|  |  |
| --- | --- |
| 교육 제목 | Decision tree, K-means clustering, Cross validation |
| 교육 일시 | 10월 20일(수) |
| 교육 장소 | 재택 |
| **교육 내용** | |
| 전체 내용 | - Decision tree  의사결정나무모형은 나무 구조러 나타내어 전체 자료를 몇 개의 소집단으로 분류하여 예측을 수행하는 분석방법이다. 상위 노드로부터 하위 노드로 나무 구조를 형성하는 매 단계마다 분류 변수와 분류 기준값의 선택이 중요하다.  데이터가 얼마나 잘 분리 되었는지를 평하기 위해서는 기준이 필욯나데 일반적으로 불순도(impurity) 기준을 사용하며, 노드에 여러 분류가 섞여 있을수록 높고, 노드에 하나의 분류만 존재할 때 가장 낮아진다. 즉 노드 분리 후 각 노드의 불순도가 낮아질수록 트리 분류가 잘 된 것으로 볼 수 있다.  상위 노드에서의 분류변수, 분류 기준값은 이기준에 의해 분기되는 하위노드에서 노드내에서는 동질성이, 노드 간에는 이질성이 가장 커지도록 선택 된다. 나무 모형의 크기는 과대적합 또는 과소적합 되지 않도록 합리적 기준에 의해 적당히 조절되어야 한다.  제일 상위의 마디를 뿌리마디(root node)라 하고 이는 분류 대상이 되는 모든 자료집단을 포함한다. 상위 마디가 하위 마디로 분기될 때, 상위 마디를 부모 마디 또는 부모 노드락 하고하위 마디를 자식 마디 또는 자식 노드라하며 더 이 상 분기되 지 않는 마디를 최종 마디 (terminal node)라고 한다.  가지분할(spilit)은 나무의 가지를 생성하는 과정이고 가지치기(prunig)는 생성된 가지를 잘라내어 모형을 단순화하나는 과정을 말한다.  의사결정나무는 분류와 회귀 모두 다 가능하다. 즉 범주나 연속형 수치 모두 예측 할 수 있다. 목표 변수가 유한한 수의 값을 가지면 분류목적의 분류 나무이고 실수값을 가지면 수치예측 회귀 나무가 된다.  - K-means clustering  k-평균 알고리즘은 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다. 이 알고리즘은 EM 알고림을 이용한 클러스터링과 비슷한 구조를 가지고 있다.  - Cross validation  교차검증에서 우리는 데이터 셋의 서로 다른 부분 집합에 대해 모델을 평가하고 여러개의 모델 품질 척도를 얻는다. 교차 검증은 보다 정확한 모델 품질 척도를 제공하는데, 이는 모델을 결정하는 데에 특히 중요하다. 그러나 이는 여러 모델의 각의 폴드를 조사하므로, 시간이 더 오래 걸릴 수 있다. 따라서 계산 부담이 적은 작은 데이터 셋에서 교차 검증을 실행해야 한다. 더 큰 데이터 셋의 경우 단일 검증 셋만으로 충분하다. 그게 더 빠르고, 충분한 데이터가 있다면 교차 검즈을 해야할 필요성이 떨어진다. 큰 데이터 셋과 작은 데이터 셋을 구하하는 특별한 기준은 없다. 하지만 만약 모델이 실행되는데 몇 분 밖에 걸리지 않는다면, 교차 검증으로 전환하는것이 좋다. 또한 교차 검증을 실시하고 각 실험의 점수들이 비슷한지 확인할 수 있다. 각 실험에서 동일한 결과가 나온다면, 이는 단일 검증 셋을 사용했어도 충분했다는 뜻이다. |
|  |  |