|  |  |
| --- | --- |
| 교육 제목 | **ML modeling process**  **ML 전처리** |
| 교육 일시 | 2021.10.13 |
| 교육 장소 | 비대면 |
| **교육 내용** | |
| 오전 | 범주형 변수  연속형 변수  **-단순 회귀(simple linear regression)**    **-최소 제곱 추적 법 (OLS ordinary least square estimation)**    이미 이차함수 형태  함수의 미니멈은 에러가 최소화 되는 점  **-손실함수(lost function, j)**      **머신 러닝**  Anaconda prompt 상에서 scikit-learn 설치 proceed -> [y]    필요한 모듈을 먼저 install |
| 오후 | **ML modeling process**  머신 러닝  -학습: 사람의 뇌 (Memory) -> 기계 (패턴 인식)  -  AI (artificial intelligence) e.g) decision tree ,l f  ML (machine learning) 통계적인 방법 계발되기 시작, e.g) SVM, random forest  DL (deep learning) 컴퓨팅 파워의 덕 2010부터 e.g) CNN, RNN  -Image classification : RGB매트릭스를 이미지로 인식해서 구분하는데 탁월  -기계학습 분류    **ML 전 처리**  -Feature 들에 대한 전 처리(preprocessing)  -초 매개변수 조절(tuning hyperparameters)  -모델 성능평가(assessing model performance)  **Training set and test set**  -ml 모델의 성능평가를 위해 자료분할    **일반적인 ML 예측 과정**  -학습세트(training set): 머신 러닝 모델을 학습할 때 사용  -검증세트(validation set) 하이 퍼 파라메터 경정할 때  -테스트(test): 학습된 모델을 평가할 때    **치우침(bias)- 분산(variance) trade-off**  너무 단조로우면 bias 너무 복잡하면 variance      **Training set 과 test set 분할**  -무작위 샘플링(random sampling)  -계층화 샘플링 (stratified random sampling): 각 계층에서 일정한 비율  **기계학습 모델 평가**  -k-fold 교차 검증 (CV -cross validation)  -k는 10번 정도면 충분하다 ->10개의 값 |