|  |  |
| --- | --- |
| 교육 제목 | 머신러닝 모델링 |
| 교육 일시 | 10월 14일(금) |
| 교육 장소 | 재택 |
| **교육 내용** | |
| 전체 내용 | - Bootstrapping  부트스트래핑 샘플은 복원추출을 이용한 데이터의 무작위 샘플  부트스트래핑은 선택한 샘플을 기반으로 모델을 구축하고 OOB(Out-of-Bag)샘플을 이용하여 모델을 평가  - 초매개변수 조절(Hyperparameter tuning)  초매개변수는 학습 과정을 제어하는 데 사용되는 매개 변수를 의미  초매개변수는 모델 학습과정이 아닌 모델 개발에 의해서 지정됨  - K-nearest neighbors classification  지도학습으로서 분류(Classification) 나 회귀(Regression)에 사용되는 비모수적 방법  파라메터 학습을 위한 훈련과정이 없으나 훈련집합은 필요  각 데이터 간에 거리를 계산하기 위한 거리척도가 필요  초매개변수k를 설정해야 함  거리에 대한 가중치  - 기하학적 거리  유클리안, 맨해튼, 민코프스키, 고워  - 반응변수 전처리(Target engineering)  Log transformation, Feature 표준화, 결측치의 처리, 결측치 대체, 중요하지 않은 Feature 제거, 제로 분산 feature,  범주형 데이터 engineering  Lumping(재범주화) : 매우 작은 빈도를 갖는 범주들을 모아서 하나의 범주로 재범주화  one-hot & dummy encoding : 각 범주를 1 또는 0(True of False)로 표시  Label encoding : 각 범주 자료를 연속형 변수로 바꾸어 표현(순서형 자료의경우)  차원축소 : 여러개의 feature에서 불필요한 feature들 제거하는 방법  - 모델평가 지표  회귀분석 모델(Regression models)  MSE, RMSE, MAE  분류모델(Classification models)  Misclassification, Mean per class error, MSE, Cross entropy, Gini Index  confusion matrix(오분류표)  ROC(Receiver Operation Characeristic curve)와 AUC(Area under the curve)  좋은 분류모델은 높은 정밀도와 감도 가지게 되고 오분류율(위양성 또는 위음성)을 최소화 함 |