Euron 1주차 예습과제_개념정리_한송희

Chapter1 파이썬 기반의 머신러닝과 생태계 이해

01. 머신러닝의 개념

머신러닝: 애플리케이션을 수정하지 않고도 데이터를 기반으로 패턴을 학습하고 결과를 예측하는 알고리즘 기법. 데이터를 기반으로 숨겨진 패턴을 인지해 문제를 해결함

<분류>

-지도학습: 분류, 회귀, 추천 시스템, 시각/음성 감지/인지, 텍스트분석, NLP

-비지도학습: 클러스터링, 차원 축소, 강화학습

*머신러닝 알고리즘과 모델 파라미터 구축 능력도 중요하지만.....머신러닝은 데이터에 매우 의존적이기 때문에 데이터 이해, 가공, 처리, 추출 하는 능력이 더 중요할 것임

<파이썬vsR>

R: 통계전용 프로그램 언어→개발 언어는 익숙하지 않지만 통계분석에 능한 사람 추천

파이썬: 다양한 영역에서 사용되는 개발 전문 프로그램→머신러닝을 시작하는 개발자에게 추천(대부분의 프로그램이 파이썬을 기반으로 만들어지므로 굳이 배울거면 파이썬 추천)

02. 파이썬 머신러닝 생태계를 구성하는 주요 패키지

머신러닝 패키지: 사이킷런(Scikit-Learn)

행렬/선형대수/통계 패키지: 넘파이(Numpy), 사이파이(SciPy)

데이터 핸들링: 판다스(Pandas)

시각화: 맷플롯립(Matplotlib), 시본(Seaborn)

03. 넘파이

Numpy=Numerical Python: 파이썬에서 선형대수 기반의 프로그램을 쉽게 만들 수 있도록 지원하는 패키지, 배열 기반 연산 및 데이터 핸들링 제공

넘파이 이해는 파이썬 기반 머신러닝에서 매우 중요함

<실습>

import numpy as np : 약어로 모듈을 표현

np.array() : ndarray로 변환을 원하는 객체를 인자로 입력하면 ndarray로 반환

.shape: 행과 열의 수를 튜플 형태로

.ndim: array의 차

.dtype: 데이터 타입 확인

.astype():ndarray 내 데이터값 타입 변환

arange(): 0부터 인자값 -1까지 순차적으로 ndarray 데이터 값으로 변환

zeros(): 튜플 형태의 shape값을 입력하면 모든 값을 0으로 채운 ndarray 반환

ones(): 튜플 형태의 shape값을 입력하면 모든 값을 1로 채운 ndarray 반환

reshape(): 차원과 크기 변환 + -1을 인자로 사용하면 원래 ndarray와 호환되는 새로운 shape으로 자동으로 변환

인덱싱

-슬라이싱: 연속된 인덱스 상 ndarray 출력

-팬시 인덱싱: 일정한 인덱싱 집합을 리스트, ndarray형태로 저장

-불린 인덱싱: True/False 값 인덱싱 집합을 기반으로 True에 해당하는 인덱스의 ndarray 반환

sort(): np.sort()는 기존 행렬을 유지한 채 정렬된 행렬을 반환, ndarray.sort()는 기존 행렬 자체를 정려란 형태로 변환하고 None를 반환

argsort(): 정렬 행렬의 원본 행렬 인덱스를 ndarray형으로 반환

numpy를 이용해 선형대수계산도 가능

-transpose(): 전치행렬

04. 판다스

데이터 핸들링을 편하게 하기 위한 패키지.

Index: 개별 데이터를 고유하게 식별하는 key값

Series: 칼럼이 하나뿐인 데이터 구조체

DataFrame: 칼럼이 여러 개인 데이터 구조체=여러개의 Series로 구성

<실습>

df=pd.read_csv('경로'): dataframe으로 불러오기

DataFrame.drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise'): 데이터 삭제+inplace=True로 하면 자신

의 DataFrame의 데이터를 삭제함

reset_index(): 새롭게 인덱스를 연속 숫자형으로 할당(기존은 인덱스는 새로운 칼럼명 index로 추가)

value_counts(): 고유값 개수 반환

+)DataFrame 뒤에 있는 [ ]는 칼럼만 지정하는 연산자로 생각

ix[ ]: 칼럼 명칭 기반 인덱싱. 현재는 사라졌음

iloc[행,열]연산자: 위치 기반 인덱싱

loc[인덱스값,칼럼명]연산자: 명칭 기반 인덱싱 + 슬라이싱을 '시작점:종료점'으로 지정할 경우 시작점에서 종료점을 포함한 위치에 있는 데이터를 반환함

불린 인덱싱: 조건 적용해 참인 경우만 추출

sort_values(): 정렬

aggregation함수: min, max, sum, count, mean → 모든 함수에 적용

groupby(): 파라미터를 by칼럼에 입력하면 대상 칼럼으로 groupby된다

isna(): 결손 데이터 여부 확인

fillna(): 결손 데이터 대체