|  |  |
| --- | --- |
| 교육 제목 | 머신러닝 |
| 교육 일시 | 2021-10-14 |
| 교육 장소 | C5 강의실 |
| **교육 내용** | |
| 오전 | 기계학습 모델 평가    10-fold CV(n=32)    초매개변수 조절(Hyperparameter tuning)   * 초매개변수는 학습 과정을 제어하는데 사용되는 매개변수를 의미 * 초매개변수는 모델 학습과정이 아닌 모델 개발자에 의해서 결정     KNN (K-nearest neighbors classification)   * 지도학습으로서 분류나 회기에 사용되는 비모수적 방법 * 파라메타 학습을 위한 훈련과정이 없으나 훈련집합은 필요 * 각 데이터 간에 거리를 계산하기 위한 거리척도가 필요 * 초매개변수 k를 설정해야 함 * 거리에 대한 가중치       Feature표준화   * 각 각 Feature의 측정 단위에 대한 보정 (각 아파트 값을 추정하기 위한 feature들 중 평수 (30평)과 주변지역 땅값(3,000,000/평) * Centering and scaling을 통하여 평균이 0, 표준편차가 1이 되도록 변환 |
| 오후 | 결측치의 종류   * 완전무작위 결측치 (MCAR: Missing Completely At Random)   단순한 결측치   * 무작위 결측치 (MAR: Missing At Random)   예) 여성(X1)의 체중(X2)에 대한 답이 없음  - 비무작위 결측치 (NMAR: Not Missing At Random)  예) 체중(X2) 무거운 사람은 체중(X2)에 대한 답이 없음  결측치 대체   * 결측치를 “최상의 추측” 값으로 대체 * Estimated statistic (e.g, Mean, Median, Mode, Regression) * KNN * Tree-based         제로분산 feature를 판단하는 일반적인 기준   * 전체 샘플중에 서로 다른 관측값의 비율이 낮은 경우 (약 10%이하) * 가장 빈도 높은 관측값과 두 번째로 높은 관측값의 비가 높은 경우(약 20배 이상)   범주형 데이터   * Lumping : 매우 작은 빈도를 갖는 범주를 모아 재범주화 * One-hot & dummy encoding : 각 범주를 1 또는 0 (True or False)로 표시 * Label encoding : 각 범주 자료를 연속형 변수로 바꾸어 표현   Eg) very High(=5), high(=4), moderate(=3), low(=2), very low(=1)  차원 축소   * 여려개의 Feature에서 불필요한 Feature를 제거하는 방법   예) 주성분 분석(PCA, principa components analysis)    모델 평가 지표 |