|  |  |
| --- | --- |
| 교육 제목 | **머신러닝** |
| 교육 일시 | 2021-10-1 |
| 교육 장소 | C06 |
| **교육 내용** | |
| 오전 | 기계학습모델 평가  : Bootstrapping 샘플은  복원 추출물을 이용한 데이터의 무작위 샘플  선택한 샘플을 기반으로 모델 구축 OOB(Out-of-Bag) 샘플을 이용하여 모델을 평가  OOB : 한번도 뽑히지 않은 데이터 샘플  초매개변수 조절(Hyperparameter tuning)  : 학습과정을 제어하는데 사용되는 매개변수를 의미  모델 학습과정이 아닌 모델 개발자에 의해서 지정됨  KNN – K-nearest neighbors classification  지도학습으로 분류, 회귀에 사용되는 비모수적 방법  파라메터 학습을 위한 훈련과정이 없으나 훈련집합은 필요  각 데이터 간에 거리를 계산하기 위한 거리척도 필요  초매개변수 k를 설정 해야함  거리에 대한 가중치  Overfitting 과적합  Underfitting  새로운 값(예측점) 가장 가까운 값 데이터 집합으로 변환    K = 10 개 일 때 예외처리 해야 됨  이때 초 매개변수 : k 값, 데이터 간 거리  기하학적 거리    공간위에 AB직선의 길이는        확인문제    일 때 new 는 A, B중 무엇일까?    계산 방법    결과 값  K = 1 new -> a  K = 3 new -> a  K = 5 new -> b  확인문제    K = 2 일 때  KNR :    초매개변수 조절위한 알고리즘 : 격자탐색(grid search)      기계적 최저점  데이터 값 동그라미 부분부터 떨어질 일 없음  지금까지 의 문제점 : 데이터의 단위, 범위가 다르면 값을 구하지 못함  예) x1 = 아파트 평수 ( 20~ 50) x2 : 땅값(3000 ~ 5000)  Target  Feature  반응변수 전처리 ( Target engineering)  : 주로 parametric model에서 예측 및 모델 적용을 위해서 사용   1. Log transformation   오른쪽으로 치우친 분포가 정규분포로 변환     1. Box-cox transformation   Feature 표준화 (Standardization)  각각 feature의 측정 단위에 대한 보정 -> 예) 위에 단위 문제 해결방법      결측치의 처리  **결측치 종류**  무작위 결측치 (Random missing value) -> 단순한 결측지  완전무작위 결측치 (MCAR: Missing Completely At Random)  무작위 결측치 (MAR: Missing At Random)  비무작위 결측치 (NMAR: Not Missing At Random) - - |
| 오후 | 결측치대체(Imputation)  결측치를 “최상의 추측”값으로 대체  Estimated statistic (e.g., Mean, Median, Mode, Regression)  K-nearest neighbor  Tree-base  중요하지않은 feature 제거 하는방법 : filtering  의미없는 변수들을 포함했을 때 RMSE의 변화  제로 분산 : 판단하는 일반적 기준  전체 샘플중에 서로 다른 관측값의 비율이 낮은 경우 (약 ≤ 10%)  가장 빈도가 높은 관측값과 두 번째로 높은 관측값 과의 비가 높은 경우 (약 ≥ 20배)  차원축소 : 여러 개의 feature에서 불필요한 feature들 제거  주성분 분석 (PCA, principal components analysis)  모델의 지표      분류 모델  Misclassification    Mean per class error 전체중에 평균    MSE  Cross entropy  Gini Index  Confusionmatrix예제    Sensitivity = Recall    간단하게 있는 그래도 계산하면 됨  모델 평가지표  연속형 -> R\*\*2 자주 사용함  범주형 -> 혼동행렬 자주 사용함  ROC(Receiver OperatingCharacteristic curve)  AUC(Area underthe curve) : 면적 다 맞추면 1 반만 맞추면 0.5  좋은 분류 모델은 높은 정밀도와 감도 가지게 되고 오분류율을 최소화함 |