|  |  |
| --- | --- |
| 교육 제목 | **Modeling process** |
| 교육 일시 | 2021.10.14 |
| 교육 장소 | C6 |
| **교육 내용** | |
| 오전 | 기계학습 모델 평가  -**bootstrapping** /여러 번 뽑히거나, 한번도 안 뽑히거나  -샘플은 복원추출을 이용한 데이터의 무작위 샘플 (확률을 항상 유지)  -선택한 샘플을 기반으로 모델을 구축하고 oob샘플을 이용하여 모델을 평가  (한번도 안뽑리는)  -**초매개변수(hyperparameter tuning**)  -학습과정을 제어하는 데 사용되는 매개변수를 의미  -모델학습과정이 아닌 모델 개발자에 의해서 지정됨    -**K-nearest neighbors classification** = KNN regression  모델이 없고 데이터를 기반,k가 적을수록 overfitting      **기하학적 거리(Geometric distance measures)**    **초 매개변수 조절을 위한 격자탐색(grid search)알고리즘**    문제가 되는 것:자료의 전처리 공평하게 해야하는데 한쪽 데이터를 기준으로 할 때    **반응변수 전처리(target engineering)**  퍼짐의 정도가 다 다르면 좋은 모양이 안나온다      Log transformation: 오른쪽으로 치우친 분포가 정규 분포로 변환  Feature 표준화(standardization) 모든 모형을 통틀어 필요하다      **결측 치 종류**  무작위 결측 치 (Random missing value)  -완전 무작위 결측 치 (MCAR: Missing Completely At Random)  - 예) 단순한 결측 치  -무작위 결측 치 (MAR: Missing At Random)  - 예) 여성(X1)의 경우 체중(X2)에 대한 답이 없음  -비 무작위 결측 치 (NMAR: Not Missing At Random)  - 체중(X2) 무거운 사람은 체중(X2)에 대한 답이 없음 |
| 오후 | -**결측 치 개체 방법에 따른 비교 imputation**    -**중요하지 않은 feature 제거 (filtering)**    **제로분산feature (Zero variance features)**  : 제로 분산 features를 판단하는 일반적 기준  전체 샘플 중에 서로 다른 관측 값의 비율이 낮은 경우 (약 ≤ 10%)  가장 빈도가 높은 관측 값과 두 번째로 높은 관측 값 과의 비가 높은 경우 (약 ≥ 20배)  =discreto  **범주형데이터(Categorical feature)engineering**  -재범주화 (Lumping) = 작은 빈도의 범주들을 모아서 하나의 범주로 재범주화  -One-hot & dummy encoding = 각 범주를 1 또는 0 으로 표시(수학적 의미x)  -Label encoding =범주 자료를 연속형 변수로 바꾸어 표현 (순서형 자료의 경우)  -Replacing with the mean or proportion  차원축소(Dimension reduction)  여러 개의 feature에서 불필요한 feature들 제거하는 방법  예) 주성분 분석 (PCA, principal components analysis)  **모델평가 지표**  **---**    ---    **분류 모델 (Classification models)**  -Misclassification  -Mean per class error  -MSE  -Cross entropy  -Gini Index    -ROC(Receiver Operating Characteristic curve)와 AUC(Area under the curve)  좋은 분류모델은 높은 정밀도와 감도 가지게 되고 오분류 율 (위양성 또는 위음성)을 최소화 함 |