Chapter 01

넘파이

1장 - 파이썬 기반의 머신러닝과 생태계

머신러닝 개요

머신러닝 : 데이터를 활용하여 패턴을 학습하고 예측, 분류, 군집화 등의 작업을 수행하는 기술 분야

파이썬과 머신러닝

파이썬 : 머신러닝 분야에서 널리 사용되는 프로그래밍 언어

머신러닝을 위한 주요 라이브러리로는 NumPy, pandas, scikit-learn, Matplotlib 등이 있다.

넘파이

넘파이 기반 데이터 타입 : ndarray

Array() : 다양한 인자 입력받은 후 ndarray로 변환

Ndarray.shape : ndarray의 차원과 크기를 튜플 형태로 나타냄

Ndarray를 0또는 1로 초기화해 쉽게 생성해야 할 때

Arrange(), zeroes(), ones()이용

Reshape()메서드 : ndarray를 특정 차원 및 크기로 변환

인덱싱(Indexing): 일부 데이터 세트나 특정 데이터만을 선택 가능

-slicing: 연속된 인덱스상의 ndarray 추출

-Fancy indexing : 일정한 인덱싱 집합 리스트 또는 ndarray로 지정해 데이터의 ndarray 반환

-Boolean indexing : 특정 조건의 True/False 값 인덱싱 집합을 기반으로 True에 해당하는 데이터

판다스

행과 열로 이루어진 2차원 데이터를 효율적으로 가공 및 처리

판다스의 핵심 객체 DAtaFrame

DataFrame: 여러 개의 행과 열로 이뤄진 2차원 데이터를 담는 데이터 구조체

Values: DataFrame을 넘파이 ndarray로 변환

Drop() 메서드 : DataFrame의 데이터 삭제

판다스의 index 객체 : DataFram, Series의 레코드를 고유하게 식별하는 객체

Reset index() 인덱스가 연속된 int 숫자형 데이터가 아닐 경우에 다시 이를 연속 int 숫자형 데이터로 만들 때 주로 사용

DataFrame의 로우나 칼럼을 지정하여 데이터 선택하는 인덱싱 방식으로 iloc[]와 loc[] 제고

Loc: 명칭기반 인덱싱 방식

Iloc[] : 위치 기반 인덱싱 방식

Sort_values() : dataframe과 series의 정렬

결손 데이터 처리

Isna(): 결손 데이터 여부 확인

Fillna(): 결손 데이터를 편리하게 다른 값으로 대체

Chapter02

2장 - 사이킷런으로 시작하는 머신러닝

사이킷런 소개

파이썬의 머신러닝 라이브러리로 간단하고 효과적인 API를 제공

사이킷런의 주요 모듈

데이터셋 로딩, 데이터 분할, 모델 학습, 예측, 평가 등을 지원하는 모듈이 있음

GridSearchCV

교차 검증과 최적 하이퍼 파라미터 튜닝을 한 번에

주요 파라미터

-estimator'

-scoring: 예측 성능을 측정할 평가 방법을 지정

Cv: 교차 검증을 위해 분할되는 학습/테스트 세트의 개수를 지정

Refit: 디폴트가 true이며 true로 생성 시 가장 최적의 하이퍼 파라미터를 찾은 뒤 입력된 estimator객체를 해당 하이퍼 파라미터로 재학습

Ex) 키의 데이터 세트 구성

Data : 피처의 데이터 세트

Taget : 분류 시 레이블 값, 회귀일 때는 숫자 결괏값 데이터 세트

Target_name : 개별 레이블의 이름

Feature_names : 피처의 이름

DESCR : 데이터 세트에 대한 설명과 각 피처의 설명

<u>데이터 전처리</u>

데이터 인코딩

레이블 인코딩- LAbelEncoder를 객체로 생성한 후 fit()와 transform()을 호출해 레이블 인코딩 수행

원 핫 인코딩- 피처 값의 유형에 따라 새로운 피처를 추가해 고유 값에 해당하는 칼럼에만 1표시 나머지 칼럼에는 0표시

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

사이킷런에서 대표적인 스켈링 클래스는 StandarSaclaer와 MinmaxScaler가 있음

Chapter03

3장 - 평가

평가 지표

분류와 회귀 모델의 성능을 평가하기 위한 다양한 평가 지표

분류에서는 정확도, 정밀도, 재현율, F1 스코어 등을 다룸

정밀도와 재현율

Positive 데이터 세트의 예측 성능에 좀 더 초점을 맞춘 평가 지표

정밀도= TP / (FP+TP)

재현율= TP / (FN+TP)

정밀도와 재현율의 Tradeoff : 서로 상호 보완적인 평가 지표이기 때문에 어느 한 쪽을 강제로 높이면 다른 하나의 수치는 떨어짐

F1 Score : 정밀도와 재현율을 결합한 지표

- 정밀도와 재현율이 어느 쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값

Roc 곡선과 AUc : 이진 분류의 예측 성능 측정에서 중요하게 사용되는 지표

Roc 곡선은 FRR이 변할 때 TPR이 어떻게 변화하는지 나타냄

특이성 (TNR)

FPR=FP/(FP+TN)=1-TNR=1-특이성

사이킷런은 ROC 곡선을 구하기 위해 roc_curve() API를 제공

Roc_curve의 주요 파라미터

입력 파라미터

-y_true : 실제 클래스 값 array

-y_score: predict_proba()의 반환 값array에서 Positive 칼럼의 에측 확률이 보통 사용

반환값

-fpr: fpr값을 array로 반환

-tpr: tpr값을 array로 반환

-Thresholds: threshold 값 array

타이타닉 생손자 예측 모델 FPR,TPR, 임계값 – roc_curve 이용해 구함

AUC 값은 ROC 곡선의 밑의 면적을 구한 것으로 일반적으로 에 가까울수록 좋은수치

마지막으로 get_clif_eval() 함수에 roc_auc_score 이용해 ROC AUC값 측정