**교육일지**

**교육 제목 : 머신러닝**

**교육 장소 : YGL C6 강의실**

**교육 일시 : 2021/10/13**

**머신 러닝(machine learning)**

**기계학습의 분류**

**-supervised learning(학습데이터에 레이블(정답) 포함)**

**-unsupervised learning(학습데이터에 래이블 없음)**

**-semi supervised learning(레이블이 있는 데이터 & 없는 데이터) -> supervised & unsupervised의 조합으로 이루어짐**

**-reinforcement learning(에이전트(Agent)가 상태(State)의 보상(reward)을 최대화 하기 위해서 행동함)**

**머신러닝 수업 진행**

**ML전처리**

**-Feature들에 대한 적절한 전처리(pre-processing)**

**-초매개변수 조절(Tuning hyperparemeters)**

**-모델 성능 평가(Assessing model performance)**

**ML알고리즘**

**-ML 알고리즘 종류**

**-모수 (매개변수) 학습 (Parameter Learning)**

**-파이썬을 이용한 실습**

**Traning set and test set**

**·ML 모델의 성능평가를 위해서 자료를 분할**

**·Training set: 모델의 알고리즘 learning, 모델에 사용될 feature들을 결정, 초매개변수 조절 (약 전체 자료수의 70% 로 설정)**

**-Training set: 모델의 알고리즘 learning**

**-Validaation set: 모델에 사용될 feature들을 결정, 초매개변수 조절, 과적합 (Over-fitting) 방지**

**·Test set: 최종 선택된 모델의 성능평가 (약 전체 자료수의 30% 로 설정), 자료의 수가 적을경우 생략 가능**

**일반적인 ML예측 과정**

**-학습세트(Traning set): 머신러닝 모델을 학습할 떄 사용**

**-검증세트(Validation set): 하이퍼 파라메터 결정할 때 사용**

**-테스트세트(testing set): 학습된 모델을 평가할 때 사용**

**치우침(Bias)-분산(variance) trade- off**

**-치우침(Bias)**

**·치우침은 모델의 실제값(또는 평균)과 예측지 간의 차이를 의미**

**·과소적합(underfitting)은 치우침이 높은 모델이 발생하기 쉬움**

**-분산(variance)**

**·분산은 주어진 데이터에서 모델 예측은 변이로 정의**

**·과적합(overfitting)은 분산이 높은 모델이 발생하기 쉬움**

**기계학습 모델 평가**

**k-fold 교차 검증 (k-fold cross validation(CV)**

**·k-fold 교차검증 (일명 k-fold CV)은 훈련 데이터를 동일한 크기의 k그룹으로 무작위로 나누는 리샘플링 방법**

**·k-fold CV 추정치는 k테스트 오류를 평균화하여 계산**