5주차

데이터셋 선정

* 종속변수가 Label 되어 있는지

있다면 🡪 지도학습, 분류 문제 등

없다면 🡪 비지도학습

* 수치, 텍스트, 이미지, 음성, 비디오 데이터
* 데이터 별 특성에 기반하여 어떤 분석을 할지, 어떤 알고리즘을 적용할지 선택
* 수치 데이터: 양적분석

텍스트 데이터: 질적분석 🡪 이슈와 연관지어 생각해보기

앙상블 머신러닝: 여러 모델들을 합쳐서 적용하는 것

논문 작성 Tip

* 제목, 요약, Figure 등은 영문 해석 첨부
* 수식 🡪 인덱스하기
* 참고문헌: 이름은 풀네임, 성은 약어 + “논문명” + 저널명(이탤릭체) + 발행연도 + 인용면수
* 저자로서의 책임을 지기! (표절 등)

연구방법론: 연구&논문&문제해결능력 향상

1. 문제정의

* 최근 트렌드를 반영한 가치있는 문제인가?
* 내가 제일 흥미롭고 풀어보고 싶고 재미있어하는 분야가 맞는가? (전공, 관심사, 경험 …)

🡪 인내심+끈기 🡪 원동력 🡪 연구, 논문, 꿈, 기회, 직업 🡪 “행복”

* 현실성 반영(개인의 전공, 선수과목 + 경험 + 프로그래밍 수준 + 머신러닝(지도/비지도), 딥러닝, 강화학습…)

1. 선택한 문제에서 활용할 데이터 타입을 고민(예: 크롤링 수집, 영어로 구글링+다운로드, csv 및 txt 파일)

* 데이터 타입

a. 수치

b. 텍스트

c. 이미지

d. 비디오

1. 기존 최신 대가들의 국제/국내 논문들을 읽고, 분석하고, 비교 테이블 및 서술형으로 요약을 할 수 있는 문해력

* 비교테이블

1. 제안하는 기법(아이디어 제안) + 파이썬/R 구현(프로그래밍 능력) 🡪 결과 그래프 시각화+해석(domain expert)
2. 글을 논리적으로 작성할 수 있는 능력

예시 논문

앙상블 모델

* 첫번째 모델(하드 보팅): 각각의 classifier가 예측한 class 수가 더 높은 값 = 최종값
* 두번째 모델(소프트 보팅): 각각의 classifier가 확률적으로 labeling expectation의 average probability의 합 = 최종값

\*모델 간 가중치 부여를 통한 위계를 설정하는 것 또한 tip.