숫자형

▶ 데이터 인코딩

- 레이블 인코딩 : 카테고리 피처 → 코드형 숫자

- 원-핫 인코딩 : 행 형태 피처 고유값 → 열 형태 / 고유 값에 해당하는 칼럼에만 1, 나머지 칼럼은 0

▶ 피처 스케일링

: 서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업

- 표준화
- 정규화

▶ StandardScaler

- 개별 피처를 평균이 0이고 분산이 1인 값으로 변환
- 가우시안 정규 분포를 가질수 있도록 함

▶ MinMaxScaler

- 데이터값을 0과 1사이의 범위값으로 변환
- 음수 값이 있다면 -1 ~ 1

3. 평가

▶ 분류 성능 평가 지표

- 정확도 : 예측 결과가 동일한 데이터 건수 / 전체 예측 데이터 건수
  - = (TN + TP)/(TN + FP + FN + TP)
- 오차행렬 : 이진 분류의 예측 오류가 얼마인지 + 어떠한 유형의 예측 오류가 발생하고 있는지
  - TN, FP, FN, TP
- 정밀도 = TP / (FP + TP)

    - 정밀도 100% 방법 : 확실한 기준이 되는 경우만 P 나머지는 모두 N
- 재현율 = TP / (FN + TP)
    - 재현율 100% 방법: 모든 환자를 P로 예측
- F1 스코어 : 정밀도와 재현율이 한 쪽으로 치우치지 않을 때 👍
- ROC 곡선 , AUC